Statistica Sinica **21** (2011), 369-389

THEORY OF GAUSSIAN VARIATIONAL APPROXIMATION FOR A POISSON MIXED MODEL

Peter Hall, J. T. Ormerod and M. P. Wand

University of Melbourne, University of Sydney and University of Wollongong

Abstract: Likelihood-based inference for the parameters of generalized linear mixed models is hindered by the presence of intractable integrals. Gaussian variational approximation provides a fast and effective means of approximate inference. We provide some theory for this type of approximation for a simple Poisson mixed model. In particular, we establish consistency at rate $m^{-1/2} + n^{-1}$, where m is the number of groups and n is the number of repeated measurements.

Key words and phrases: Asymptotic theory, generalized linear mixed models, Kullback-Liebler divergence, longitudinal data analysis, maximum likelihood estimation.

1. Introduction

Variational approximation has become a central component in inference in Machine Learning and other areas of Computer Science. Recent summaries of variational approximation methodology and theory are provided by Jordan (2004), Titterington (2004), and Bishop (2006). The `Infer.NET` software project (Minka et al. (2009)) is a manifestation of the numerous areas in which variational approximations are being applied. Almost all of this work involves Bayesian inference.

Statistical areas such as longitudinal data analysis have issues that are similar to those arising in Machine Learning. Recently, we have explored the transferral of variational approximation technology to statistical settings. One of these is likelihood-based, rather than Bayesian, inference for generalized linear mixed models. A particularly appealing approach in this context is *Gaussian variational approximation*, which involves minimum Kullback-Liebler divergence from a family of Gaussian densities. Details on Gaussian variational approximation for generalized linear mixed models are given in Ormerod and Wand (2009, 2010).

The present article is concerned with theoretical aspects of Gaussian variational approximations to maximum likelihood estimation. Almost all of the variational approximation theory of which we are aware treats Bayesian inferential settings (e.g., Humphreys and Titterington (2000); Wang and Titterington (2006)). An exception is Hall, Humphreys and Titterington (2002), who treat likelihood-based inference for Markovian models with missingness. As we shall see, in the case of generalized linear mixed models, rigorous asymptotics for variational approximation maximum likelihood estimation is delicate and involved. For this reason attention is restricted to a simple generalized linear mixed model setting which we call the *simple Poisson mixed model* and formally define in Section 2. In Poisson mixed models, the Gaussian variational approximation admits explicit forms that allow us to study its properties quite deeply. We show that the exact maximum likelihood estimators are well-defined. We then prove that their variational approximations are 'root-m' consistent, in the sense that their discrepancy from the true parameter values decreases at a rate proportional to the inverse square-root of the number of groups – denoted by m. However, this property requires the number of repeated measurements, n, to be at least as large as the square root of m. Without that condition the convergence rate is $O_p(n^{-1})$ rather than $O_p(m^{-1/2})$. Hence, consistency of Gaussian variational approximation requires that both the number of groups m and the number of repeated measures n be allowed to increase. While this excludes some longitudinal data analysis settings, such as matched paired designs, there are others where it is reasonable for n to grow. Ormerod and Wand (2009, 2010) show Gaussian variational approximation to be quite accurate for $n \simeq 5$. Our results also have something important to say in non-asymptotic cases, where n is small – Section 3.5 shows that Gaussian variational approximation can be inconsistent unless n, as well as m, is large.

The maximum likelihood problem and its Gaussian variational approximate solution are described in Section 2. Section 3 contains our theoretical results and accompanying commentary. All proofs are given in Section 4. We conclude with some discussion in Section 5.

2. Simple Poisson Mixed Model

We now describe the simple Poisson mixed model and Gaussian variational approximate maximum likelihood estimation of its parameters. The simple Poisson mixed model is a special case of the generalized linear mixed model where the fixed effects are a simple linear relationship, and the random effects correspond to a random intercept. The responses, conditional on the random effects, are assumed to be Poisson.

The observed data are (X_{ij}, Y_{ij}), $1 \leq i \leq m$, $1 \leq j \leq n$, where the Y_{ij}s are non-negative integers and the X_{ij}s are unrestricted real numbers. The simple Poisson mixed model is

$$Y_{ij}|X_{ij}, U_i \ \text{ independent Poisson with mean } \ \exp(\beta_0 + \beta_1 X_{ij} + U_i),$$
$$U_i \ \text{ independent } \ N(0, \sigma^2).$$

In biomedical applications the $1 \leq i \leq m$ corresponds to m patients, and $1 \leq j \leq n$ corresponds to n repeated measures on those patients, where typically $m \gg n$. The random intercepts U_i invoke a within-patient correlation. See, for example, McCulloch, Searle, and Neuhaus (2008) for details of this model and its longitudinal data analysis connections.

Let $\boldsymbol{\beta} \equiv (\beta_0, \beta_1)$ be the vector of fixed effects parameters. The conditional log-likelihood of $(\boldsymbol{\beta}, \sigma^2)$ is the logarithm of the joint probability mass function of the Y_{ij}'s, given the X_{ij}'s, as a function of the parameters. It admits the expression

$$\begin{aligned}\ell(\boldsymbol{\beta}, \sigma^2) = &\sum_{i=1}^{m}\sum_{j=1}^{n}\{Y_{ij}(\beta_0 + \beta_1 X_{ij}) - \log(Y_{ij}!)\} - \tfrac{m}{2}\log(2\pi\sigma^2) \\ &+ \sum_{i=1}^{m}\log\int_{-\infty}^{\infty}\exp\Big(\sum_{j=1}^{n}Y_{ij}u - e^{\beta_0+\beta_1 X_{ij}+u} - \frac{u^2}{2\sigma^2}\Big)\,du. \end{aligned} \tag{2.1}$$

The maximum likelihood estimates of $\boldsymbol{\beta}$ and σ^2 are then

$$(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2) = \underset{\boldsymbol{\beta}, \sigma^2}{\operatorname{argmax}}\, \ell(\boldsymbol{\beta}, \sigma^2).$$

In practice, computation of $(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2)$ and corresponding inference is hindered by the fact that the m integrals in (2.1) cannot be reduced. In this simple setting the integrals are univariate and quadrature can be entertained. However, in more elaborate grouped data generalized linear mixed models, such as those described in Ormerod and Wand (2010), the integrals are multidimensional and quadrature is more challenging.

Gaussian variational approximation offers a remedy since it results in a closed form approximation to $\ell(\boldsymbol{\beta}, \sigma^2)$. So-called *variational parameters* can be chosen to optimize the quality of the approximation. Let $\boldsymbol{u}$, $\boldsymbol{x}$, and $\boldsymbol{y}$, respectively, denote the random vectors containing the U_i's, the X_{ij}'s, and the Y_{ij}'s. Also, let p be the generic symbol for density or probability mass function. Then

$$\ell(\boldsymbol{\beta}, \sigma^2) = \log p(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\beta}, \sigma^2).$$

Hence, for arbitrary density functions q on $\mathbb{R}^m$,

$$\begin{aligned}\ell(\boldsymbol{\beta}, \sigma^2) &= \log p(\boldsymbol{y}|\boldsymbol{x})\int_{\mathbb{R}^m} q(\boldsymbol{u})\,d\boldsymbol{u} = \int_{\mathbb{R}^m}\log p(\boldsymbol{y}|\boldsymbol{x})q(\boldsymbol{u})\,d\boldsymbol{u} \\ &= \int_{\mathbb{R}^m}\log\left\{\frac{p(\boldsymbol{y},\boldsymbol{u}|\boldsymbol{x})/q(\boldsymbol{u})}{p(\boldsymbol{u}|\boldsymbol{y},\boldsymbol{x})/q(\boldsymbol{u})}\right\}q(\boldsymbol{u})\,d\boldsymbol{u} \\ &= \int_{\mathbb{R}^m} q(\boldsymbol{u})\log\left\{\frac{p(\boldsymbol{y},\boldsymbol{u}|\boldsymbol{x})}{q(\boldsymbol{u})}\right\}d\boldsymbol{u} + \int_{\mathbb{R}^m} q(\boldsymbol{u})\log\left\{\frac{q(\boldsymbol{u})}{p(\boldsymbol{u}|\boldsymbol{y},\boldsymbol{x})}\right\}d\boldsymbol{u}.\end{aligned}$$

The second term is the Kullback-Leibler distance between $q(\boldsymbol{u})$ and $p(\boldsymbol{u}|\boldsymbol{y},\boldsymbol{x})$. Since this is always non-negative (Kullback and Leibler (1951)) we get

$$\ell(\boldsymbol{\beta},\sigma^2) \geq \int_{\mathbb{R}^m} q(\boldsymbol{u}) \log\left\{\frac{p(\boldsymbol{y},\boldsymbol{u}|\boldsymbol{x})}{q(\boldsymbol{u})}\right\} d\boldsymbol{u}.$$

Now take q to be the m-variate Gaussian density function with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Lambda}$. This leads to

$$\ell(\boldsymbol{\beta},\sigma^2) \geq \underline{\ell}(\boldsymbol{\beta},\sigma^2,\boldsymbol{\mu},\boldsymbol{\Lambda}), \tag{2.2}$$

where

$$\begin{aligned}\underline{\ell}(\boldsymbol{\beta},\sigma^2,\boldsymbol{\mu},\boldsymbol{\Lambda}) \equiv & \sum_{i=1}^m \sum_{j=1}^n \{Y_{ij}(\beta_0+\beta_1 X_{ij}+\mu_i) - e^{\beta_0+\beta_1 X_{ij}+\mu_i+\lambda_i/2} - \log(Y_{ij}!)\} \\ & -\frac{m}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^m (\mu_i^2+\lambda_i) + \tfrac{1}{2}\log|\boldsymbol{\Lambda}| \end{aligned} \tag{2.3}$$

is the *Gaussian variational approximation* to $\ell(\boldsymbol{\beta},\sigma^2)$. Here $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_m)$ and $\boldsymbol{\lambda} = (\lambda_1,\ldots,\lambda_m)$ are the diagonal entries of $\boldsymbol{\Lambda}$. Since (2.2) holds for all choices of $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$, we obtain the tightest lower bound by maximizing over these *variational* parameters. Theorem 1 in Section 3.1 implies that the off-diagonal entries of $\boldsymbol{\Lambda}$ do not improve the lower bound, so there is no loss from working with

$$\begin{aligned}\underline{\ell}(\boldsymbol{\beta},\sigma^2,\boldsymbol{\mu},\boldsymbol{\lambda}) \equiv & \sum_{i=1}^m \sum_{j=1}^n \{Y_{ij}(\beta_0+\beta_1 X_{ij}+\mu_i) - e^{\beta_0+\beta_1 X_{ij}+\mu_i+\lambda_i/2} - \log(Y_{ij}!)\} \\ & -\frac{m}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^m (\mu_i^2+\lambda_i) + \tfrac{1}{2}\sum_{i=1}^m \log(\lambda_i). \end{aligned} \tag{2.4}$$

The Gaussian variational approximate maximum likelihood estimators are:

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

Note that maximization over $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ makes the lower bound as tight as possible, and hence improves the accuracy of the variational approximation.

3. Theoretical Results

In this section we provide several theoretical results concerned with the maximum likelihood problem presented in Section 2 and its approximate solution via Gaussian variational approximation. All proofs are deferred to Section 4.

3.1. Sufficiency of a diagonal covariance matrix

Theorem 1. *If $\mathbf{\Sigma}$ is a symmetric, positive definite matrix then, given the components down the main diagonal of $\mathbf{\Sigma}$, $|\mathbf{\Sigma}|$ is uniquely maximized by taking the off-diagonal components to vanish.*

Theorem 1 provides a justification for dropping the off-diagonal terms of $\mathbf{\Lambda}$ between (2.3) and (2.4). This means that the optimal q density factorizes into a product of m univariate normal densities. This result is in accordance with the fact that the integral over $\boldsymbol{u}$ in the exact log-likelihood (2.1) reduces to m univariate integrals.

3.2. Similarities between the log-likelihood and its lower bound

In this section we give formulae for the log-likelihood and its approximation. Assume that the X_{ij}'s and U_i's are totally independent, the X_{ij}'s are identically distributed as X, and the U_i's are all normal $N(0, \sigma^2)$. Also, for $1 \le i \le m$, let

$$Y_{i\bullet} \equiv \sum_{j=1}^{n} Y_{ij} \quad \text{and} \quad B_i = B_i(\beta_0, \beta_1) \equiv \sum_{j=1}^{n} \exp(\beta_0 + \beta_1 X_{ij}).$$

Then the log-likelihood and its approximation are:

$$\ell(\boldsymbol{\beta}, \sigma^2) = \ell_0(\boldsymbol{\beta}, \sigma^2) + \ell_1(\boldsymbol{\beta}, \sigma^2) + DATA\,,$$

and

$$\underline{\ell}(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\lambda}) = \ell_0(\boldsymbol{\beta}, \sigma^2) + \ell_2(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\lambda}) + DATA\,,$$

where $DATA$ denotes a quantity depending on the data alone, not on $\boldsymbol{\beta}$ or σ^2,

$$\ell_0(\boldsymbol{\beta}, \sigma^2) \equiv \sum_{i=1}^{m} \sum_{j=1}^{n} \{Y_{ij}\,(\beta_0 + \beta_1 X_{ij})\} - \tfrac{m}{2} \log \sigma^2\,,$$

$$\ell_1(\boldsymbol{\beta}, \sigma^2) \equiv \sum_{i=1}^{m} \log \left\{ \int_{-\infty}^{\infty} \exp\left(Y_{i\bullet}\, u - B_i\, e^u - \tfrac{1}{2}\, \sigma^{-2}\, u^2\right) du \right\} \qquad (3.1)$$

and

$$\ell_2(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\lambda}) \equiv \sum_{i=1}^{m} \left\{\mu_i\, Y_{i\bullet} - B_i\, \exp\left(\mu_i + \tfrac{1}{2}\, \lambda_i\right)\right\}$$
$$- \tfrac{1}{2\sigma^2} \textstyle\sum_{i=1}^{m} \left(\mu_i^2 + \lambda_i\right) + \tfrac{1}{2} \textstyle\sum_{i=1}^{m} \log \lambda_i. \qquad (3.2)$$

A first step is to find $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ to maximize $\ell_2(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\lambda})$. It is clear from the definition of ℓ_2, as a series in functions of (λ_i, μ_i), that if we keep β_0, β_1, and σ^2 fixed then the resulting μ_i will be a function of λ_i, and vice versa.

3.3. Properties of the variational parameters

Here we discuss relationships among the parameters that produce an extremum of $\ell_2(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\lambda}, \boldsymbol{\mu})$.

Theorem 2. *If* $\sigma^2 > 0$ *then:* (i) $\ell_2(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\lambda})$ *has a unique maximum in* $(\boldsymbol{\lambda}, \boldsymbol{\mu})$*;* (ii) *the maximum occurs when*

$$\mu_i = \sigma^2 \, Y_{i\bullet} + 1 - \sigma^2 \, \lambda_i^{-1} \,, \quad \textit{for} \quad 1 \le i \le m\,; \tag{3.3}$$

(iii) *at the maximum, the parameter values satisfy*

$$0 < \lambda_i < \sigma^2 \quad \textit{and} \quad \mu_i < \sigma^2 \, Y_{i\bullet}\,; \tag{3.4}$$

and (iv) μ_i *is defined uniquely, in terms of* B_i *and* $Y_{i\bullet}$*, by*

$$\sigma^2 \, B_i \, \exp\left\{\mu_i + \tfrac{1}{2}\, \sigma^2 \, (\sigma^2 \, Y_{i\bullet} + 1 - \mu_i)^{-1}\right\} = \sigma^2 \, Y_{i\bullet} - \mu_i\,. \tag{3.5}$$

It is worth noting that the values of the components (λ_i, μ_i) at which the maximum in $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ of $\ell_2(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\lambda})$ occurs, are determined index-by-index and do not require a more complex maximization. Of course, this is an immediate consequence of the diagonalization noted in Theorem 1.

3.4. "True Values" of parameters

In this section we derive the almost-sure limits of the values of β and σ^2 that maximize $\ell(\boldsymbol{\beta}, \sigma^2)$ and $\underline{\ell}(\boldsymbol{\beta}, \sigma^2)$. First, however, we derive the limits of $m^{-1}\,\ell_j$ for $j = 0, 1, 2$. For this purpose we impose the following conditions:

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

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

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

(A4) each Y_{ij}, conditional on both X_{ij} and U_i, is Poisson-distributed with mean $\exp(\beta_0^0 + \beta_1^0 \, X_{ij} + U_i)$, where β_0^0 and β_1^0 denote the true values of β_0 and β_1;

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

(A6) the moment generating function of X, $\phi(t) = E\{\exp(tX)\}$, is finite for $|t| < 2c$ for some $c > 0$, and that $|\beta_1^0| < c$.

Let $(B, Y_\bullet) = (B_1, Y_{1\bullet})$. Note that B is a function of β_0 and β_1, although $Y_\bullet$ is not. Define $Q = Q(\beta_0, \beta_1)$ to be the unique solution of the equation

$$\sigma^2 B \exp\left\{Q + \tfrac{1}{2}\sigma^2(\sigma^2 Y_\bullet + 1 - Q)^{-1}\right\} + Q - \sigma^2 Y_\bullet = 0\,. \tag{3.6}$$

The case $Q \geq$ is easly to handle. Hence forth, we assume the more difficult $Q < 0$ case. Let:

$$\ell_0^0(\boldsymbol{\beta}, \sigma^2) \equiv n \exp\left(\beta_0^0 + \tfrac{1}{2}(\sigma^2)^0\right)\left\{\beta_0\,\phi\big(\beta_1^0\big) + \beta_1\,\phi'\big(\beta_1^0\big)\right\} - \tfrac{1}{2}\log(\sigma^2), \tag{3.7}$$

$$\ell_1^0(\boldsymbol{\beta}, \sigma^2) \equiv E\left[\log\left\{\int_{-\infty}^{\infty} \exp\left(Y_\bullet u - B e^u - \tfrac{1}{2}\sigma^{-2}u^2\right) du\right\}\right], \tag{3.8}$$

and

$$\begin{aligned}\ell_2^0(\boldsymbol{\beta}, \sigma^2) \equiv\ & E(Q\,Y_\bullet) - \sigma^{-2} E(\sigma^2 Y_\bullet - Q) \\ & - \tfrac{1}{2\sigma^2} E\left\{Q^2 + \sigma^2(\sigma^2 Y_\bullet + 1 - Q)^{-1}\right\} \\ & + \tfrac{1}{2}\log(\sigma^2) - \tfrac{1}{2} E\{\log(\sigma^2 Y_\bullet + 1 - Q)\}\,. \end{aligned} \tag{3.9}$$

Note that the terms in $\log(\sigma^2)/2$, in both ℓ_0^0 and ℓ_2^0, cancel from $\ell_0^0 + \ell_2^0$.

Since $\phi(t) < \infty$ for $|t| < 2c$ then $E\{\exp(t\,|X|)\} < \infty$ for $0 < t < 2c$, and therefore $|\phi'(t)| \leq E\{|X| \exp(t\,|X|)\} < \infty$ for $|t| < 2c$. Therefore $\ell_0^0(\boldsymbol{\beta}, \sigma^2)$ is well-defined and finite provided that $|\beta_1^0| < 2c$ and $\sigma^2 > 0$. The theorem below implies that $\ell_2^0(\boldsymbol{\beta}, \sigma^2)$ is finite if $|\beta_1^0| < c$ and $\sigma^2 > 0$. Clearly, $\ell_1^0(\boldsymbol{\beta}, \sigma^2)$ is well-defined and finite whenever $\sigma^2 > 0$.

Theorem 3. *Assume conditions* (A1)$-$(A6). *Then $\ell_2^0(\boldsymbol{\beta}, \sigma^2)$ is well-defined and finite if $|\beta_1^0| < c$ and $\sigma^2 > 0$. Moreover, with probability* 1, *as $m \to \infty$ and for fixed n, we have $m^{-1}\,\ell_j(\boldsymbol{\beta}, \sigma^2) \to \ell_j^0(\boldsymbol{\beta}, \sigma^2)$ for $j = 0, 1$, and $m^{-1} \sup_{\boldsymbol{\lambda}, \boldsymbol{\mu}}$ $\ell_2(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\lambda}) \to \ell_2^0(\boldsymbol{\beta}, \sigma^2)$, uniformly in*

$$\beta_0 \in \left[\beta_0^{(1)}, \beta_0^{(2)}\right], \quad \beta_1 \in \left[\beta_1^{(1)}, \beta_1^{(2)}\right], \quad \sigma^2 \in \left[(\sigma^2)^{(1)}, (\sigma^2)^{(2)}\right], \tag{3.10}$$

provided that

$$-\infty < \beta_0^{(1)} < \beta_0^{(2)} < \infty, \quad -c < \beta_1^{(1)} < \beta_1^{(2)} < c, \quad 0 < (\sigma^2)^{(1)} < (\sigma^2)^{(2)} < \infty. \tag{3.11}$$

Recall from Section 3.2 that the log-likelihood ℓ, and its approximate form $\underline{\ell}$, satisfy $\ell = \ell_0 + \ell_1$ and $\underline{\ell} = \ell_0 + \ell_2$. Therefore, provided the maximizations are taken over values in a range permitted by (3.10) and (3.11), the almost sure limits of the estimators of β_0, β_1, and σ^2 that maximize ℓ and $\underline{\ell}$, are the values of the quantities that maximize $\ell_0^0(\boldsymbol{\beta}, \sigma^2) + \ell_1^0(\boldsymbol{\beta}, \sigma^2)$ and $\ell_0^0(\boldsymbol{\beta}, \sigma^2) + \ell_2^0(\boldsymbol{\beta}, \sigma^2)$, respectively.

By exploiting formulae (3.7)$-$(3.9) it is possible to choose the distribution of X such that (A1)–(A6) hold but the value of $(\boldsymbol{\beta}, \sigma^2)$ that maximizes $\ell_0^0 + \ell_1^0$ is different from that which maximizes $\ell_0^0+\ell_2^0$. The maximum likelihood estimators, based on maximizing ℓ, are consistent and converge at rate $m^{-1/2}$, even if n is held fixed. Therefore, in the present context and for fixed n, the Gaussian variational approximate estimators, based on maximizing $\underline{\ell}$, are inconsistent. However, as we show in Section 3.5, permitting m and n to diverge together leads to consistency, in fact at rate $m^{-1/2}+n^{-1}$.

3.5. Consistency at rate $m^{-1/2}+n^{-1}$

We are now in a position to state our main results on the consistency and convergence rates of estimators of the model parameters based on Gaussian variational approximation. Recall from Section 2 that the Gaussian variational approximate maximum likelihood estimators are

$$(\underline{\widehat{\beta}}_0, \underline{\widehat{\beta}}_1, \underline{\widehat{\sigma}}^2) = (\beta_0, \beta_1, \sigma^2) \text{ component of } \underset{\beta_0,\beta_1,\sigma^2,\boldsymbol{\mu},\boldsymbol{\lambda}}{\operatorname{argmax}} \ \underline{\ell}(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\lambda}).$$

We impose the following conditions:

(A7) the moment generating function of X, $\phi(t) = E\{\exp(tX)\}$, is well-defined on the whole real line;

(A8) the mapping that takes β to $\phi'(\beta)/\phi(\beta)$ is invertible;

(A9) in some neighbourhood of β_1^0 (the true value of β_1), $(d^2/d\beta^2)\log\phi(\beta)$ does not vanish;

(A10) for a constant $C>0$, $m = O(n^C)$ as m and n diverge;

(A11) the true values β_0^0, β_1^0, and $(\sigma^2)^0$ of β_0, β_1, and σ^2, respectively, lie in $(-\infty,\infty)$, $(-\infty,\infty)$, and $(0,\infty)$, respectively, and when choosing $(\underline{\widehat{\beta}}_0$, $\underline{\widehat{\beta}}_1$, $\underline{\widehat{\sigma}}^2)$ we search in the rectangular region $[-C_1, C_1]\times[-C_1, C_1]\times[C_1^{-1}, C_1]$, where C_1 is a constant satisfying $C_1 > \max(|\beta_0^0|, |\beta_1^0|, (\sigma^2)^0, 1/(\sigma^2)^0)$.

Theorem 4. *If* (A1)–(A6) *and* (A7)–(A11) *hold then, as m and n diverge,*

$$\underline{\widehat{\beta}}_0 = \beta_0^0 + O_p(m^{-1/2}+n^{-1}),\ \underline{\widehat{\beta}}_1 = \beta_1^0 + O_p(m^{-1/2}+n^{-1}),$$

and

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

4. Proofs

Theorem 1, which is proved in Section 4.1, reduces the parametric complexity of the variational problem from $O(m^2)$ to $O(m)$ in respect of the number of groups. The proof of Theorem 2 is then relatively conventional. That theorem is then applied to prove Theorem 3 by eliminating $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ from the variational likelihood. The proof of Theorem 4 is conducted in a sequence of three steps, each of which is essentially a lemma for the next. In particular, the first step (given in Section 4.4.1) establishes consistency of estimators of β_0, β_1, and σ^2; the second step (Section 4.4.2) uses the conclusion of Step 1 to control remainder terms, so that the consistency property can be extended to a rate of convergence that is almost, but not quite, as good as the rate stated in Theorem 4. Finally, in Section 4.4.3, the conclusion of Step 2 is used to give still better control of remainders, so that the full theorem can be derived.

4.1. Proof of Theorem 1

Let $\boldsymbol{\Sigma}$ be $p \times p$, and let $\boldsymbol{\Sigma}_1$ be the $(p-1)\times(p-1)$ matrix, $\boldsymbol{a}$ be the p-vector, and b be the scalar such that

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{a} \\ \boldsymbol{a}^{\mathrm{T}} & b \end{bmatrix}.$$

Then,

$$|\boldsymbol{\Sigma}| = |\boldsymbol{\Sigma}_1| \left(b - \boldsymbol{a}^{\mathrm{T}} \boldsymbol{\Sigma}_1^{-1} \boldsymbol{a} \right).$$

We prove the theorem by induction over p, and so we may assume that $|\boldsymbol{\Sigma}_1|$ is uniquely maximized by taking the off-diagonal components of $\boldsymbol{\Sigma}_1$ to vanish. (The theorem clearly holds when $p = 1$.) Since $\boldsymbol{\Sigma}_1$ is positive definite then, for $\boldsymbol{\Sigma}_1$ and b fixed, $b - \boldsymbol{a}^{\mathrm{T}}\boldsymbol{\Sigma}_1^{-1}\boldsymbol{a}$ is uniquely maximized by taking $\boldsymbol{a} = \mathbf{0}$, and then $|\boldsymbol{\Sigma}| = |\boldsymbol{\Sigma}_1|\, b$. The induction hypothesis now implies that $|\boldsymbol{\Sigma}|$ is uniquely maximized by taking the off-diagonal components of $\boldsymbol{\Sigma}$ to equal zero.

4.2. Proof of Theorem 2

Note that

$$\frac{\partial \ell_2(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\lambda})}{\partial \mu_i} = Y_{i\bullet} - B_i \exp\left(\mu_i + \tfrac{1}{2}\lambda_i\right) - \sigma^{-2}\mu_i\,, \tag{4.1}$$

$$\frac{\partial \ell_2(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\lambda})}{\partial \lambda_i} = -\tfrac{1}{2} B_i \exp\left(\mu_i + \tfrac{1}{2}\lambda_i\right) - \tfrac{1}{2}\sigma^{-2} + \tfrac{1}{2}\lambda_i^{-1}\,. \tag{4.2}$$

Equating both equations to zero to obtain a turning point in (λ_i, μ_i), and subtracting twice the second equation from the first, we see that (3.3) holds.

Using (3.3) to express λ_i in terms of μ_i, the right-hand sides (4.1) and (4.2), when set equal to zero, are both equivalent to (3.5). The left-hand side there increases from zero to $\sigma^2 B_i \exp\{\sigma^2(\sigma^2 Y_{i\bullet}+1)^{-1}/2\}$, and the right-hand side decreases from ∞ to 0, as μ_i increases from $-\infty$ to 0. Moreover, $\sigma^2 B_i \exp\{\sigma^2(\sigma^2 Y_{i\bullet}+1)^{-1}/2\} > 0$, provided that $\sigma^2 > 0$. Therefore, if $\sigma^2 > 0$ then (3.5) has a unique solution in μ_i, and so (i) holds.

The fact that the equation formed by setting the left-hand side of (4.1) equal to zero has a solution means that $\sigma^2 Y_{i\bullet} - \mu_i > 0$, which is the second part of (3.4). It therefore follows from (3.3), and the fact that $\lambda_i > 0$ since $\mathbf{\Lambda}$ must be positive definite, that $\lambda_i \in (0, \sigma^2)$, which is the first part of (3.4).

4.3. Proof of Theorem 3

First we establish the finiteness of $\ell_2^0(\boldsymbol{\beta}, \sigma^2)$. Assume that $\sigma^2 > 0$. Since $Q \le 0$ and $Y_\bullet \ge 0$ then $\sigma^2 Y_\bullet + 1 - Q \ge 1$, and so it suffices to ensure that $E(Y_\bullet) + E(|Q|\, Y_\bullet) + E(Q^2) < \infty$. Now, (3.6) implies that $\sigma^2 B \exp(-|Q|) \le \sigma^2 Y_\bullet - Q \le \sigma^2 B \exp(-|Q| + \sigma^2/2)$. Therefore,

$$(\sigma^2 Y_\bullet)^2 + 2\sigma^2 Y_\bullet |Q| + Q^2 = (\sigma^2 Y_\bullet - Q)^2 \le (\sigma^2 B)^2 \exp(-2|Q| + \sigma^2) \le (\sigma^2 B)^2 e^{\sigma^2},$$

and so,

$$E\big\{(\sigma^2 Y_\bullet)^2 + 2\sigma^2 Y_\bullet |Q| + Q^2\big\} \le \sigma^4 \exp(2\beta_0 + \sigma^4)\{n(n-1)\phi^2(\beta_1) + n\phi(2\beta_1)\} < \infty$$

provided $|\beta_1^0| < c$. Hence, the latter condition implies that, $E(Y_\bullet) + E(|Q|\, Y_\bullet) + E(Q^2) < \infty$, and therefore that $\ell_2^0(\boldsymbol{\beta}, \sigma^2) < \infty$.

Since

$$E(Y_{ij} \,|\, X_{ij}) = E\{\exp(\beta_0^0 + \beta_1^0 X_{ij} + U_i) \,|\, X_{ij}\} = \exp\big(\beta_0^0 + \beta_1^0 X_{ij} + \tfrac{1}{2}(\sigma^2)^0\big)$$

and $E\{X \exp(\beta_1 X)\} = \phi'(\beta_1)$, then

$$\begin{aligned} E\{Y_{ij}(\beta_0 + \beta_1 X_{ij})\} &= E\Big\{\exp\big(\beta_0^0 + \beta_1^0 X_{ij} + \tfrac{1}{2}(\sigma^2)^0\big)(\beta_0 + \beta_1 X_{ij})\Big\} \\ &= \exp\big(\beta_0^0 + \tfrac{1}{2}(\sigma^2)^0\big)\big\{\beta_0\, \phi(\beta_1^0) + \beta_1\, \phi'(\beta_1^0)\big\}. \end{aligned}$$

Therefore, if we take m to diverge to infinity and keep n fixed, then by the Law of Large Numbers, and with probability 1, $m^{-1} \ell_j(\boldsymbol{\beta}, \sigma^2) \to \ell_j^0(\boldsymbol{\beta}, \sigma^2)$ for $j = 0$, and analogously the result when $j = 1$ also holds. This establishes pointwise convergence. Uniform convergence follows from equicontinuity of the functions ℓ_0 and ℓ_1 if they are interpreted as indexed by different values of their random arguments. For example, in the case of ℓ_0 we have, for different versions $(\beta_0', \beta_1', (\sigma^2)')$ and $(\beta_0'', \beta_1'', (\sigma^2)'')$ of $(\beta_0, \beta_1, \sigma^2)$,

$$m^{-1} |\ell_0(\beta', (\sigma^2)') - \ell_0(\beta'', (\sigma^2)'')| \le (|\beta_0' - \beta_0''| + |\beta_1' - \beta_1''|)\, S + \tfrac{1}{2}\, |\log\{\frac{(\sigma^2)'}{(\sigma^2)''}\}|\,, \tag{4.3}$$

where $S = m^{-1} \sum_i \sum_j Y_{ij}(1 + |X_{ij}|)$ and converges almost surely to a finite limit as $m \to \infty$. (Here we have used the fact that $|\beta_1^0| < c$, where c is as in (A6).) It follows from (4.3) that the almost sure limit as $m \to \infty$, of the supremum of

$|\ell_0(\beta',(\sigma^2)')-\ell_0(\beta'',(\sigma^2)'')|$ over $|\beta_0'-\beta_0''|+|\beta_1'-\beta_1''|+|(\sigma^2)'-(\sigma^2)''|\le\epsilon$, where the parameter values are constrained to satisfy (3.10) and (3.11), converges to zero as $\epsilon\downarrow 0$. The case of ℓ_1 is similar.

To treat the convergence of $\ell_2(\boldsymbol{\beta},\sigma^2,\boldsymbol{\mu},\boldsymbol{\lambda})$ we note that, in view of Theorem 2 and particularly (3.5), with probability 1,

$$m^{-1}\sup_{\boldsymbol{\lambda},\boldsymbol{\mu}}\ell_2(\boldsymbol{\beta},\sigma^2,\boldsymbol{\mu},\boldsymbol{\lambda})\to E(Q\,Y_\bullet)-E\big\{B\,\exp\big(Q+\tfrac12\,R\big)\big\}$$
$$-\tfrac12\,\sigma^{-2}\,E\big(Q^2+R\big)+\tfrac12\,E(\log R)\,,$$

where the random variables B, Q, and R are jointly distributed such that $R=\sigma^2\,(\sigma^2\,Y_\bullet+1-Q)^{-1}$ and Q solves (3.6). In this notation, and with probability 1,

$$m^{-1}\sup_{\boldsymbol{\lambda},\boldsymbol{\mu}}\ell_2(\boldsymbol{\beta},\sigma^2,\boldsymbol{\mu},\boldsymbol{\lambda})\to E(Q\,Y_\bullet)-E\big[B\,\exp\big\{Q+\tfrac12\,\sigma^2\,(\sigma^2\,Y_\bullet+1-Q)^{-1}\big\}\big]$$
$$-\tfrac12\,\sigma^{-2}\,E\big\{Q^2+\sigma^2\,(\sigma^2\,Y_\bullet+1-Q)^{-1}\big\}$$
$$+\tfrac12\,\log(\sigma^2)-\tfrac12\,E\{\log(\sigma^2\,Y_\bullet+1-Q)\}\,.$$

That is equivalent to asserting that, with probability 1, $m^{-1}\sup_{\boldsymbol{\lambda},\boldsymbol{\mu}}\ell_2(\boldsymbol{\beta},\sigma^2,\boldsymbol{\mu},\boldsymbol{\lambda})\to\ell_2^0(\boldsymbol{\beta},\sigma^2)$. Again, uniform convergence follows via an equicontinuity argument.

4.4. Proof of Theorem 4

4.4.1. Consistency

For $1\le i\le m$, let $\underline{\widehat{\lambda}}_i$ and $\underline{\widehat{\mu}}_i$ denote the entries of $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ that maximize $\underline{\ell}(\boldsymbol{\beta},\sigma^2,\boldsymbol{\mu},\boldsymbol{\lambda})$. Equating the right-hand sides of (4.1) and (4.2) to zero, and dividing by n, we see that $\underline{\widehat{\beta}}_0$, $\underline{\widehat{\beta}}_1$, $\underline{\widehat{\sigma}}^2$, $\underline{\widehat{\lambda}}_i$, and $\underline{\widehat{\mu}}_i$ satisfy

$$\frac{\underline{\widehat{\mu}}_i}{(n\,\underline{\widehat{\sigma}}^2)}+\exp\big(\underline{\widehat{\mu}}_i+\tfrac12\,\underline{\widehat{\lambda}}_i+\underline{\widehat{\beta}}_0\big)\,\frac1n\sum_{j=1}^n\exp(\underline{\widehat{\beta}}_1\,X_{ij})-\frac1n\,Y_{i\bullet}=0 \tag{4.4}$$

$$\text{and}\quad\frac{1}{n\,\underline{\widehat{\sigma}}^2}+\exp\big(\underline{\widehat{\mu}}_i+\tfrac12\,\underline{\widehat{\lambda}}_i+\underline{\widehat{\beta}}_0\big)\,\frac1n\sum_{j=1}^n\exp(\underline{\widehat{\beta}}_1\,X_{ij})-\frac{1}{n\,\underline{\widehat{\lambda}}_i}=0. \tag{4.5}$$

Since X has a finite moment generating function, ϕ, on the real line, Markov's inequality can be used to prove that for all $C_1,C_2>0$ and $\rho\in(0,1/2)$,

$$\sup_{\beta_1:\,|\beta_1|\le C_1}P\bigg\{\bigg|\frac1n\sum_{j=1}^n\exp(\beta_1\,X_{ij})-\phi(\beta_1)\bigg|>n^{-\rho}\bigg\}=O\big(n^{-C_2}\big)\,.$$

Therefore, if $\mathcal{G}(n)$ is a grid of points in the interval $[-C_1,C_1]$ containing no more than $O(n^C)$ points, for some $C>0$, then

$$P\bigg\{\sup_{\beta_1\in\mathcal{G}(n)}\bigg|\frac1n\sum_{j=1}^n\exp(\beta_1\,X_{ij})-\phi(\beta_1)\bigg|>n^{-\rho}\bigg\}=O\big(n^{-C_2}\big)$$

for all $C_2 > 0$. Choosing C sufficiently large, and using Hölder continuity of the exponential function and of ϕ, we can extend this bound from $\mathcal{G}(n)$ to the whole of the interval:

$$P\left\{ \sup_{\beta_1 : |\beta_1| \le C_1} \left| \frac{1}{n} \sum_{j=1}^{n} \exp(\beta_1 X_{ij}) - \phi(\beta_1) \right| > n^{-\rho} \right\} = O\big(n^{-C_2}\big) .$$

Therefore, provided $m = O(n^C)$ for some $C > 0$,

$$P\left\{ \max_{1\le i\le m} \sup_{\beta_1 : |\beta_1| \le C_1} \left| \frac{1}{n} \sum_{j=1}^{n} \exp(\beta_1 X_{ij}) - \phi(\beta_1) \right| > n^{-\rho} \right\} = O\big(n^{-C_2}\big) \tag{4.6}$$

for all $C_1, C_2 > 0$. The probability statement in each of the three displays immediately above (4.6) could have been preceded by $\max_{1\le i\le m}$, although we chose not to do so because the distribution of X_{ij} does not depend on i. Nevertheless, the passage to (4.6) can be interpreted as moving the operator $\max_{1\le i\le m}$ from outside to inside the probability statement.

Recall that, conditional on X_{ij} and U_i, Y_{ij} is Poisson-distributed with mean $\exp(\beta_0^0 + \beta_1^0 X_{ij} + U_i)$. It therefore can be proved using Markov's inequality that for all $C_1, C_2 > 0$ and $\rho \in (0, 1/2)$,

$$\max_{1\le i\le m} P\left[\left| \frac{1}{n} \sum_{j=1}^{n} \left\{ Y_{ij} - \exp\left(\beta_0^0 + \beta_1^0 X_{ij} + U_i\right) \right\} \right| > n^{-\rho} \right] = O\big(n^{-C_2}\big) .$$

Hence, since $m = O(n^C)$ for some $C > 0$,

$$P\left[\max_{1\le i\le m} \left| \frac{1}{n} \sum_{j=1}^{n} \left\{ Y_{ij} - \exp\left(\beta_0^0 + \beta_1^0 X_{ij} + U_i\right) \right\} \right| > n^{-\rho} \right] = O\big(n^{-C_2}\big) . \tag{4.7}$$

The properties $m = O(n^C)$ and

$$P\left\{ \max_{1\le i\le m} U_i \ge \sigma\, (2D\, \log m)^{1/2} \right\} \le m^{1-D} \tag{4.8}$$

imply that there exists $C' > 0$ such that

$$P\left[\max_{1\le i\le m} \exp(U_i) > \exp\left\{ C' \,(\log n)^{1/2} \right\} \right] \to 0 . \tag{4.9}$$

Combining (4.6), (4.7), and (4.9) we see that for all $C_1, C_2 > 0$ and $\rho \in (0, 1/2)$,

$$P\left\{ \max_{1\le i\le m} \left| \frac{1}{n} Y_{i\bullet} - \exp\left(\beta_0^0 + U_i\right) \phi(\beta_1^0) \right| > n^{-\rho} \right\} = O\big(n^{-C_2}\big) . \tag{4.10}$$

From (4.6) and (4.10) we deduce that, for each $\rho \in (0, 1/2)$, $\widehat{\underline{\lambda}}_i$ and $\widehat{\underline{\mu}}_i$ satisfy

$$\frac{\widehat{\underline{\mu}}_i}{\sigma^2 n} + \exp\big(\widehat{\underline{\mu}}_i + \tfrac{1}{2}\,\widehat{\underline{\lambda}}_i + \beta_0\big)\,\big\{\phi(\beta_1) + O_p\big(n^{-\rho}\big)\big\}$$
$$= \exp(\beta_0^0 + U_i)\,\big\{\phi(\beta_1^0) + O_p\big(n^{-\rho}\big)\big\}\,, \tag{4.11}$$
$$\frac{1}{\sigma^2 n} + \exp\big(\widehat{\underline{\mu}}_i + \tfrac{1}{2}\,\widehat{\underline{\lambda}}_i + \beta_0\big)\,\big\{\phi(\beta_1) + O_p\big(n^{-\rho}\big)\big\} = \frac{1}{n\widehat{\underline{\lambda}}_i}\,, \tag{4.12}$$

where the $O_p(n^{-\rho})$ remainders are of the stated orders uniformly in $1 \le i \le m$ and in $\sigma^2, |\beta_0|, |\beta_1| \le C_1$, and $\sigma^2 \ge C_1^{-1}$ with $C_1 > 0$ arbitrary but fixed.

By (3.4), $0 < \lambda_i < \sigma^2$. Therefore, the left-hand side of (4.11) equals $\{\mu_i/(\sigma^2 n)\} + \exp(\mu_i + \omega_i)\,\{\phi(\beta_1) + O_p(n^{-\rho})\}$, where $\beta_0 - \sigma^2/2 \le \omega_i \le \beta_0 + \sigma^2/2$; call this result ($\mathrm{R}_1$). We confine σ^2 and β_0 to compact sets, in particular to values satisfying $\sigma^2, |\beta_0| \le C_1$, and $\sigma^2 \ge C_1^{-1}$, and so $|\omega_i|$ is bounded uniformly in i. This property, (R_1), and (4.11) imply that

$$\exp\big(\widehat{\underline{\mu}}_i + \tfrac{1}{2}\,\widehat{\underline{\lambda}}_i + \beta_0\big)\,\{\phi(\beta_1) + o_p(1)\} = \exp(\beta_0^0 + U_i)\,\{\phi(\beta_1^0) + o_p(1)\}\,, \tag{4.13}$$

uniformly in $1 \le i \le m$ and in σ^2, β_0 and β_1 satisfying $\sigma^2, |\beta_0|, |\beta_1| \le C_1$, and $\sigma^2 \ge C_1^{-1}$, for any fixed $C_1 > 1$. To establish (4.13) in detail, note first that for each $C_1 > 0$,

$$\phi(\beta_1) \text{ is bounded away from zero and infinity uniformly in } |\beta_1| \le C_1\,. \tag{4.14}$$

If (4.13) fails for some $\eta > 0$, then it can be shown from (4.11) that "for some i in the range $1 \le i \le m$, both $\mu_i < 0$ and $|\mu_i|/(\sigma^2 n) > \eta \exp(\mu_i + \omega_i)$, for all σ^2, β_0, and β_1 satisfying $\sigma^2, |\beta_0|, |\beta_1| \le C_1$, and $\sigma^2 \ge C_1^{-1}$," where the property within quotation marks holds with probability bounded away from zero along an infinite subsequence of values of n. In this case, since $|\omega_i|$ is bounded then $\mu_i < -\log n + O(\log\log n)$, and so, in view of (4.14), the left-hand side of (4.11), which can be nonnegative, is, for this $i = i(n)$, of order n^{-c} for some $c > 0$. Hence for this i, $\exp(\beta_0^0 + U_i) = O_p(n^{-c})$, and so the probability that the latter bound holds for some $1 \le i \le m$ must itself be bounded away from zero along an infinite subsequence of values of n diverging to infinity; call this result (R_2). Since the distribution of U_i is symmetric then (4.9) holds if, on the left-hand side we replace U_i by $-U_i$, and so

$$P\Big[\min_{1 \le i \le m} \exp(U_i) \ge \exp\big\{-C'\,(\log n)^{1/2}\big\}\Big] \to 1\,.$$

Since $\exp\{-C'\,(\log n)^{1/2}\}$ is of strictly larger order than n^{-c} for any $c > 0$, property (R_2) is violated, and so the original assumption that (4.13) fails must have been false.

Results (4.13) and (4.14) imply that $\widehat{\underline{\lambda}}_i$ and $\widehat{\underline{\mu}}_i$ satisfy

$$\widehat{\underline{\mu}}_i + \tfrac{1}{2}\,\widehat{\underline{\lambda}}_i + \beta_0 + \log\phi(\beta_1) = \beta_0^0 + U_i + \log\phi(\beta_1^0) + o_p(1)\,, \tag{4.15}$$
$$\exp\big(\widehat{\underline{\mu}}_i + \tfrac{1}{2}\,\widehat{\underline{\lambda}}_i + \beta_0\big)\,\phi(\beta_1) = \exp\big(\beta_0^0 + U_i\big)\,\phi(\beta_1^0)\,\{1 + o_p(1)\}\,, \tag{4.16}$$

uniformly in $1 \le i \le m$ and in σ^2, β_0, and β_1 satisfying $\sigma^2, |\beta_0|, |\beta_1| \le C_1$, and $\sigma^2 \ge C_1^{-1}$, for each fixed $C_1 > 0$. Substituting (4.16) into (4.12), and noting (4.14), we deduce that the $\widehat{\underline{\lambda}}_i$ satisfy

$$\frac{1}{\sigma^2 n} + \exp\big(\beta_0^0 + U_i\big)\,\phi(\beta_1^0)\,\{1 + o_p(1)\} = \frac{1}{n\widehat{\underline{\lambda}}_i}\,,$$

uniformly in $1 \le i \le m$ and in σ^2, β_0, and β_1 satisfying $\sigma^2, |\beta_0|, |\beta_1| \le C_1$, and $\sigma^2 \ge C_1^{-1}$. This result, (4.8), and the version of (4.8) with U_i replaced by $-U_i$ imply that

$$-\log(n\widehat{\underline{\lambda}}_i) = \beta_0^0 + U_i + o_p(1)\,, \tag{4.17}$$

uniformly in the same sense. In particular, $\sup_{1\le i\le m} \widehat{\underline{\lambda}}_i \to 0$ in probability. Hence, by (4.15),

$$\widehat{\underline{\mu}}_i = U_i + \beta_0^0 - \beta_0 + \log\{\frac{\phi(\beta_1^0)}{\phi(\beta_1)}\} + o_p(1)\,, \tag{4.18}$$

uniformly in $1 \le i \le m$ and in σ^2, β_0, and β_1 satisfying $\sigma^2, |\beta_0|, |\beta_1| \le C_1$, and $\sigma^2 \ge C_1^{-1}$.

The following property follows using the argument leading to (4.6):

$$P\bigg\{\max_{1\le i\le m}\ \sup_{\beta_1:\,|\beta_1|\le C_1}\ \bigg|\frac{1}{n}\sum_{j=1}^{n} X_{ij}\,\exp(\beta_1\,X_{ij}) - \phi'(\beta_1)\bigg| > n^{-\rho}\bigg\} = O\big(n^{-C_2}\big)\,; \tag{4.19}$$

the property below is a consequence of (4.10):

$$P\bigg\{\bigg|\frac{1}{mn}\sum_{i=1}^{m} Y_{i\bullet} - \exp\big(\beta_0^0 + \tfrac{1}{2}\,(\sigma^2)^0\big)\,\phi(\beta_1^0)\bigg| > n^{-\rho}\bigg\} = O\big(n^{-C_2}\big)\,; \tag{4.20}$$

and the following property can be derived analogously:

$$P\bigg\{\bigg|\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}\,Y_{ij} - \exp\big(\beta_0^0 + \tfrac{1}{2}\,(\sigma^2)^0\big)\,\phi'(\beta_1^0)\bigg| > n^{-\rho}\bigg\} = O\big(n^{-C_2}\big)\,. \tag{4.21}$$

Each of (4.19)–(4.21) holds for all $C_1, C_2 > 0$, and all $\rho \in (0, 1/2)$. Formula

(3.2) for $\ell_2(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\lambda})$ implies that

$$
\begin{aligned}
-\frac{1}{mn} \frac{\partial \ell_2(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\lambda})}{\partial \beta_0} &= \frac{1}{mn} \sum_{i=1}^{m} \exp\left(\beta_0 + \mu_i + \tfrac{1}{2}\lambda_i\right) \sum_{j=1}^{n} \exp(\beta_1 X_{ij}), \quad (4.22)\\
-\frac{1}{mn} \frac{\partial \ell_2(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\lambda})}{\partial \beta_1} &= \frac{1}{mn} \sum_{i=1}^{m} \exp\left(\beta_0 + \mu_i + \tfrac{1}{2}\lambda_i\right) \sum_{j=1}^{n} X_{ij} \exp(\beta_1 X_{ij})\\
&= \exp\left\{\beta_0^0 + \tfrac{1}{2}(\sigma^2)^0\right\} \phi(\beta_1^0) \phi(\beta_1)^{-1} \phi'(\beta_1) + o_p(1), \quad (4.23)
\end{aligned}
$$

where the second identity in (4.23) follows from (4.18) and (4.19). The second identity in (4.23) holds uniformly in values of λ_i and μ_i that solve (4.4) and (4.5), and for σ^2, β_0, and β_1 satisfying $\sigma^2, |\beta_0|, |\beta_1| \le C_1$, and $\sigma^2 \ge C_1^{-1}$. Also, by (3.1), (4.20), and (4.21),

$$
\begin{aligned}
\frac{1}{mn} \frac{\partial \ell_0(\boldsymbol{\beta}, \sigma^2)}{\partial \beta_0} &= \frac{1}{mn} \sum_{i=1}^{m} Y_{i\bullet} = \exp\left(\beta_0^0 + \tfrac{1}{2}(\sigma^2)^0\right) \phi(\beta_1^0) + o_p(1), \quad (4.24)\\
\frac{1}{mn} \frac{\partial \ell_0(\boldsymbol{\beta}, \sigma^2)}{\partial \beta_1} &= \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} X_{ij} Y_{ij}\\
&= \exp\left(\beta_0 + \tfrac{1}{2}(\sigma^2)_0\right) \phi'(\beta_1^0) \phi(\beta_1) \phi(\beta_1)^{-1} + o_p(1), \quad (4.25)
\end{aligned}
$$

uniformly in the same sense. Combining (4.23) and (4.25) we deduce that:

$$
\begin{aligned}
\frac{1}{mn} \frac{\partial \underline{\ell}(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\lambda})}{\partial \beta_1} &= \exp\{\beta_0^0 + \tfrac{1}{2}(\sigma^2)^0\} \phi(\beta_1)^{-1} \{\phi'(\beta_1^0)\phi(\beta_1) - \phi(\beta_1^0)\phi'(\beta_1)\}\\
&\quad + o_p(1), \quad (4.26)
\end{aligned}
$$

uniformly in the following sense:

> uniformly in σ^2, β_0, β_1, $\boldsymbol{\lambda}$, and $\boldsymbol{\mu}$ that solve $\partial\underline{\ell}/\partial\beta_0 = \partial\underline{\ell}/\partial\beta_1 = 0$, $\partial\underline{\ell}/\partial\boldsymbol{\lambda} = \partial\underline{\ell}/\partial\boldsymbol{\mu} = \mathbf{0}$, and which satisfy $\sigma^2, |\beta_0|, |\beta_1| \le C_1$, and $\sigma^2 \ge C_1^{-1}$, provided that C_1 is so large that $|\beta_0^0|$, $|\beta_1^0| < C_1$, and $C_1^{-1} < (\sigma^2)^0 < C_1$. (4.27)

Recall from (4.1), (4.22) and (4.24) that

$$
\begin{aligned}
\frac{\partial \ell_2(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\lambda})}{\partial \mu_i} &= Y_{i\bullet} - B_i \exp\left(\mu_i + \tfrac{1}{2}\lambda_i\right) - \sigma^{-2}\mu_i, \quad (4.28)\\
\frac{\partial \underline{\ell}(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\lambda})}{\partial \beta_0} &= \sum_{i=1}^{m} \sum_{j=1}^{n} Y_{ij} - \sum_{i=1}^{m} \exp\left(\beta_0 + \mu_i + \tfrac{1}{2}\lambda_i\right) \sum_{j=1}^{n} \exp(\beta_1 X_{ij})\\
&= \sum_{i=1}^{m} \left\{Y_{i\bullet} - B_i \exp\left(\mu_i + \tfrac{1}{2}\lambda_i\right)\right\}. \quad (4.29)
\end{aligned}
$$

Adding (4.28) over i, and subtracting from (4.29), we deduce that

$$\sum_{i=1}^{m} \widehat{\underline{\mu}}_i = 0\,. \tag{4.30}$$

Hence, by (4.18),

$$\beta_0^0 - \widehat{\underline{\beta}}_0 + \log\{\frac{\phi(\beta_1^0)}{\phi(\widehat{\underline{\beta}}_1)}\} = o_p(1)\,, \tag{4.31}$$

uniformly in the sense of (4.27).

Formula (4.26) (respectively, (4.31)) equates the value of $\phi'(\beta_1)/\phi(\beta_1)$ (respectively, $\exp(\beta_0)\,\phi(\beta_1)$) to its true value, plus a negligible remainder. Since, by assumption, there is a one-to-one relation between values of β and values of $\phi'(\beta)/\phi(\beta)$ (respectively, $\exp(\beta)$). Therefore consistency of estimation of β_0 and β_1 follows from (4.26) and (4.31).

Note too that

$$\begin{aligned}\frac{2}{m}\,\frac{\partial\underline{\ell}(\boldsymbol{\beta},\sigma^2,\boldsymbol{\mu},\boldsymbol{\lambda})}{\partial\sigma^2} &= \sigma^{-4}\,\frac{1}{m}\sum_{i=1}^{m}\big(\lambda_i+\mu_i^2\big)-\sigma^{-2}\\ &= \sigma^{-4}\,\frac{1}{m}\sum_{i=1}^{m}\big[U_i+\beta_0^0-\beta_0+\log\{\frac{\phi(\beta_1^0)}{\phi(\beta_1)}\}\big]^2-\sigma^{-2}+o_p(1)\,,\end{aligned} \tag{4.32}$$

where the first identity follows from (3.1) and (3.2), and the second comes from (4.18). Therefore if β_0 and β_1 solve the variational approximate likelihood equations then

$$\widehat{\underline{\sigma}}^2 = \frac{1}{m}\sum_{i=1}^{m} U_i^2 + o_p(1) = (\sigma^2)^0 + o_p(1)\,.$$

and $(\sigma^2)^0$ is estimated consistently.

4.4.2. Convergence rate for variational approximate estimators equals $O_p(m^{-1/2}+n^{-\rho})$

By (4.17),

$$\widehat{\underline{\lambda}}_i = n^{-1}\,\exp\big\{-\big(\beta_0^0+U_i\big)+o_p(1)\big\}\,, \tag{4.33}$$

where the remainder is of the stated order uniformly in $1 \le i \le m$. Also, by (4.8), the version of (4.8) for $-U_i$ rather than U_i, and (4.18),

$$\max_{1\le i\le m} |\widehat{\underline{\mu}}_i| = O_p\big(n^{\eta}\big)\,, \tag{4.34}$$

for each $\eta > 0$. Hence, by (4.11), (4.33) and the consistency of the estimator of $(\sigma^2)^0$, proved in Section 4.4.1,

$$\exp(\widehat{\underline{\mu}}_i+\widehat{\underline{\beta}}_0)\,\big\{\phi(\widehat{\underline{\beta}}_1)+O_p\big(n^{-\rho}\big)\big\} = \exp(\beta_0^0+U_i)\,\big\{\phi(\beta_1^0)+O_p\big(n^{-\rho}\big)\big\}\,, \tag{4.35}$$

uniformly in $1 \le i \le m$ and for each $\rho \in (0, 1/2)$.

Equation (4.35), together with (4.14), imply the following stronger form of (4.18):

$$\widehat{\underline{\mu}}_i = U_i + \beta_0^0 - \widehat{\underline{\beta}}_0 + \log\left\{\frac{\phi(\beta_1^0)}{\phi(\widehat{\underline{\beta}}_1)}\right\} + O_p\big(n^{-\rho}\big)\,, \tag{4.36}$$

uniformly in $1 \le i \le m$ and for each $\rho \in (0, 1/2)$.

The argument from (4.19) down, but with (4.18) replaced by (4.36), can now be used to prove that for each $\rho \in (0, 1/2)$,

$$\begin{aligned} &\widehat{\underline{\sigma}}^2 = (\sigma^2)^0 + O_p(m^{-1/2} + n^{-\rho})\,,\ \widehat{\underline{\beta}}_0 = \beta_0^0 + O_p(m^{-1/2} + n^{-\rho})\,,\\ &\widehat{\underline{\beta}}_1 = \beta_1^0 + O_p(m^{-1/2} + n^{-\rho})\,. \end{aligned} \tag{4.37}$$

The term in $m^{-1/2}$ derives from the standard deviations of means of functions of the U_is.

4.4.3. Concise convergence rate for variational approximate estimators

Without loss of generality, $\phi'(t) \neq 0$ in some compact neighbourhood of β_1^0. (If not, add a constant to the random variable X to ensure that $\phi'(\beta_1^0) \neq 0$.) We take β_1^0 to lie in that neighbourhood, and assume too that $\widehat{\underline{\beta}}_1$ is there. In view of the already-proved consistency of $\widehat{\underline{\beta}}_1$ for β_1^0, this assumption too can be made without loss of generality.

Using (4.6), (4.7), (4.14), (4.33), (4.34), and the consistency of $\widehat{\underline{\sigma}}^2$ for $(\sigma^2)^0$, it can be proved that

$$\frac{1}{n}\sum_{j=1}^{n} \exp(\beta_1\, X_{ij}) = \phi(\beta_1)\, \exp\{\Delta_{i1}(\beta_1)\}\,, \tag{4.38}$$

$$\frac{1}{n}\sum_{j=1}^{n} X_{ij}\, \exp(\beta_1\, X_{ij}) = \phi'(\beta_1)\, \exp\{\Delta_{i2}(\beta_1)\}\,, \tag{4.39}$$

$$\frac{1}{n}\, Y_{i\bullet} - \frac{\mu_i}{\sigma^2 n} = \exp\{U_i + \beta_0^0 + \Delta_{i3}(\beta_0, \beta_1)\}\, \phi(\beta_1^0)\,, \tag{4.40}$$

uniformly in i, and where Δ_{i1}, Δ_{i2} and Δ_{i3} satisfy, for all $C_1 > 0$ and each $\rho \in (0, 1/2)$

$$\max_{1\le i\le m}\ \sup_{|\beta_0|,|\beta_1|\le C_1} |\Delta_{ik}(\beta_0, \beta_1)| = O_p\big(n^{-\rho}\big)\,. \tag{4.41}$$

(When $k = 1$ or 2 the dependence of $\Delta_{ik}(\beta_0, \beta_1)$ on β_0 is degenerate. Note that, by (3.4), the left-hand side of (4.40) is strictly positive.) It can also be proved that

$$\max_{k=1,2,3}\max_{r=1,2}\sup_{|\beta_0|,|\beta_1|\le C_1}\frac{1}{m}\left|\sum_{i=1}^{m}\exp(U_i)\Delta_{ik}(\beta_0,\beta_1)^r\right| = O_p\big(m^{-1/2}+n^{-1}\big), \quad (4.42)$$

$$\max_{k=1,2,3}\max_{r_1=0,1}\max_{r_2=1,2}\sup_{|\beta_0|,|\beta_1|\le C_1}\frac{1}{m}\left|\sum_{i=1}^{m}U_i^{r_1}\Delta_{ik}(\beta_0,\beta_1)^{r_2}\right| = O_p\big(m^{-1/2}+n^{-1}\big). \quad (4.43)$$

In the notation of (4.38)–(4.40), (4.4) is equivalent to:

$$\exp\left\{\beta_0+\mu_i+\tfrac{1}{2}\lambda_i+\Delta_{i1}(\beta_1)\right\}\phi(\beta_1) = \exp\{\beta_0^0+U_i+\Delta_{i3}(\beta_0,\beta_1)\}\,\phi(\beta_1^0)\,. \quad (4.44)$$

Taking logarithms of both sides of (4.44), and adding over i, we deduce that if β_0 and β_1 satisfy that equation,

$$\beta_0-\beta_0^0+\log\left\{\frac{\phi(\beta_1)}{\phi(\beta_1^0)}\right\} = \frac{1}{m}\sum_{i=1}^{m}\Big\{U_i-\mu_i-\tfrac{1}{2}\lambda_i+\Delta_{i3}(\beta_0,\beta_1)-\Delta_{i1}(\beta_1)\Big\}\,. \quad (4.45)$$

Additionally, $m^{-1}\sum_i U_i = O_p(m^{-1/2})$; by (4.30), $\sum_i \underline{\widehat{\mu}}_i = 0$; using (4.33), $m^{-1}\sum_i \underline{\widehat{\lambda}}_i = O_p(n^{-1})$; and by (4.43),

$$\sup_{|\beta_0|,|\beta_1|\le C_1}\frac{1}{m}\left|\sum_{i=1}^{m}\{\Delta_{i3}(\beta_0,\beta_1)-\Delta_{i1}(\beta_1)\}\right| = O_p\big(m^{-1/2}+n^{-1}\big)\,.$$

Combining the results from (4.45) down we deduce that

$$\underline{\widehat{\beta}}_0-\beta_0^0+\log\left\{\frac{\phi(\underline{\widehat{\beta}}_1)}{\phi(\beta_1^0)}\right\} = O_p\big(m^{-1/2}+n^{-1}\big)\,. \quad (4.46)$$

Observe that

$$\Delta \equiv \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}X_{ij}\left\{Y_{ij}-\exp(\beta_0^0+\beta_1^0X_{ij}+U_i)\right\} = O_p\big(m^{-1/2}\big)\,. \quad (4.47)$$

Using (4.23) and (4.25) we deduce that the equation $\partial\underline{\ell}/\partial\beta_1=0$ is equivalent to:

$$\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}X_{ij}\left\{Y_{ij}-\exp\left(\beta_0+\mu_i+\tfrac{1}{2}\lambda_i+\beta_1X_{ij}\right)\right\} = 0\,,$$

which, in turn, is equivalent to

$$\Delta+\exp(\beta_0^0)\,\phi'(\beta_1^0)\,\frac{1}{m}\sum_{i=1}^{m}\exp\{U_i+\Delta_{i2}(\beta_1^0)\}$$
$$= \exp(\beta_0)\,\phi'(\beta_1)\,\frac{1}{m}\sum_{i=1}^{m}\exp\{\mu_i+\tfrac{1}{2}\lambda_i+\Delta_{i2}(\beta_1)\}\,. \quad (4.48)$$

But by (4.44),

$$\underline{\widehat{\mu}}_i + \tfrac{1}{2}\,\underline{\widehat{\lambda}}_i = \beta_0^0 - \underline{\widehat{\beta}}_0 + \log\left\{\frac{\phi(\beta_1^0)}{\phi(\underline{\widehat{\beta}}_1)}\right\} + U_i + \Delta_{i3}(\underline{\widehat{\beta}}_0, \underline{\widehat{\beta}}_1) - \Delta_{i1}(\underline{\widehat{\beta}}_1)\,, \qquad (4.49)$$

and therefore

$$\exp\left\{\mu_i + \tfrac{1}{2}\,\lambda_i + \Delta_{i2}(\underline{\widehat{\beta}}_0)\right\} = \exp\{\beta_0^0 - \underline{\widehat{\beta}}_0 + U_i + \Delta_{i3}(\underline{\widehat{\beta}}_0, \underline{\widehat{\beta}}_1) - \Delta_{i1}(\underline{\widehat{\beta}}_1) \\ + \Delta_{i2}(\underline{\widehat{\beta}}_1)\}\,\phi(\beta_1^0)/\phi(\underline{\widehat{\beta}}_1)\,. \qquad (4.50)$$

Combining (4.48) and (4.50) we deduce that

$$\Delta\,\exp(-\beta_0^0)\,\phi(\beta_1^0)^{-1} + \phi'(\beta_1^0)\,\phi(\beta_1^0)^{-1}\,\frac{1}{m}\sum_{i=1}^{m}\exp\{U_i + \Delta_{i2}(\beta_1^0)\}$$

$$= \phi'(\underline{\widehat{\beta}}_1)\,\phi(\underline{\widehat{\beta}}_1)^{-1}\,\frac{1}{m}\sum_{i=1}^{m}\exp\Big\{U_i + \Delta_{i3}(\underline{\widehat{\beta}}_0, \underline{\widehat{\beta}}_1) - \Delta_{i1}(\underline{\widehat{\beta}}_1) + \Delta_{i2}(\underline{\widehat{\beta}}_1)\Big\}. \quad (4.51)$$

Together, (4.37), (4.42), (4.43), (4.47), and (4.51) imply that

$$\phi'(\beta_1^0)\,\phi(\beta_1^0)^{-1}\,\frac{1}{m}\sum_{i=1}^{m}\exp(U_i)$$

$$= \phi'(\underline{\widehat{\beta}}_1)\,\phi(\underline{\widehat{\beta}}_1)^{-1}\,\frac{1}{m}\sum_{i=1}^{m}\exp(U_i) + O_p\big(m^{-1/2} + n^{-1}\big)\,. \qquad (4.52)$$

(To derive (4.52) we Taylor-expanded quantities $\exp(U_i+\Delta_i)$ as $\exp(U_i)\,(1+\Delta_i+\Delta_i^2/2)$, plus a remainder dominated by $(1/6)|\Delta_i|^3\exp(U_i + |\Delta_i|) = O_p(n^{\eta-(3/2)})$, uniformly in $1 \le i \le m$ for all $\eta > 0$; here we used (4.41).) Result (4.52) implies that

$$\phi'(\beta_1^0)\,\phi(\beta_1^0)^{-1} = \phi'(\underline{\widehat{\beta}}_1)\,\phi(\underline{\widehat{\beta}}_1)^{-1} + O_p\big(m^{-1/2} + n^{-1}\big)\,. \qquad (4.53)$$

Together, (4.46) and (4.53) imply that

$$\underline{\widehat{\beta}}_0 = \beta_0^0 + O_p\big(m^{-1/2} + n^{-1}\big)\,, \quad \underline{\widehat{\beta}}_1 = \beta_1^0 + O_p\big(m^{-1/2} + n^{-1}\big)\,. \qquad (4.54)$$

Results (4.33), (4.49), and (4.54) imply that

$$\underline{\widehat{\mu}}_i = U_i + \Delta_{i3}(\underline{\widehat{\beta}}_0, \underline{\widehat{\beta}}_1) - \Delta_{i1}(\underline{\widehat{\beta}}_1) + O_p\big(m^{-1/2} + n^{-1}\big)\,, \qquad (4.55)$$

uniformly in $1 \le i \le m$. Combining (4.43) and (4.55) we obtain

$$\frac{1}{m}\sum_{i=1}^{m}\mu_i^2 = \frac{1}{m}\sum_{i=1}^{m}U_i^2 + O_p\big(m^{-1/2} + n^{-1}\big) = (\sigma^2)^0 + O_p\big(m^{-1/2} + n^{-1}\big)\,. \quad (4.56)$$

From (4.32), (4.33) and (4.56) we deduce that $\underline{\widehat{\sigma}}^2 = (\sigma^2)^0 + O_p(m^{-1/2} + n^{-1})$. The theorem follows from this property and (4.54).

5. Discussion

The preceding two sections represent an important first step in understanding the theoretical properties of variational approximations in likelihood-based inference. The simple Poisson mixed model lends itself to a deep understanding of such properties since it is the one of the simplest generalized linear mixed model that still is complicated enough to benefit from approximation methods. Of course, there are several extensions that could be entertained: non-equal sample sizes within groups, multiple fixed effects, multiple variance components, more general covariance structures, and other generalized response families. The case of unequal sample sizes is straightforward to address provided we take the sizes to lie between two constant multiples of n; more disparate sample sizes lead to more complex results. General covariance structures and response families are more challenging to address, but a problem of arguably greater interest is that of finding practical approximations to the distributions of estimators. Methods for solving that problem are currently being developed. Asymptotic distribution theory is another extension of interest. The current article is likely to provide a basis for such future work.

Acknowledgements

This research was partially supported by Australian Research Council grants to University of Melbourne and University of Wollongong. The authors are grateful to Shen Wang for checking the proofs.

References

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

Hall, P., Humphreys, K. and Titterington, D. M. (2002). On the adequacy of variational lower bound functions for likelihood-based inference in Markovian models with missing values. *J. Roy. Statist. Soc. Ser. B* **64**, 549-564.

Humphreys, K. and Titterington, D. M. (2000). Approximate Bayesian inference for simple mixtures. In *Proceedings of Computational Statistics* 2000 (Edited by J. G. Bethlehem and P. G. M. van der Heijden), 331-336. Physica, Heidelberg.

Jordan, M. I. (2004). Graphical models. *Statist. Sci.* **19**, 140-155.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.* **22**, 79-86.

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

Minka, T., Winn, J. M., Guiver, J. P. and Kannan, A. (2009). Infer.NET 2.3. Microsoft Research Cambridge, Cambridge, UK.

Ormerod, J. T. and Wand, M. P. (2009). Comment on paper by Rue, Martino and Chopin. *J. Roy. Statist. Soc. Ser. B* **71**, 377-378.

Ormerod, J. T. and Wand, M. P. (2010). Gaussian variational approximate inference for generalized linear mixed models. Unpublished manuscript.

Titterington, D. M. (2004). Bayesian methods for neural networks and related models. *Statist. Sci.* **19**, 128-139.

Wang, B. and Titterington, D. M. (2006). Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model. *Bayesian Analysis* **1**, 625-650.

Department of Mathematics and Statistics, University of Melbourne, Melbourne 3000, Australia.

E-mail: halpstat@ms.unimelb.edu.au

School of Mathematics and Statistics, University of Sydney, Sydney 2006, Australia.

E-mail: jormerod@sydney.edu.au

School of Mathematics and Applied Statistics, University of Wollongong, Wollongong 2522, Australia.

E-mail: mwand@uow.edu.au

(Received May 2009; accepted September 2009)