Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

On the error in Laplace approximations of high-dimensional integrals

Helen Ogden

School of Mathematical Sciences, University of Southampton, UK

Correspondence
Helen Ogden, School of Mathematical Sciences, University of Southampton, S017 1BJ, UK.
Email: h.e.ogden@soton.ac.uk

Abstract

Laplace approximations are commonly used to approximate high-dimensional integrals in statistical applications, but the quality of such approximations as the dimension of the integral grows is not well understood. In this paper, we provide a new result on the size of the error in first- and higher-order Laplace approximations, in terms of the rate of growth of information about each of the integrated variables. By contrast with many existing results, we allow for variation in the rate of information growth among the different integrated variables. We apply our results to investigate the quality of Laplace approximations to the likelihood in some generalized linear mixed models.

KEYWORDS:
Asymptotic approximation; Intractable likelihood; Generalized linear mixed model; Latent Gaussian model; Marginal likelihood

1 | INTRODUCTION

Integrals of the form

$$L = \int_{\mathbb{R}^d} \exp\{-g(u)\}du \tag{1}$$

are frequently encountered in statistical applications, where $g(.)$ is a smooth function with a unique minimum. For example, the likelihood function for a generalized linear mixed model is of this form, where u is a vector of random effects. Integrals of this type are also common in Bayesian applications, for example as marginal likelihoods used for model comparison.

Laplace approximations are often used to approximate integrals of form (1). For instance, Laplace approximations are used to approximate the likelihood for a generalized linear mixed model in the `lme4` R package (Bates, Mächler, Bolker, & Walker 2015), and to approximate posterior moments or marginal likelihoods in Bayesian applications (Tierney & Kadane 1986). Laplace approximations are sometimes referred to as saddlepoint approximations: Goutis and Casella (1999) give a clear overview of the connection between saddlepoint and Laplace approximations, and Reid (1988) and Butler (2007) provide extensive reviews of statistical applications of saddlepoint approximations.

First-order Laplace approximations have been found to be highly accurate in many cases. For instance, Friel and Wyse (2012) compare various approximations to the marginal likelihood to choose between two possible logistic regression models for one particular dataset, and find that a first-order Laplace approximation is much more accurate than many more computationally expensive approximations in that case. However, it is difficult to generalise the results of numerical comparisons for specific cases to form broader conclusions about when we can expect a Laplace approximation to be accurate and when it may perform more poorly.

In this paper, we develop asymptotic results about the accuracy of Laplace approximations. While asymptotic results do not tell us the size of the error in the Laplace approximation in specific examples, they do provide guidance about the type of situations in which we might expect a

Laplace approximation to perform well. Formally, we consider sequences of integrals

$$L_n = \int_{\mathbb{R}^{d_n}} \exp\{-g_n(u)\}du, \qquad n \geq 1,$$

and consider the size of the error of Laplace approximations to L_n in the limit as $n \to \infty$. Typically $g(u)$ is a sum with one term for each observation, so n is the sample size. For ease of notation, we often suppress the dependence on n, and write the integral in form (1), where $L = L_n$, $g(.) = g_n(.)$ and $d = d_n$ are all implicitly allowed to vary with n. The case in which d remains fixed as $n \to \infty$ is well studied (see e.g. Small 2010), but relatively few results are available about the more general case.

Shun and McCullagh (1995) provide a formal expansion for integrals of type (1). By studying the size of various terms in this expansion, they find that the first-order Laplace approximation should be reliable if $d = o(n^{1/3})$, under the assumption that all derivatives of $g(.)$ grow at rate n. This result has been applied in a variety of contexts, such as by Rue, Martino, and Chopin (2009) in their discussion of the asymptotic error in the integrated nested Laplace approximation to the posterior distribution of parameters in a latent Gaussian model. From the result of Shun and McCullagh (1995), we might expect a Laplace approximation to perform well in all cases where the sample size is sufficiently large relative to the dimension of the integral. However, the assumption that all derivatives of $g(.)$ grow at rate n is not realistic in many situations for which Laplace approximations are commonly used, such as when approximating the likelihood for a generalized linear mixed model or other latent Gaussian models. We give an example in which $d = o(n^{1/3})$ but the error in the first-order Laplace approximation grows with n. There are situations in which the sample size is very large relative to the dimension of the integral, but for which the Laplace approximation will still perform poorly.

In Section 2, assuming alternative conditions on $g(.)$, we develop a new result on the error in Laplace approximations of various orders to integrals of type (1). Our result is motivated by a two-level random intercept model with n_j observations on items in the jth cluster, for which the likelihood factorizes into a product of terms

$$L = \prod_{j=1}^{d} \int_{-\infty}^{\infty} \exp\{-g_j(u_j)\}du_j,$$

where each $g_j(u_j)$ is a sum over n_j terms. In this case, we could use existing results on the error of Laplace approximations to one-dimensional integrals to show that the error in the first-order Laplace approximation to the integral is $O(\sum_{j=1}^{d} n_j^{-1})$. We show that a version of this result also holds more generally, and find similar expressions for the error in higher-order Laplace approximations. For our results to make sense, we first need to define what n_j means in the general case. Roughly, we may think of n_j as the rate of growth of information about u_j provided by the integrand. We measure the information about u via an integrand information matrix $[g^{(2)}]^{-1}$, the inverse of the Hessian matrix of $g(.)$ evaluated at its minimizer $\hat{u}$.

In Section 3, we apply these results to study the quality of Laplace approximations of the likelihood for some generalized linear mixed models, including a multilevel random intercept model with any number of levels of hierarchy.

2 | ERROR IN THE LOG-INTEGRAL APPROXIMATION

2.1 | A series expansion for the log-integral

Shun and McCullagh (1995) give a series expansion for the log-integral $\ell = \log L$. We use their expansion here, expressed with slightly different notation. We write

$$\ell = \tilde{\ell}_1 + \sum_{l=1}^{\infty} e_l, \tag{2}$$

where $\tilde{\ell}_1$ is the first-order Laplace approximation to the log-integral, and e_l are contributions to the error in this approximation of size decreasing with l, which we will shortly define.

The first-order Laplace approximation to the log-integral is

$$\tilde{\ell}_1 = -\frac{1}{2}\log\det(g^{(2)}) + \frac{d}{2}\log(2\pi) - g(\hat{u})$$

where $\hat{u} = \arg\min_{u\in\mathbb{R}^d}\{g(u)\}$ and $g^{(2)} = g''(\hat{u})$ is the matrix of second derivatives of $g(.)$ with respect to u, evaluated at $\hat{u}$.

Based on the decomposition (2), we may also define an order-k Laplace approximation to the log-integral, for $k \geq 2$, as

$$\tilde{\ell}_k = \tilde{\ell}_1 + \sum_{l=1}^{k-1} e_l.$$

What is meant by the order of a Laplace approximation is not standard across the literature: our definition is made by grouping together terms in a series expansion to the log-integral in terms of their asymptotic order. This is a different notion of order than that used by Raudenbush, Yang, and Yosef (2000), who group together terms according to the number of derivatives required to compute them.

In this paper, we study the errors in these Laplace approximations to the log-integral

$$\epsilon_k = \tilde{\ell}_k - \ell = -\sum_{l=k}^{\infty} e_l.$$

Shun and McCullagh (1995) give a series expansion for the log-integral in terms of particular bipartitions. For positive integers v and m, define the set of M-bipartitions $\mathcal{M}_{v,m}$ to be all (P, Q) such that $P = (p_1|\dots|p_v)$ and $Q = (q_1|\dots|q_m)$ are both partitions of $\{1, \dots, 2m\}$, such that each block of P contains at least three elements and each block of Q contains exactly two elements.

For each $(P, Q) \in \mathcal{M}_{v,m}$, define a corresponding graph $\mathcal{G}(P, Q)$ with vertices $1, \dots, 2m$, and an edge between each pair of vertices contained in the same block of either P or Q. If $\mathcal{G}_{P,Q}$ is a connected graph, say that (P, Q) is a connected bipartition, and write $(P, Q) \in \mathcal{M}^C_{v,m}$. We define the level of $(P, Q) \in \mathcal{M}_{v,m}$ to be $l = m - v$, and write $\mathcal{M}^C_l$ for all connected level-l M-bipartitions.

For a vector of indices I, write $g_I(u) = \nabla_{u_I} g(u)$ for the partial derivative of g with respect to the coordinates u_I, and let $g_I = g_I(\hat{u})$. Let $g^{(k)}$ be the k-dimensional array with entries $g^{(k)}_{j_1,\dots,j_k} = g_{j_1,\dots,j_k}$, and write $g^{jk} = \left(g^{(2)}\right)^{-1}_{jk}$. Then define

$$e_{P,Q} = \frac{(-1)^v}{(2m)!} \sum_{j \in [1:d]^{2m}} g_{j_{p_1}} \cdots g_{j_{p_v}} g^{j_{q_1}} \cdots g^{j_{q_m}}, \tag{3}$$

where $[1:d]^{2m} = \{(j_1, \dots, j_{2m}) : j_l \in \{1, \dots, d\}\}$, and j_p is the sub-vector of $j = (j_1, \dots, j_{2m})$ corresponding to the indices in p.

We may write the level-l contribution to the log-integral e_l as a sum of contributions from each connected level-l M-bipartition, as

$$e_l = \sum_{(P,Q) \in \mathcal{M}^C_l} e_{P,Q}. \tag{4}$$

To demonstrate these definitions, we find the level-1 contribution e_1, used in the second-order Laplace approximation. There are three types of bipartitions in $\mathcal{M}^C_1$: (P_1, Q_1), where $P_1 = (1\,2\,3\,4)$ and $Q_1 = (1\,2\,|\,3\,4)$; (P_2, Q_2), where $P_2 = (1\,2\,3\,|\,4\,5\,6)$ and $Q_2 = (1\,2\,|\,3\,4\,|\,5\,6)$; and (P_3, Q_3) where $P_3 = P_2$ and $Q_3 = (1\,4\,|\,2\,5\,|\,3\,6)$. While there are other bipartitions in $\mathcal{M}^C_1$, they are all similar to one of these three, in that they may be obtained by rearranging the labels $\{1, \dots, 2m\}$, and so give the same contribution $e_{P,Q}$. For example, the bipartition $P^*_1 = (1\,2\,3\,4)$, $Q^*_1 = (1\,3\,|\,2\,4)$ may be obtained from (P_1, Q_1) by exchanging 2 and 3, and $e_{P^*_1,Q^*_1} = e_{P_1,Q_1}$. From (3), we have

$$\begin{aligned}
e_{P_1,Q_1} &= -\frac{1}{4!} \sum_{j_1,\dots,j_4} g_{j_1j_2j_3j_4} g^{j_1j_2} g^{j_3j_4} \\
e_{P_2,Q_2} &= \frac{1}{6!} \sum_{j_1,\dots,j_6} g_{j_1j_2j_3} g_{j_4j_5j_6} g^{j_1j_2} g^{j_3j_4} g^{j_5j_6} \\
e_{P_3,Q_3} &= \frac{1}{6!} \sum_{j_1,\dots,j_6} g_{j_1j_2j_3} g_{j_4j_5j_6} g^{j_1j_4} g^{j_2j_5} g^{j_3j_6}.
\end{aligned} \tag{5}$$

McCullagh (1987) lists 4 bipartitions similar to (P_1, Q_1), 9 similar to (P_2, Q_2) and 6 similar to (P_3, Q_3), so the level-1 contribution is $e_1 = 3e_{P_1,Q_1} + 9e_{P_2,Q_2} + 6e_{P_3,Q_3}$, and the second-order Laplace approximation to the log-likelihood is $\tilde{\ell}_2 = \tilde{\ell}_1 + e_1$.

There may be more efficient ways to compute e_1 than direct computation of the sums in (5). For example, Zipunnikov and Booth (2011) describe a more efficient method for computing these terms for a generalized linear mixed model.

2.2 | Error in log-integral approximations

In order to establish our main result, we assume some conditions on $g(.)$ in (1). To describe these conditions, we require some notation. Write $a = \Theta(b)$ if $a = O(b)$ and $a^{-1} = O(b^{-1})$, so a grows at the same rate as b. For a random variable A, write $A = \Theta_p(b)$ if $A = O_p(b)$ and $A^{-1} = O_p(b^{-1})$, so that A grows at rate b in probability.

We use a particular notion of a random array being order 1 in probability. Suppose A is a k-dimensional array, with entries $A_{j_1,\dots,j_k}$ for each $j_i \in \{1, \dots, d\}$. If $k = 1$, say $A = O^*_p(1)$ if $A_j = O_p(1)$ for each $j = 1, \dots, d$. If $k \geq 2$, let

$$A^i_j = \sum_{j_1=1}^{d} \cdots \sum_{j_{i-1}=1}^{d} \sum_{j_{i+1}=1}^{d} \cdots \sum_{j_k=1}^{d} |A_{j_1,\dots,j_{i-1},j,j_{i+1},\dots,j_k}|,$$

and say $A = O^*_p(1)$ if $A^i_j = O_p(1)$ for each $i = 1, \dots, k$ and $j = 1, \dots, d$. In the simple case where A is a diagonal array, $A = O^*_p(1)$ if the diagonal entries $A_{j,j,\dots,j} = O_p(1)$.

For a given choice of normalizing terms $n_1, \dots, n_d$, and for each vector of indices I, define the normalized derivatives

$$f_I = g_I \prod_{j \in I} n_j^{-1/|I|},$$

and write $f^{(k)}$ for the k-dimensional array with entries $f^{(k)}_{j_1,\dots,j_k} = f_{j_1,\dots,j_k}$. We write $f^{jk} = [(f^{(2)})^{-1}]_{jk}$.

Having established this notation, we may now state the conditions required for our main result.

Condition 1. $g(.)$ is a smooth function with a unique minimum.

Condition 2. There is some choice of normalizing terms $n_1, \dots, n_d$ such that the normalized derivative arrays $f^{(k)}$ satisfy $f^{(k)} = O_p^*(1)$ for all $k \geq 3$, and $[f^{(2)}]^{-1} = O_p^*(1)$.

The normalizing terms are often chosen so that $g_{jj} = \Theta_p(n_j)$, and we may think of n_j as an effective sample size for u_j. We state here our main result, which is proved in Appendix A.

Theorem 1. Suppose L is of form (1), where $g(.)$ satisfies Conditions 1 and 2, for some choice of normalizing terms $n_1, \dots, n_d$. Then the error in the order-k Laplace approximation to $\log L$ is $\epsilon_k = O_p(\sum_{j=1}^d n_j^{-k})$.

Laplace approximations are invariant to linear reparameterizations. That is, if $v = Au$, where A is an invertible $d \times d$ matrix, then writing $g_A(v) = g(A^{-1}v) + \log\det(A)$, and

$$L^{(A)} = \int_{\mathbb{R}^d} \exp\{-g_A(v)\}dv,$$

we have $L^{(A)} = L$, and the order-k Laplace approximation of L is unchanged by the reparameterization, so that $\tilde{L}_k^{(A)} = \tilde{L}_k$. In many situations, Condition 2 does not hold in the original parameterization, but does hold after making a suitable linear reparameterization, so we may still apply Theorem 1. We give an example of this in Section 3.3.

3 | APPLICATION TO LIKELIHOOD APPROXIMATION FOR GENERALIZED LINEAR MIXED MODELS

3.1 | Generalized linear mixed models

In a generalized linear mixed model, the distribution of the response $Y = (Y_1, \dots, Y_n)$ is determined by a linear predictor $\eta = (\eta_1, \dots, \eta_n)$. Conditional on η, the components Y_i of the response are independent, with known density function $f(y_i|\eta_i)$. We assume an exponential family with canonical link, so that

$$\log f(y_i|\eta_i) = \frac{y_i\eta_i - b(\eta_i)}{a_i(\phi)},$$

where $b(.)$ is a smooth and convex function, $a_i(\phi) > 0$, and ϕ is the dispersion parameter, which we assume here to be known. The linear predictor is modelled as $\eta = X\beta + Zu$, where $X \in \mathbb{R}^{n\times p}$ and $Z \in \mathbb{R}^{n\times d}$ are design matrices, $\beta \in \mathbb{R}^p$ is a vector of fixed effects, and $u \in \mathbb{R}^d$ is a vector of random effects. We assume that $u \sim N_d(0, \Sigma(\psi))$, where $\psi \in \mathbb{R}^q$ is an unknown parameter, and write $\theta = (\beta, \psi)$ for the full vector of unknown parameters.

The likelihood for this model is

$$L(\theta) = \int_{\mathbb{R}^d} \exp\{-g(u;\theta)\}du, \tag{6}$$

where

$$g(u;\theta) = h(u;\beta) - \log\phi_d(u; 0, \Sigma(\psi)), \tag{7}$$

$$h(u;\beta) = \sum_{i=1}^n -\log f(y_i|\eta_i = {X_i}^T\beta + {Z_i}^T u) = \sum_{i=1}^n \frac{b({X_i}^T\beta + {Z_i}^T u) - y_i({X_i}^T\beta + {Z_i}^T u)}{a_i(\phi)} \tag{8}$$

and $\phi_d(.;\mu,\Sigma)$ is the $N_d(\mu,\Sigma)$ density function. The d-dimensional integral in (6) is typically intractable, except in the special case of a linear mixed model where $Y_i|\eta_i$ are normally distributed. Because of this intractability, it is common to use some numerical approximation $\tilde{L}(\theta)$ to the likelihood, and first-order Laplace approximation is often used. For example, by default the `lme4` R package (Bates et al. 2015) uses a first-order Laplace approximation to the likelihood for inference, and the integrated nested Laplace approximations of Rue et al. (2009) is a Bayesian approach based on a Laplace approximation to the likelihood.

In order to apply Theorem 1 to the likelihood of a generalized linear mixed model, we will first have to show that $g(.)$ as defined in (7) satisfies Conditions 1 and 2. We drop θ from the notation, so that (6) is of form (1).

We can show Condition 1 holds in all cases. The proof is in Appendix B.

Proposition 1. Let $g(u)$ be as defined in (7), where Σ is a positive definite matrix. Then $g(.)$ satisfies Condition 1.

We need to show that Condition 2 holds on a case-by-case basis. In our examples, we choose the normalizing term n_j to be the number of observations which involve u_j.

3.2 | A two-level random intercept model

We consider a two-level random intercept model, which is a special case of the generalized linear mixed model of Section 3.1 in which each observation i is contained in a cluster c(i). Observations in the same cluster j are correlated by a shared random effect u_j. The linear predictor is $\eta_i = x_i^T\beta + u_{c(i)}$ $(i = 1, \dots, n)$, where we suppose the u_j are independent $N(0, \sigma^2)$ random variables. In the notation of Section 3.1, we have $Z_{i,c(i)} = 1$, and $Z_{i,j} = 0$ if $j \neq c(i)$ and $\Sigma = \sigma^2 I$, where I is an identity matrix.

In this special case, the likelihood (6) simplifies into a product of one-dimensional integrals

$$L(\theta) = \prod_{j=1}^{d} \int \prod_{i:c(i)=j} f(y_i|\eta_i = {x_i}^T\beta + u_j)\phi(u_j; 0, \sigma^2)du_j.$$

The log-likelihood may be written as a sum

$$\ell(\theta) = \sum_{j=1}^{d} \log \int \prod_{i:c(i)=j} f(y_i|\eta_i = {x_i}^T\beta + u_j)\phi(u_j; 0, \sigma^2)du_j, \tag{9}$$

so ϵ_k is a sum of separate error terms.

Proposition 2. Suppose we have a two-level random intercept model, with n_j observations on cluster j, for $j = 1, \dots d$. The error in the order-k Laplace approximation to the log-likelihood is

$$\epsilon_k(\theta) = \tilde{\ell}_k(\theta) - \ell(\theta) = O_p\Big(\sum_{j=1}^{d} n_j^{-k}\Big).$$

Proof. The derivative arrays $g^{(k)}$ are diagonal for all k, with diagonal entries $g_{j\dots j} = \Theta_p(n_j)$, so Condition 2 holds with normalizing terms $n_1, \dots, n_d$. Theorem 1 gives that $\epsilon_k(\theta) = O_p(\sum_{j=1}^{d} n_j^{-k})$, as required. □

In the balanced case, where all $n_j = nd^{-1}$, $\epsilon_k = O_p(d^{k+1}n^{-k})$. This tends to zero as $n \to \infty$ if $d = o(n^{k/(k+1)})$. The error in the first-order Laplace approximation tends to zero if $d = o(n^{1/2})$.

In an unbalanced case, the result can be quite different. As an extreme example, suppose

$$n_j = \begin{cases} \log d & \text{if } j = 1, \dots, d-1 \\ n - (d-1)\log d & \text{if } j = d, \end{cases}$$

where $n > d \log d$. Then

$$\epsilon_1 = O_p\big((d-1)(\log d)^{-1} + (n - (d-1)\log d)^{-1}\big) = O_p\big(d(\log d)^{-1}\big).$$

This upper bound on the error in the first-order Laplace approximation tends to infinity as $d \to \infty$, no matter how large n is relative to d. For example, if $n = d^4$, then $d = o(n^{1/3})$, but $d(\log d)^{-1} \to \infty$.

In Section 3.4, we investigate the error ϵ_1 numerically. We find that the upper bound is met, so that ϵ_1 scales as $\sum_{j=1}^{d} n_j^{-1}$. In the unbalanced case just described, this means that $\epsilon_1 \to \infty$, no matter how large n is relative to d.

We have shown that a large sample size relative to the dimension of the integral is no guarantee that a Laplace approximation will be accurate. Instead, Theorem 1 tells us that we need to consider the amount of information provided by the integrand about each u_j. This finding is important in practice because in many cases where Laplace approximations are used, such as in likelihood approximation for generalised linear mixed models and many other latent Gaussian models, the amount of information about each u_j remains small even when the overall sample size, n, is large.

3.3 | A multilevel random intercept model

Suppose that each observation i is contained in a level-2 cluster $c_2(i)$, and that each level-2 cluster j is itself contained within a hierarchy of higher-level clusters, $c_l(j)$, $j = 3, \dots, L$. The clusters are nested within one another, so that if $c_l(j) = c_l(k)$, then $c_{l+1}(j) = c_{l+1}(k)$. The linear predictor is

$$\eta_i = {x_i}^T\beta + u^{(2)}_{c_2(i)} + \sum_{l=3}^{L} u^{(l)}_{c_l(c_2(i))} \quad (i = 1, \dots, n),$$

where we assume $u_j^{(l)} \sim N(0, \sigma_l^2)$, $l = 2, \dots, L$, with all the u_j independent. Suppose that there are d level-2 clusters in total, and d_l level-l clusters, for each $l = 3, \dots, L$. It is no longer possible to write the log-likelihood as a sum of one-dimension log-integrals as in (9). Since an accurate approximation to the exact log-likelihood is no longer readily available, it is important to understand the quality of the Laplace approximation in this case.

Condition 2 does not hold for this parameterization, so we define a new parameterization of the model. Let $v_j = u_j^{(2)} + \sum_{l=3}^{L} u_{c_l(j)}^{(l)}$ for $j = 1, \ldots, d$. We have $\eta_i = x_i^T \beta + v_{c_2(i)}$, where there are now a total of d random effects, rather than $d + d_3 + \ldots + d_L$ in the original parameterization. We have reduced the structure to the two-level random intercept model of Section 3.2, except now $v \sim N_d(0, \Sigma)$, where

$$\Sigma_{jk} = \begin{cases} \sigma_2^2 + \sigma_3^2 + \ldots + \sigma_L^2 & \text{if } j = k \\ \sigma_3^2 + \ldots + \sigma_L^2 & \text{if } j \neq k, \text{ but } c_3(j) = c_3(k) \\ \vdots & \vdots \\ \sigma_l^2 + \ldots + \sigma_L^2 & \text{if } c_{l-1}(j) \neq c_{l-1}(k), \text{ but } c_l(j) = c_l(k) \\ \vdots & \vdots \\ \sigma_L^2 & \text{if } c_{L-1}(j) \neq c_{L-1}(k), \text{ but } c_L(j) = c_L(k) \\ 0 & \text{if } c_L(j) \neq c_L(k). \end{cases}$$

Proposition 3. Suppose we have an L-level random intercept model with independent random effects, with n_j observations in level-2 cluster j, for $j = 1, \ldots d$. The error in the order-k Laplace approximation to the log-likelihood is

$$\epsilon_k(\theta) = \tilde{\ell}_k(\theta) - \ell(\theta) = O_p\Big(\sum_{j=1}^{d} n_j^{-k}\Big).$$

The proof is in Appendix B. The asymptotic order of the error in a Laplace approximation to the log-likelihood depends on the number of observations in each of the level-2 clusters, but not on how these level-2 clusters are grouped into higher-level clusters.

3.4 | Numerical demonstration

We simulate from a simple two-level random intercept model

$$Y_i | u_{c(i)} \sim \text{Bernoulli}(p_{c(i)}), \quad \text{logit}(p_j) = u_j, \quad u_j \sim N(0, \sigma^2), \quad j = 1, \ldots, d.$$

Since the error in an approximation to the loglikelihood may be rewritten as a sum over clusters, it is sufficient to study the error in the log-likelihood for a fixed d, allowing the number of observations n_j on each cluster to vary. For this experiment, we will fix $\sigma = 1$ and $d = 1000$. For a range of values of n_j, we simulate a single dataset from the model, and calculate the error ϵ_1 in the first-order Laplace approximation to the loglikelihood. We choose a relatively large d in order to make the random error due to variations in the simulated data sets negligible. Figure 1 shows $\log|\epsilon_1|$ plotted against $\log(n_j)$. We see that ϵ_1 scales approximately as n_j^{-1}, meeting the upper bound derived theoretically in Proposition 2.

3.5 | Impact on approximate likelihood inference

When an approximate likelihood $\tilde{L}(\theta)$ is used for inference, the impact of the error in the likelihood approximation on the resulting inference is of more interest than the size of that error itself. If the error in the log-likelihood $\epsilon(\theta) = \log \tilde{L}(\theta) - \log L(\theta)$ tends to zero in probability, uniformly in θ, Douc, Moulines, and Rydén (2004) show that the approximate likelihood estimator $\tilde{\theta}$ will be fully efficient, and have the same first-order asymptotic distribution as the maximum likelihood estimator. In our examples, if

$$\sum_{j=1}^{d} n_j^{-k} \to 0 \text{ as } d \to \infty \tag{10}$$

we expect the order-k Laplace estimator to be fully efficient. In order to make the argument rigorous, we would need to show that the supremum of the error in log-likelihood in some region around the true parameter value tends to zero.

However, condition (10) is likely to be stronger than necessary for a order-k Laplace estimator to be fully efficient. Ogden (2017) gives conditions on the size of the error in the score function $\nabla_\theta \epsilon(\theta)$ which ensure that inference with an approximate likelihood retains the same first-order properties as inference with the exact likelihood. By studying this error in score, it should be possible to show that the order-k Laplace estimator is fully efficient under a weaker condition than (10). Some modification of the results of Ogden (2017) would be required before they could be used in this case, as information on different components of the parameter vector may grow at different rates (Nie 2007).

While our asymptotic results provide guidance about the type of situations in which we might expect a Laplace approximation to perform well, they do not directly tell us whether inference obtained by using a first-order Laplace approximation will be reliable for a given model and dataset. In cases where it is feasible to compute the second-order Laplace approximation, one possibility would be to compare the inference obtained with the first- and second-order approximations, deeming the inference reliable if the two sets of conclusions match one another closely. The ability of this approach to detect situations in which the first-order Laplace approximation fails is worthy of future investigation.

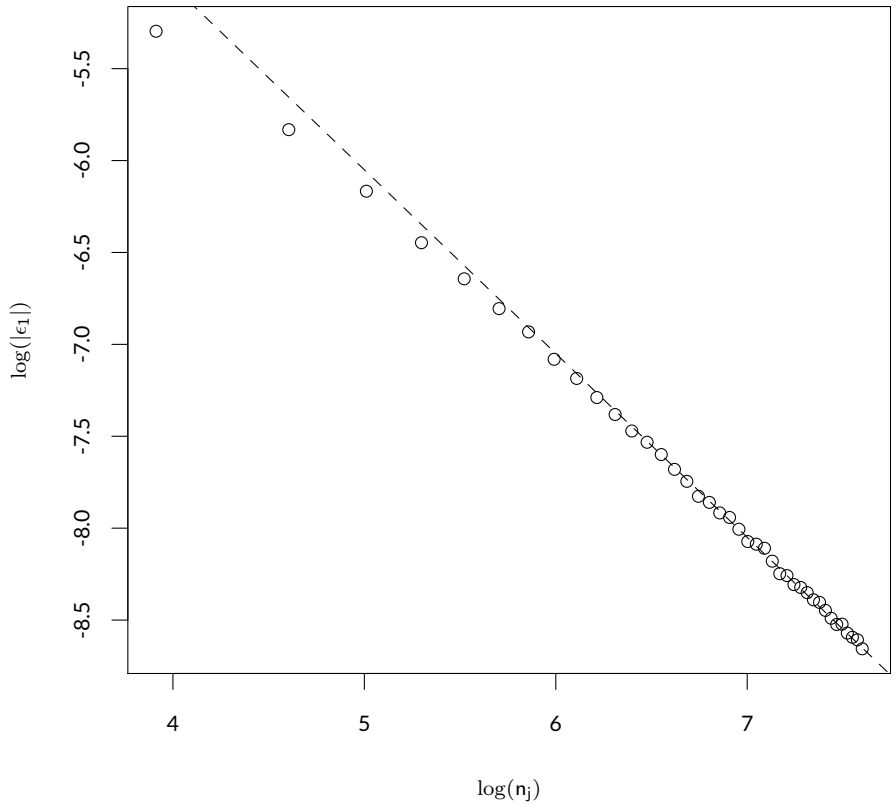

FIGURE 1 The log absolute error in the first-order Laplace approximation to the log-likelihood against the log of the number of observations in each cluster ($\mathsf{n_j}$), for a range of choices of $\mathsf{n_j}$. The overlaid dashed line has intercept -1.05 and slope -1.

Data sharing statement

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

APPENDIX

A PROOF OF MAIN RESULT

To prove Theorem 1, we aim to find the size of the contribution from each bipartition (P, Q).

Lemma 1. Suppose Condition 2 holds. For each fixed bipartition $(\mathsf{P}, \mathsf{Q}) \in \mathcal{M}_{\mathsf{l}}^{\mathsf{C}}$

$$e_{P,Q} = O_p\Big(\sum_{j=1}^{d} n_j^{-l}\Big).$$

Given Lemma 1, the proof of Theorem 1 is straightforward:

Proof of Theorem 1. By Lemma 1, we have $\mathsf{e}_{\mathsf{P},\mathsf{Q}} = \mathsf{O}_\mathsf{p}(\sum_{\mathsf{j}=1}^{\mathsf{d}} \mathsf{n}_\mathsf{j}^{-\mathsf{l}})$ for each fixed bipartition $(\mathsf{P}, \mathsf{Q}) \in \mathcal{M}_{\mathsf{l}}^{\mathsf{C}}$. Combining the contributions from each bipartition in $\mathcal{M}_{\mathsf{l}}^{\mathsf{C}}$, we have $\mathsf{e}_\mathsf{l} = \mathsf{O}_\mathsf{p}(\sum_{\mathsf{j}=1}^{\mathsf{d}} \mathsf{n}_\mathsf{j}^{-\mathsf{l}})$, so $\epsilon_\mathsf{k} = -\sum_{\mathsf{l}=\mathsf{k}}^{\infty} \mathsf{e}_\mathsf{k} = \mathsf{O}_\mathsf{p}(\sum_{\mathsf{j}=1}^{\mathsf{d}} \mathsf{n}_\mathsf{j}^{-\mathsf{k}})$, as required. □

In order to prove Lemma 1, we need some auxiliary results.

Proposition 4. Let (P, Q) be a fixed $(v, 2m)$ bipartition. For each $j = (j_1, \dots, j_{2m}) \in [1:d]^{2m}$, write $A_{P,Q}(j) = f_{j_{p_1}} \dots f_{j_{p_v}} f^{j_{q_1}} \dots f^{j_{q_m}}$. Then

$$e_{P,Q} = \frac{(-1)^v}{(2m)!} \sum_{j \in [1:d]^{2m}} n_{j_1}^{c_1} \dots n_{j_{2m}}^{c_{2m}} A_{P,Q}(j)$$

where $\sum_{j=1}^{2m} c_j = -l$, and each $c_j < 0$.

Proof. We may write

$$e_{P,Q} = \frac{(-1)^v}{(2m)!} \sum_{j \in [1:d]^{2m}} \prod_{p \in P} \prod_{k \in j_p} n_k^{1/|p|} f_{j_p} \prod_{q \in Q} \prod_{l \in j_q} n_l^{-1/2} f^{j_q} = \frac{(-1)^v}{(2m)!} \sum_{j \in [1:d]^{2m}} n_{j_1}^{c_1} \dots n_{j_{2m}}^{c_{2m}} \prod_{p \in P} f_{j_p} \prod_{q \in Q} f^{j_q}$$

for some $c_1, \dots c_{2m}$. We have $c_i = -\frac{1}{2} + \frac{1}{|p|}$, for whichever p contains i, so $c_i < 0$ as $|p| \geq 3$. We have

$$\sum_{i=1}^{2m} c_i = -m + \sum_{p \in P} \sum_{i \in p} \frac{1}{|p|} = -m + \sum_{p \in P} 1 = -m + v = -l$$

which gives the result. □

Proposition 5. Suppose $A = O_p^*(1)$ and $B = O_p^*(1)$, and C is the k-dimensional array with entries $C_j = A_{j_S} B_{j_T}$, where $j = (j_1, \dots, j_k)$, and $S, T \subseteq \{1, \dots, k\}$, such that $S \cup T = \{1, \dots, k\}$. If $S \cap T \neq \emptyset$, then $C = O_p^*(1)$.

Proof. We proceed by induction on $k = \dim(C) = |S \cup T|$. In the case $k = 1$, we have $S = T$, since $S \cap T \neq \emptyset$. So $C_{j_1} = A_{j_1} B_{j_1} = O_p(1)$, so $C = O_p^*(1)$. Now we suppose the hypothesis is true for $\dim(C) = k - 1$, and consider $\dim(C) = k \geq 2$. We have

$$C_{j_i}^i = \sum_{j_l : l \neq i} |C_j| = \sum_{j_l : l \neq i, a} \sum_{j_a} |C_j|.$$

Writing $j_{-a} = (j_1, \dots, j_{a-1}, j_{a+1}, \dots, j_k)$ and $C_{j_{-a}}^{-a} = \sum_{j_a} |C_j|$, if we can show that $C^{-a} = O_p^*(1)$ for some $a \neq i$, then

$$C_{j_i}^i = \sum_{j_l : l \neq i, a} C_{j_{-a}}^{-a} = O_p(1),$$

so that $C = O_p^*(1)$.

C^{-a} has entries

$$C_{j_{-a}}^{-a} = \sum_{j_a} |A_{j_S} B_{j_T}| = \begin{cases} |B_{j_T}| \sum_{j_a} |A_{j_S}| & \text{if } a \in S, a \notin T \\ |A_{j_S}| \sum_{j_a} |B_{j_T}| & \text{if } a \notin S, a \in T \\ \sum_{j_a} |A_{j_S} B_{j_T}| & \text{if } a \in S, a \in T \end{cases}$$

In the first case, we must have $\dim(A) \geq 2$, otherwise $S = \{a\}$ and $S \cap T = \emptyset$, which would be a contradiction. Since $A = O_p^*(1)$, the array A^{-a} with entries $A_{j_{S \setminus a}}^{-a} = \sum_{j_a} |A_{j_S}|$ must also be $O_p^*(1)$. So the array C^{-a} with entries $C_{j_{-a}}^{-a} = |B_{j_T}| A_{j_{S \setminus a}}^{-a}$ is $O_p^*(1)$, by the induction hypothesis, since $\dim(C^{-a}) = k - 1$. Similarly, in the second case $C^{-a} = O_p^*(1)$. In the third case,

$$C_{j_{-a}}^{-a} = \sum_{j_a} |A_{j_S} B_{j_T}| \leq \sum_{j_a} |A_{j_S}| \sum_{j_a} |B_{j_T}| = A_{j_{S \setminus a}}^{-a} B_{j_{T \setminus a}}^{-a}$$

by the Cauchy–Schwarz inequality. So $C^{-a} = O_p^*(1)$, by the induction hypothesis.

In all cases $C^{-a} = O_p^*(1)$, so $C = O_p^*(1)$, as required. □

Proposition 6. Suppose Condition 2 holds. Let (P, Q) be a fixed $(v, 2m)$ bipartition. Then $A_{P,Q} = O_p^*(1)$.

Proof. We have $A_{P,Q}(j) = \prod_{p \in P} f_{j_p} \prod_{q \in Q} f^{j_q}$. Since $f^{(k)} = O_p^*(1)$ for $k \geq 3$ and $[f^{(2)}]^{-1} = O_p^*(1)$, $A_{P,Q}$ is a product of $O_p^*(1)$ arrays.

We build up this product one term at a time, at each step applying Proposition 5 to show that the product remains $O_p^*(1)$.

We start with an arbitrary $p_1 \in P$, and choose $q_1 \in Q$ such that one element of q_1 is in p_1, and the other is not in p_1, and therefore must be in some other block $p_2 \in P$. If $v > 1$, it will always be possible to find such a q, because (P, Q) is a connected bipartition, so blocks of P (which form disjoint clusters in $\mathcal{G}_{P,Q}$) are connected by blocks of Q.

Let S^1 be the array with entries $S_{j_{p_1 \cup q_1}}^1 = f_{j_{p_1}} f^{j_{q_1}}$. Then $S^1 = O_p^*(1)$ by Proposition 5, as $p_1 \cap q_1 \neq \emptyset$. Let T^1 be the array with entries $T_{j_{p_1 \cup p_2}}^1 = f_{j_{p_2}} S_{j_{p_1 \cup q_1}}^1$. Then $T^1 = O_p^*(1)$, as $p_2 \cap (p_1 \cup q_1) = p_2 \cap q_1 \neq \emptyset$.

We continue to choose alternating terms from blocks of Q and P, at step k choosing a block q_k with one entry in $p_1 \cup \dots \cup p_k$, and the other entry in a new block p_{k+1}. At each stage k we have $S_{j_{p_1 \cup \dots \cup p_k \cup q_k}}^k = f^{j_{q_k}} T_{j_{p_1 \cup \dots \cup p_k}}^{k-1}$ and $T_{j_{p_1 \cup \dots \cup p_k \cup p_{k+1}}}^k = f_{j_{p_{k+1}}} S_{j_{p_1 \cup \dots \cup p_k \cup q_k}}^k$, where $S^k = O_p^*(1)$ and $T^k = O_p^*(1)$.

We continue until we have included all blocks of P, and have $T^{v-1}_{j_{p_1} \cup j_{p_2} \cup \ldots \cup j_{p_v}} = T^{v-1}_{j}$ where $T^{v-1} = O^*_p(1)$. We have already included terms from $v-1$ blocks of Q. We may multiply in the remaining $2m-v+1$ blocks of Q while retaining an $O^*_p(1)$ array by Proposition 5 as T^{v-1} is an array on all indices $j_1, \ldots, j_{2m}$, and $q \cap (1:2m) = q \neq \emptyset$ for each $q \in Q$. So $A_{P,Q} = O^*_p(1)$, as required. □

Proof of Lemma 1. By Proposition 4

$$|e_{P,Q}| = \frac{1}{(2m)!} \left| \sum_{j \in [1:d]^{2m}} n_{j_1}^{c_1} \ldots n_{j_{2m}}^{c_{2m}} A_{P,Q}(j) \right| \leq \frac{1}{(2m)!} \sum_{j \in [1:d]^{2m}} n_{j_1}^{c_1} \ldots n_{j_{2m}}^{c_{2m}} |A_{P,Q}(j)|. \tag{A1}$$

We apply the weighted form of the inequality of arithmetic and geometric means, which states that given non-negative numbers $x_1, \ldots, x_n$ and non-negative weights $w_1, \ldots, w_n$ with $\sum_i w_i = 1$,

$$\prod_{i=1}^{n} x_i^{w_i} \leq \sum_{i=1}^{n} w_i x_i.$$

Here, we let $n = 2m$, $x_i = n_{j_i}^{-l}$ and $w_i = -c_i/l$, to give that

$$n_{j_1}^{c_1} \ldots n_{j_{2m}}^{c_{2m}} \leq \sum_{i=1}^{2m} w_i n_{j_i}^{-l}. \tag{A2}$$

Putting (A2) back into (A1) gives

$$\begin{aligned} |e_{P,Q}| &\leq \frac{1}{(2m)!} \sum_{j \in [1:d]^{2m}} \sum_{i=1}^{2m} w_i n_{j_i}^{-l} |A_{P,Q}(j)| = \frac{1}{(2m)!} \sum_{i=1}^{2m} \sum_{j_i=1}^{d} w_i n_{j_i}^{-l} \sum_{j_1, \ldots, j_{i-1}, j_{i+1}, \ldots, j_{2m}} |A_{P,Q}(j)| \\ &= \frac{1}{(2m)!} \sum_{i=1}^{2m} \sum_{j_i=1}^{d} w_i n_{j_i}^{-1} A_{j_i}^{i} = \frac{1}{(2m)!} \sum_{i=1}^{2m} O_p\Big(\sum_{j_i=1}^{d} n_{j_i}^{-l} \Big) = O_p\Big(\sum_{j=1}^{d} n_j^{-l} \Big) \end{aligned}$$

since m is fixed as $d \to \infty$. □

B PROOFS FOR EXAMPLES

Proof of Proposition 1. The matrix of second derivatives of $g(.)$ with respect to u is $g^{(2)}(u) = h^{(2)}(u) + \Sigma^{-1}$, where $h^{(2)}(u)$ is the matrix of second derivatives of $h(.)$ with respect to u, and $h(.)$ is defined in (8). We have $h^{(2)}(u) = Z^T W(u) Z$, where $W(u)$ is a diagonal matrix with diagonal entries

$$W_{ii}(u) = \frac{b''(X_i^T \beta + Z_i^T u)}{a_i(\phi)}.$$

But $a_i(\phi) > 0$, and since $b(.)$ is a convex function $b''(X_i^T \beta + Z_i^T u) \geq 0$, so $W_{ii}(u) \geq 0$ for all u. So $W(u)$ is a non-negative definite matrix, and for any $x \in \mathbb{R}^d$, $x^T h^{(2)}(u) x = (xZ)^T W(Zx) \geq 0$, which means that $h^{(2)}(u)$ is non-negative definite. Since Σ^{-1} is positive definite, this means that $g^{(2)}(u)$ is positive definite for all u, so $g(.)$ is strictly convex, and therefore has a unique minimum. Since $b(.)$ is a smooth function, so is $g(.)$, so Condition 1 holds. □

Proof of Proposition 3. To prove the result, we need to show that after reparameterization Condition 2 holds with normalizing terms $n_1, \ldots, n_d$, so that we can apply Theorem 1. For $k \geq 3$, $g^{(k)}$ is diagonal with diagonal terms $g_{j \ldots j} = \Theta_p(n_j)$, so $f^{(k)} = O^*_p(1)$ for $k \geq 3$. It remains to show that $[f^{(2)}]^{-1} = O^*_p(1)$. Write

$$\Sigma^{[l]}_{jk} = \begin{cases} \sigma_2^2 + \sigma_3^2 + \ldots + \sigma_l^2 & \text{if } j = k \\ \sigma_3^2 + \ldots + \sigma_l^2 & \text{if } j \neq k, \text{ but } c_3(j) = c_3(k) \\ \vdots & \vdots \\ \sigma_l^2 & \text{if } c_{l-1}(j) \neq c_{l-1}(k), \text{ but } c_l(j) = c_l(k), \end{cases}$$

so that $\Sigma = \Sigma^{[L]}$. $\Sigma^{[l]}$ is a block-diagonal matrix, with d_l blocks, one for each level-l cluster. We have

$$\Sigma^{[l]}_{jk} = \begin{cases} \Sigma^{[l-1]}_{jk} + \sigma_l^2 & \text{if } c_l(j) = c_l(k) \\ 0 & \text{otherwise} \end{cases}$$

Write $\Sigma^{jk}_{[l]} = (\Sigma^{[l]})^{-1}_{jk}$. Applying the Sherman–Morrison formula to invert each block of $\Sigma^{[l]}$ gives

$$\Sigma^{jk}_{[l]} = \begin{cases} \Sigma^{jk}_{[l-1]} - \frac{\sigma_l^2 r_j r_k}{1 + \sigma_l^2 s_{c_l(j)}} & \text{if } c_l(j) = c_l(k) \\ 0 & \text{otherwise,} \end{cases} \tag{B3}$$

where

$$r_j = \sum_{k:c_l(k)=c_l(j)} \Sigma^{jk}_{[l-1]}, \quad s_c = \sum_{j:c_l(j)=c} r_j. \tag{B4}$$

We hypothesize that

$$\Sigma^{jk}_{[l]} = \begin{cases} \Theta(1) & \text{if } j = k \\ \Theta\big((\mathsf{d}^3_{\mathsf{c_3(j)}})^{-1}\big) & \text{if } j \neq k, \text{but } c_3(j) = c_3(k) \\ \vdots & \vdots \\ \Theta\big((\mathsf{d}^{\mathsf{l}}_{\mathsf{c_l(j)}})^{-1}\big) & \text{if } c_{l-1}(j) \neq c_{l-1}(k), \text{but } c_l(j) = c_l(k) \\ 0 & \text{otherwise,} \end{cases} \tag{B5}$$

and prove this by induction on l. This claim is true for $\mathsf{l} = 2$, as $\Sigma^{-1}_{[2]} = \sigma_2^{-2}\mathsf{I}$. For $\mathsf{l} \geq 2$, applying the induction hypothesis to (B4), we find $\mathsf{r_j} = \Theta(1)$, so $\mathsf{s_c} = \Theta(\mathsf{d}^{\mathsf{l}}_{\mathsf{c}})$ and

$$\frac{\sigma_l^2 r_j r_k}{1 + \sigma_l^2 s_{c_l(j)}} = \Theta\big((d_c^l)^{-1}\big). \tag{B6}$$

Substituting (B6) into (B3) proves (B5).

Now write $\mathsf{g}^{[\mathsf{l}]}_{\mathsf{jk}} = \mathsf{h_{jk}} - \Sigma^{\mathsf{jk}}_{[\mathsf{l}]}$, so that $\mathsf{g_{jk}} = \mathsf{g}^{[\mathsf{L}]}_{\mathsf{jk}}$. Again, $\mathsf{g}^{[\mathsf{l}]}$ is block-diagonal, and

$$g^{[l]}_{jk} = g^{[l-1]}_{jk} - (\Sigma^{jk}_{[l]} - \Sigma^{jk}_{[l-1]}) = \begin{cases} \mathsf{g}^{[\mathsf{l-1}]}_{\mathsf{jk}} + \frac{\sigma^2_{\mathsf{l}} \mathsf{r_j r_k}}{1+\sigma^2_{\mathsf{l}} \mathsf{s_{c_l(j)}}} & \text{if } c_l(j) = c_l(k) \\ 0 & \text{otherwise.} \end{cases}$$

Write $\mathsf{g}^{\mathsf{jk}}_{[\mathsf{l}]} = (\mathsf{g}^{[\mathsf{l}]})^{-1}_{\mathsf{jk}}$. Applying the Sherman–Morrison formula to invert each block of $\mathsf{g}^{[\mathsf{l}]}$ gives

$$g^{jk}_{[l]} = \begin{cases} \mathsf{g}^{\mathsf{jk}}_{[\mathsf{l-1}]} - \frac{\alpha \mathsf{a_j a_k}}{1+\alpha \mathsf{b_{c_l(j)}}} & \text{if } c_l(j) = c_l(k) \\ 0 & \text{otherwise,} \end{cases} \tag{B7}$$

where

$$\alpha = \frac{\sigma_l^2}{1 + \sigma_l^2 s_{c_l(j)}} = \Theta\big((d_l^{c_l(j)})^{-1}\big), \quad a_j = \sum_{k:c_l(k)=c_l(j)} r_j g^{jk}_{[l-1]}, \quad b_c = \sum_{j,k:c_l(j)=c_l(k)=c} r_j r_k g^{jk}_{[l-1]}. \tag{B8}$$

We hypothesize that

$$g^{jk}_{[l]} = \begin{cases} \mathsf{O_p}(\mathsf{n_j^{-1}}) & \text{if } j = k \\ \mathsf{O_p}\big((\mathsf{d}^3_{\mathsf{c_3(j)}})^{-1}\mathsf{n_j^{-1}}\mathsf{n_k^{-1}}\big) & \text{if } j \neq k, \text{but } c_3(j) = c_3(k) \\ \vdots & \vdots \\ \mathsf{O_p}\big((\mathsf{d}^{\mathsf{l}}_{\mathsf{c_l(j)}})^{-1}\mathsf{n_j^{-1}}\mathsf{n_k^{-1}}\big) & \text{if } c_{l-1}(j) \neq c_{l-1}(k), \text{but } c_l(j) = c_l(k) \\ 0 & \text{otherwise,} \end{cases} \tag{B9}$$

and prove this by induction on l. This claim is true for $\mathsf{l} = 2$, as $\mathsf{g}^{[2]}$ is diagonal, with diagonal entries $\mathsf{h_{jj}} + \sigma_2^{-2} = \Theta_{\mathsf{p}}(\mathsf{n_j})$. For $\mathsf{l} \geq 2$, applying the induction hypothesis to (B8), recalling that $\mathsf{r_j} = \Theta(1)$, we find

$$a_j = \sum_{k:c_l(k)=c_l(j)} O_p\big((d^l_{c_l(j)})^{-1} n_j^{-1} n_k^{-1}\big) = O_p(n_j^{-1})$$

and

$$b_c = \sum_{k:c_l(k)=c_l(j)=c} O_p\big((d^l_c)^{-1} n_j^{-1} n_k^{-1}\big) = O_p(1),$$

so

$$\frac{\alpha\, a_j a_k}{1 + \alpha b_{c_l(j)}} = \frac{a_j a_k}{\alpha^{-1} + b_{c_l(j)}} = O_p\big((d_l^{c_l(j)})^{-1} n_j^{-1} n_k^{-1}\big) \tag{B10}$$

Substituting (B10) into (B7) proves (B9). Normalizing,

$$f^{jk} = n_j^{1/2} n_k^{1/2} g^{jk} = n_j^{1/2} n_k^{1/2} g^{jk}_{[L]} = \begin{cases} \mathsf{O_p}(1) & \text{if } j = k \\ \mathsf{O_p}\big((\mathsf{d}^{\mathsf{l}}_{\mathsf{c_l(j)}})^{-1}\mathsf{n_j^{-1/2}}\mathsf{n_k^{-1/2}}\big) & \text{if } c_{l-1}(j) \neq c_{l-1}(k), \text{but } c_l(j) = c_l(k), \\ & \text{for } l = 3, \ldots, L \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$\sum_k |f^{jk}| = O_p\left(1 + \sum_{l=3}^{L} \sum_{k:c_{l-1}(j)\neq c_{l-1}(k),c_l(j)=c_l(k)} \left(d^l_{c_l(j)}\right)^{-1} n_j^{-1/2} n_k^{-1/2}\right)$$
$$= O_p\left(1 + \sum_{l=3}^{L} d^l_{c_l(j)} \left(d^l_{c_l(j)}\right)^{-1} n_j^{-1/2} \max_k\{n_k^{-1/2}\}\right) = O_p(1 + n_j^{-1/2}) = O_p(1),$$

and $\sum_j |f^{jk}| = \sum_k |f^{kj}| = O_p(1)$, so $f^{(2)} = O_p^*(1)$. So Condition 2 holds with normalizing terms $n_1, \ldots, n_d$, and Theorem 1 gives that $\epsilon_k(\theta) = O_p(\sum_{j=1}^{d} n_j^{-k})$, as required. □

References

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

Butler, R. W. (2007). *Saddlepoint approximations with applications*. Cambridge University Press. doi: 10.1017/CBO9780511619083

Douc, R., Moulines, E., & Rydén, T. (2004). Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *The Annals of Statistics*, *32*(5), 2254–2304.

Friel, N., & Wyse, J. (2012). Estimating the evidence - a review. *Statistica Neerlandica*, *66*(3), 288–308. doi: 10.1111/j.1467-9574.2011.00515.x

Goutis, C., & Casella, G. (1999). Explaining the Saddlepoint Approximation. *The American Statistician*, *53*(3), 216–224.

McCullagh, P. (1987). Tensor methods in statistics. In (pp. 254–256). Chapman and Hall.

Nie, L. (2007). Convergence rate of MLE in generalized linear and nonlinear mixed-effects models: Theory and applications. *Journal of Statistical Planning and Inference*, *137*(6), 1787–1804.

Ogden, H. E. (2017). On asymptotic validity of naive inference with an approximate likelihood. *Biometrika*, *104*(1), 153–164.

Raudenbush, S. W., Yang, M.-L., & Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics*, *9*(1), 141–157.

Reid, N. (1988). Saddlepoint methods and statistical inference. *Statistical Science*, *3*(2), 213–238. doi: 10.1214/ss/1177012906

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *71*(2), 319–392.

Shun, Z., & McCullagh, P. (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(4), 749–760.

Small, C. G. (2010). Expansions and asymptotics for statistics. In (chap. 6). Chapman and Hall/CRC.

Tierney, L., & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, *81*(393), 82–86.

Zipunnikov, V., & Booth, J. G. (2011). *Closed form GLM cumulants and GLMM fitting with a SQUAR-EM-LA 2 algorithm.*