

Asymptotics of numerical integration for two-level mixed models

BLAIR BILODEAU^{1,a} , ALEX STRINGER^{2,b}  and YANBO TANG^{*3,c} 

¹University of Toronto, ^ablair.bilodeau@mail.utoronto.ca

²University of Waterloo, ^balex.stringer@uwaterloo.ca

³Imperial College London, ^cyanbo.tang@imperial.ac.uk

We study mixed models with a single grouping factor where inference about unknown parameters requires optimizing a marginal likelihood defined by an intractable integral. Low-dimensional numerical integration techniques are regularly used to approximate these integrals with inferences about parameters based on the resulting approximate marginal likelihood. For a generic class of mixed models that satisfy explicit regularity conditions we derive the stochastic relative error rate incurred for both the likelihood and maximum likelihood estimator when adaptive numerical integration is used to approximate the marginal likelihood. We then specialize the analysis to well-specified generalized linear mixed models having exponential family response and multivariate Gaussian random effects, verifying that the regularity conditions hold and hence that the convergence rates apply. We also prove that for models with likelihoods satisfying very weak concentration conditions that the maximum likelihood estimators from non-adaptive numerical integration approximations of the marginal likelihood are not consistent, further motivating adaptive numerical integration as the preferred tool for inference in mixed models. Code to reproduce the simulations in this paper is provided at <https://github.com/awstringer1/aq-theory-paper-code>.

Keywords: Adaptive quadrature; approximate inference; generalized linear models

1. Introduction

1.1. Approximate integration in statistical problems

Approximate integration is often required when fitting statistical models by maximum likelihood. The concentration properties of likelihood functions as the number of data increase render such statistical problems distinct from deterministic numerical integration problems, and a distinct perspective on convergence is required. In this paper we provide comprehensive convergence theory of maximum likelihood estimation in two-level linear and nonlinear mixed models in which the integral defining the marginal likelihood is approximated using numerical integration. We show that if the maximum likelihood estimate based on the exact likelihood converges in distribution to a random variable, then so does the maximum likelihood estimate based on an approximate likelihood computed using adaptive numerical integration, as long as the order of the accuracy of the integration rule is chosen high enough. We further show the negative result that non-adaptive integration rules cannot yield consistent maximum likelihood estimators in any statistical model in which the normalized likelihood function concentrates around its mode. The conclusion is that non-adaptive numerical integration should not be used for approximate likelihood inference in any statistical model, while in mixed models adaptive integration can always be made to provide inferences asymptotically indistinguishable from inferences based on the exact likelihood by choosing an accurate enough integration rule.

*Equal contribution authorship.

In contrast to previous work (Bianconcini, 2014, Jiang, Wand and Bhaskaran, 2021, Vonesh, 1996), rather than focusing on a particular class of response and random effect distributions, in the present work we devise a set of explicit and verifiable regularity conditions about these distributions under which our convergence results hold. We then verify that these conditions hold for well-specified generalized linear models, i.e. those having exponential family response distributions and multivariate Gaussian random effects. This makes our convergence theory general and also generalizable. All that must be done in order to apply our theorems to a particular model not explicitly treated in the present work is to verify that the conditions hold. By explicitly verifying our conditions for a particular class of models we ensure that the conditions are reasonable, and not so strong as to rule out potentially interesting models.

1.2. Adaptive numerical integration

In “regular” statistical models, suitably normalized likelihood functions concentrate around their modes as more data are obtained. This concentration behaviour is what enables consistent inference and quantification of uncertainty using classical asymptotic normality results, e.g. van der Vaart (1998, Ch. 5, 7). Failure to account for this concentration behaviour can lead to failure of standard numerical integration techniques when applied to statistical problems. This phenomenon has long been observed to occur when using Gaussian quadrature to fit mixed models—e.g. Lesaffre and Spiessens (2001)—and is discussed by others (Bianconcini, 2014). However, the failure of numerically accurate integration methods in simple statistical problems has not been explained theoretically. In Theorem 1 in Subsection 3.1 we prove that any likelihood that concentrates around its mode as more data are obtained cannot lead to consistent maximum likelihood estimates when approximated using a fixed quadrature rule.

A standard solution to this problem is to adapt the quadrature rule to the location and curvature of the likelihood by shifting and scaling the points and weights in a data-dependent manner. In the statistical literature this is referred to as *adaptive quadrature* (AQ) at least as far back as Naylor and Smith (1982), although this term conflicts somewhat with the numerical analysis literature which ascribes a different meaning to “adaptive”. Adaptive quadrature rules have enjoyed common use in fitting generalized linear and nonlinear mixed models (Pinheiro and Bates, 1995) and are widely available in common software including R (package `lme4`, Bates et al. 2015; package `GLMMadaptive`, Rizopoulos 2020), SAS (PROC `NLMIXED`), STATA (function `gllamm`). However, convergence theory for these methods is limited. In Section 3 we prove that the approximate maximum likelihood estimator obtained by maximizing an adaptive quadrature approximation to the marginal likelihood in a mixed model attains higher-order asymptotic accuracy (Lemma 1), leading to a consistent approximate maximum likelihood estimator (Lemma 2) having the same large-sample statistical properties as the maximum of the exact likelihood if enough quadrature points are used (Theorem 2). We prove that these results hold under very general yet precise and verifiable regularity conditions (Assumptions 2 to 7 in Appendix A) which we formally verify for generalized linear mixed models with exponential family responses and multivariate Gaussian random effects (Proposition 1 in Section 4). Simulations in Subsection 4.3 illustrate the practical implications of the theory.

1.3. Technical approach and related work

Asymptotic convergence results for exact maximum likelihood estimators (disregarding integration error) have recently been studied for mixed models; see Bhaskaran and Wand (2023), Jiang, Wand and Bhaskaran (2021), Mastrini, Bhaskaran and Wand (2024), Nie (2007). Our Theorem 2 complements these novel results by including the error incurred by the approximate integration required to implement

maximum likelihood estimation in these models, rendering them directly applicable to practice; see Corollary 2 in Subsection 4.2.

The Laplace approximation (quadrature with a single point) is the default estimation method in the popular `lme4` software package (Bates et al., 2015) and is closely related to the penalized quasi-likelihood method (Breslow and Clayton, 1993). Vonesh (1996) studies the rate of convergence of Laplace-approximate maximum marginal likelihood estimators in Gaussian nonlinear mixed effects models. Our Theorem 2 recovers the rate of consistency of Vonesh (1996) up to a small constant, applies to a much broader class of models and likelihood approximations, and extends it in combination with Jiang, Wand and Bhaskaran (2021) to include asymptotic normality in addition to consistency. Ogden (2017) studies the much more general problem of obtaining the rate of consistency and the asymptotic normality of an approximate maximum likelihood estimator obtained by maximizing *any* approximate likelihood. Their theory is proved under the very strong conditions that two derivatives of the approximate log-likelihood converge uniformly in probability to the corresponding likelihood derivatives at a specific rate. Ogden (2017) applies this theory to Bernoulli generalized linear mixed models fit by Laplace-approximate marginal likelihood. However, the uniform convergence of derivatives is not formally established, and this appears highly nontrivial. We provide results for mixed models fit by adaptive Gaussian quadrature with any number of quadrature points, under conditions on the model and data-generating distribution which we formally verify for generalized linear mixed models with any exponential family distribution. Our rates recover those of Ogden (2017) again up to a small constant factor.

Beyond a single quadrature point, a deterministic rate of convergence for AQ was first derived by Liu and Pierce (1994) and later corrected by Jin and Andersson (2020). Bianconcini (2014) studies stochastic convergence of the approximate maximum marginal likelihood estimator using AQ in a class of latent variable models with exponential family response that are included in the class of mixed models we consider in the present paper; see Subsection 2.1 for details. Like Ogden (2017) they assume uniform convergence in probability of two derivatives of the likelihood approximation. They further assume the highly non-trivial result of uniform convergence in probability of the likelihood approximation over an unspecified region in the parameter space. We formally prove a precise version of this assumption in Lemma 1 and this yields a rate of convergence which is fundamental to the subsequent rate of consistency we derive in Theorem 2. Therein we clarify the important detail that such uniform convergence only occurs in a suitably shrinking region in the parameter space. Our rates recover those of Bianconcini (2014) up to a small constant factor.

The small additional constant factor mentioned in the previous paragraphs is due to our uniform regularity conditions. This term facilitates verifying that the conditions actually hold for well-specified exponential family models, specifically through Lemmas 10 and 11 in Appendix D.2. Previous authors do not include this term in their conditions and hence it does not show up in their convergence rates, but they also do not verify that their conditions hold. It is not clear how to verify these conditions without this small additional factor. We note that there is no apparent practical difference between our slightly looser rates and those reported by previous authors. A similar ε factor appears in the error rate of Quasi-Monte Carlo integration (Owen, 2019). This factor is not expected to be practically impactful for quadrature since the number of quadrature points can be increased in practice to obtain a faster rate if desired.

Our results depend directly on the recent work of Bilodeau, Stringer and Tang (2024a) who gave the first stochastic rate of convergence for the related task of Bayesian posterior normalization under model-based regularity conditions. Their result may be applied to our present situation pointwise for any fixed parameter value. A primary technical contribution of the present work is to upgrade their result to hold uniformly across parameter values and numbers of groups as is required to assess the convergence of the maximum likelihood estimator in mixed models. This requires (a) suitably upgraded regularity

conditions, which we formally verify for generalized linear mixed models with multivariate Gaussian random effects; and (b) an appropriate modification of the entire proof of the main result of [Bilodeau, Stringer and Tang \(2024a\)](#) to hold under these new conditions at a slightly looser rate.

A further technical challenge addressed in the present paper is that uniform convergence of the approximate likelihood cannot occur in a *fixed* neighbourhood of the true parameter value. This is because the adaptation of the quadrature points is centred around the mode of the latent variables which itself must converge in probability to some point, but this convergence can only be made to occur in a *shrinking* neighbourhood of the true parameter value. This subtlety has not been treated by previous authors, who often implicitly assume a form of uniform convergence over too large of a space. This is too strong of an assumption to verify and indeed has not been previously verified. The primary technical challenge that we solve in order to achieve the required uniform convergence is to identify the precise region in the parameter space in which this convergence holds and the rate at which it should be made to shrink as the number of data increase. This requires a balanced approach as follows. We first apply the technical lemmas of [Bilodeau, Stringer and Tang \(2024a\)](#) to provide a formal proof of the rate of uniform convergence in probability of the likelihood approximation (Lemma 1) at a rate depending on the order of the quadrature rule on an appropriately shrinking neighbourhood. We then provide a proof of consistency of the approximate maximum likelihood estimator (Lemma 2) which guarantees that it is eventually in this shrinking neighbourhood. The main difficulty of the proof involves balancing these rates where uniform consistency is easier to show for a faster shrinking ball but harder to show that the approximate maximum likelihood estimator remains in a faster shrinking neighbourhood. We then combine these two results and our regularity conditions to prove that the distance between the exact and approximate maximum likelihood estimators converges faster than does the exact estimator to its limit on the shrinking neighbourhood (Theorem 2). The result is a general and self-contained convergence theory for approximate maximum likelihood estimation for two-level mixed models.

2. Preliminaries

2.1. Two-level mixed models

We study two-level mixed models for repeated measurements data consisting of a total number of observations $N \in \mathbb{N}$, number of groups $m \in [N] = \{1, \dots, N\}$, and number of observations per group $(n_i)_{i \in [m]}$ satisfying $N = \sum_{i=1}^m n_i$. For $i \in [m]$ let $\mathbf{y}_i = (y_{ij})_{j \in [n_i]}$, $y_{ij} \in \mathbb{R}$ denote the observations from group i . Let P_N^* denote the true unknown joint distribution of $\mathbf{y} = (\mathbf{y}_1^T, \dots, \mathbf{y}_m^T)^T$. In addition for each $i \in [m]$ and $j \in [n_i]$ there are observed covariates $\mathbf{x}_{ij} \in \mathbb{R}^q$ which may be fixed or random. They are implicitly conditioned upon for the remainder of the paper.

A mixed model is determined by a random effect distribution G with known mean parameterized by variance parameters $\boldsymbol{\sigma} \in \Theta_{\text{Re}} \subseteq \mathbb{R}^s$, a conditional response distribution F parameterized by regression coefficients $\boldsymbol{\beta} \in \Theta_{\text{Mod}} \subseteq \mathbb{R}^q$, and possibly additional dispersion parameters $\boldsymbol{\phi} \in \Theta_{\text{Disp}} \subseteq \mathbb{R}^l$ such that

$$\begin{aligned} \mathbf{y}_{ij} \mid \mathbf{x}_{ij}, \mathbf{u}_i &\stackrel{\text{ind.}}{\sim} F(\mathbf{x}_{ij}, \mathbf{u}_i; \boldsymbol{\beta}, \boldsymbol{\phi}), \\ \mathbf{u}_i &\stackrel{\text{ind.}}{\sim} G(\boldsymbol{\sigma}). \end{aligned} \tag{2.1}$$

The distinction between regression and variance parameters is somewhat arbitrary in the definition and is made only for convenience. In practice different parameters will have different convergence rates on a model-specific basis. See Corollary 2 for an example from generalized linear mixed models.

Model (2.1) is quite general, and restrictions yield more familiar models. For example a generalized linear mixed model (GLMM; [Breslow and Clayton 1993](#)) is obtained by constraining F to be from an

exponential family with canonical parameter linear in $\mathbf{u}_i$ and $\boldsymbol{\beta}$ and constraining G to be Gaussian; see Section 4. The latent variable models analyzed by Bianconcini (2014) are obtained by having the canonical parameter depend linearly on $\boldsymbol{\beta}^\top \mathbf{u}$ and G to be a *spherical* Gaussian depending on no further variance parameters. A nonlinear mixed model (NLMM; Pinheiro and Bates 1995) is obtained by constraining F to be Gaussian with mean given by a known nonlinear function of $\mathbf{x}$, $\boldsymbol{\beta}$, and $\mathbf{u}$, and constraining G to be Gaussian. The regularity conditions for these distributions under which the convergence of approximate likelihoods holds are stated in Appendix A. Proposition 1 in Section 4 demonstrates that these regularity conditions are weaker than requiring the response to belong to the exponential family or the random effects to be Gaussian by demonstrating that they are satisfied by exponential family models with correlated multivariate Gaussian random effects.

We denote the model for the joint density of the group's responses and random effects by $\pi(\mathbf{y}_i, \mathbf{u}_i; \boldsymbol{\theta}) = \prod_{j=1}^{n_i} f(y_{ij} | \mathbf{u}_i; \boldsymbol{\beta}, \boldsymbol{\phi}) g(\mathbf{u}_i; \boldsymbol{\sigma})$ where f and g are the densities of F and G . Inferences about the unknown parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\sigma})$ are based on the marginal likelihood,

$$\pi(\mathbf{y}; \boldsymbol{\theta}) = \prod_{i=1}^m \pi(\mathbf{y}_i; \boldsymbol{\theta}) = \prod_{i=1}^m \int \pi(\mathbf{y}_i, \mathbf{u}_i; \boldsymbol{\theta}) d\mathbf{u}_i, \quad (2.2)$$

where the first equality follows from the independence of $\mathbf{u}_1, \dots, \mathbf{u}_m$. The joint distribution of data and random effects may be regarded as *high-dimensional* because $\dim(\mathbf{u}) = dm$ and $m \rightarrow \infty$ as more data are obtained. Numerical integration typically incurs a computational cost that is exponential in the dimension of the integrand and hence general high-dimensional numerical integration is infeasible unless some structure of the integral is exploited to develop a more efficient approximation. In two-level mixed models the dm -dimensional integral defining the marginal likelihood factors into a product of m integrals of dimension d where $d = \dim(\mathbf{u}_i)$ is typically quite small. This enables the use of accurate low-dimensional quadrature techniques within what is nominally a high-dimensional problem. Inferences are therefore based on an approximation to $\pi(\mathbf{y}; \boldsymbol{\theta})$ obtained by approximating each integral $\pi(\mathbf{y}_i; \boldsymbol{\theta})$ and then taking the product of the approximations. As the number of groups increases the number of approximate integrals being multiplied increases and the error in the approximate likelihood will grow. The accuracy of the approximation will therefore affect the statistical properties of inferences based on the resulting approximate marginal likelihood with larger data requiring integration rules that attain higher accuracy. A reviewer points out that this phenomenon also occurs when using Monte Carlo integration methods which may require a larger number of samples to attain the same accuracy on a larger set of data. This paper focuses on the statistical properties of inferences based on quadrature/cubature approximations to this marginal likelihood.

2.2. Fixed quadrature

A *quadrature rule* $\mathfrak{R}(Q, \omega)$ to approximate integral of the form in Eq. (2.2) is a collection of *points* $Q \subseteq \mathbb{R}^d$ and a *weight function* $\omega : Q \rightarrow \mathbb{R}$ and is denoted by $\mathfrak{R}(Q, \omega)$. Given such a rule an approximate marginal likelihood is:

$$\tilde{\pi}(\mathbf{y}_i; \boldsymbol{\theta}) = \sum_{\mathbf{z} \in Q} \pi(\mathbf{y}_i, \mathbf{z}; \boldsymbol{\theta}) \omega(\mathbf{z}). \quad (2.3)$$

We call any such approximation $\mathfrak{R}(Q, \omega)$ for which the points $\mathbf{z} \in Q$ and weights ω do not depend on the data to be a *non-adaptive* or *fixed* quadrature rule. We focus on quadrature rules with the following exact integration property.

Definition 1 (Definition 1 of Bilodeau, Stringer and Tang (2024a)). For any $k, d \in \mathbb{N}$, a quadrature rule $\mathfrak{R}(\mathcal{Q}, \omega)$ satisfies $\mathcal{P}(k, d)$ if for all d -dimensional real-valued polynomials P of total order $2k - 1$ or less,

$$\int_{\mathbb{R}^d} \phi(\mathbf{u}; 0, \mathbf{I}_d) P(\mathbf{u}) d\mathbf{u} = \sum_{\mathbf{z} \in \mathcal{Q}} \phi(\mathbf{z}; 0, \mathbf{I}_d) P(\mathbf{z}) \omega(\mathbf{z}), \quad (2.4)$$

where $\phi(\mathbf{u}; 0, \mathbf{I}_d)$ is the standard d -dimensional Gaussian density.

The Gauss–Hermite quadrature rule satisfies $\mathcal{P}(k, 1)$ and the multi-dimensional product rule extension $\mathfrak{R}(\mathcal{Q}^d, \omega_1 \cdots \omega_d)$ satisfies $\mathcal{P}(k, d)$ for $d > 1$. See Bilodeau, Stringer and Tang (2024a) for discussion of other rules and extensions to multiple dimensions.

Classical convergence analysis (e.g. Davis and Rabinowitz 1984) predicts small error when specific quadrature rules are applied to *deterministic* functions satisfying specific properties. Likelihood functions are random and such analysis predicts that accurate results may be obtained if a particular sample of data leads to a realized likelihood function which is well-behaved in this sense. However, statistical convergence theory requires a probabilistic analysis of the error incurred under assumptions on the data-generating distribution and model. In this paper we show that no fixed quadrature rule—no matter how accurate—can yield an approximate marginal likelihood with the correct statistical properties; see Theorem 1 in Section 3. When used in statistical problems quadrature rules must be adapted to the data.

2.3. Adaptive quadrature

Bernstein-von Mises theory (van der Vaart, 1998, Section 10.2) states that, when appropriately normalized, likelihood functions satisfying weak conditions concentrate around their modes as more data are obtained. Approximating integrals involving likelihood functions (such as Eq. (2.2)) using fixed quadrature rules ignores this physical behaviour and results in a quadrature rule that misses most of the mass of the integrand as more data are obtained; see Theorem 1. This phenomenon is the source of the empirical lack of accuracy that has frequently been observed when using GQ to fit mixed models (e.g. Lesaffre and Spiessens 2001). Consequently, *adaptive quadrature* (AQ) is popular for fitting mixed models. In this paper we show that this method does have desirable convergence properties for mixed models; see Theorem 2 in Subsection 3.3.

For each $i \in [m]$ and $\theta \in \Theta$ let $\ell_i^\theta(\mathbf{u}) = \log \pi(\mathbf{y}_i, \mathbf{u}; \theta)$, $\hat{\mathbf{u}}_i^\theta = \operatorname{argmax}_{\mathbf{u} \in \mathbb{R}^d} \ell_i^\theta(\mathbf{u})$, and $\mathbf{H}_i^\theta = -\partial_{\mathbf{u}}^2 \ell_i^\theta(\hat{\mathbf{u}}_i^\theta)$. Define $\mathbf{L}_i^\theta$ to be the lower Cholesky triangle satisfying $(\mathbf{H}_i^\theta)^{-1} = \mathbf{L}_i^\theta (\mathbf{L}_i^\theta)^\top$. For a given quadrature rule $\mathfrak{R}(\mathcal{Q}, \omega)$ that satisfies $\mathcal{P}(k, d)$, the AQ approximation to $\pi(\mathbf{y}_i; \theta)$ is:

$$\tilde{\pi}^{\text{AQ}}(\mathbf{y}_i; \theta) = |\mathbf{L}_i^\theta| \sum_{\mathbf{z} \in \mathcal{Q}} \pi(\mathbf{y}_i, \mathbf{L}_i^\theta \mathbf{z} + \hat{\mathbf{u}}_i^\theta; \theta) \omega(\mathbf{z}). \quad (2.5)$$

The corresponding likelihood approximation is $\tilde{\pi}^{\text{AQ}}(\mathbf{y}; \theta) = \prod_{i=1}^m \tilde{\pi}^{\text{AQ}}(\mathbf{y}_i; \theta)$. When $k = 1$ and $\mathfrak{R}(\mathcal{Q}, \omega)$ is Gauss-Hermite quadrature, $\tilde{\pi}^{\text{AQ}}(\mathbf{y}; \theta)$ is called a Laplace approximation. Adaptive quadrature incurs the computational cost of fixed quadrature along with the cost of obtaining the mode and Hessian of the log-integrand. The former cost usually dominates the latter in practice.

3. Theoretical properties

3.1. Non-convergence of fixed quadrature approximations to likelihood functions

We provide the first theoretical explanation for the empirical lack of accuracy of non-adaptive quadrature rules in statistical problems noted by previous authors (Bianconcini, 2014, Lesaffre and Spiessens, 2001). Theorem 1 states that any non-adaptive quadrature rule cannot yield an asymptotically convergent likelihood approximation as long as the likelihood admits weak concentration properties. Corollary 1 then applies Theorem 1 to mixed models (Eq. (2.1)).

Consider a latent variable model which assumes that the density of random variable $\mathbf{Y}$ is given by $\pi(\mathbf{y}) = \int \pi(\mathbf{y}, \mathbf{u}) d\mathbf{u}$ for some joint density $\pi(\mathbf{y}, \mathbf{u})$. This is identical to the Bayesian setup within which Bilodeau, Stringer and Tang (2024a) argued that adapting a quadrature rule as described in Subsection 2.3 is *sufficient* to achieve fast asymptotic convergence of the approximation as $N \rightarrow \infty$. Theorem 1 provides a type of converse to Bilodeau, Stringer and Tang (2024a, Theorem 1): adapting the quadrature rule to the data—somehow—is *necessary* to achieve asymptotic convergence.

Theorem 1. Fix any quadrature rule $\mathfrak{R}(Q, \omega)$ where $Q \subseteq \mathbb{R}^d$ and $\omega : Q \rightarrow \mathbb{R}$. Let the random variable $\mathbf{Y} \in \mathbb{R}^N$ have distribution P_N^* . Let the density of $\mathbf{Y}$ under the model be $\pi(\mathbf{y}) = \int \pi(\mathbf{y}, \mathbf{u}) d\mathbf{u}$, and let

$$\tilde{\pi}(\mathbf{y}) = \sum_{\mathbf{z} \in Q} \omega(\mathbf{z}) \pi(\mathbf{y}, \mathbf{z}).$$

Denote by $\xrightarrow{P}$ convergence in probability with respect to P_N^* . If there exists $\mathbf{u}_*$ such that $\pi(\mathbf{u}_* | \mathbf{y}) \xrightarrow{P} \infty$ and $\pi(\mathbf{u} | \mathbf{y}) \xrightarrow{P} 0$ for all $\mathbf{u} \neq \mathbf{u}_*$, then there exist $\kappa > 0$ and $\gamma \in (0, 1)$ such that

$$\lim_{N \rightarrow \infty} P_N^* \left\{ \left| \frac{\tilde{\pi}(\mathbf{y})}{\pi(\mathbf{y})} - 1 \right| > \kappa \right\} > \gamma.$$

Proof. Observe that

$$\frac{\tilde{\pi}(\mathbf{y})}{\pi(\mathbf{y})} = \sum_{\mathbf{z} \in Q} \omega(\mathbf{z}) \pi(\mathbf{z} | \mathbf{y}),$$

where $\pi(\mathbf{z} | \mathbf{y})$ is the posterior density of $\mathbf{u}$ evaluated at $\mathbf{z}$. We can therefore write

$$\underline{\omega} \times \max_{\mathbf{z} \in Q} \pi(\mathbf{z} | \mathbf{y}) \leq \frac{\tilde{\pi}(\mathbf{y})}{\pi(\mathbf{y})} \leq |Q| \bar{\omega} \times \max_{\mathbf{z} \in Q} \pi(\mathbf{z} | \mathbf{y}), \quad (3.1)$$

where $\underline{\omega} = \min_{\mathbf{z} \in Q} \omega(\mathbf{z})$ and $\bar{\omega} = \max_{\mathbf{z} \in Q} \omega(\mathbf{z})$.

There are two cases to consider. Suppose first that $\mathbf{u}_* \in Q$. Then by Eq. (3.1) and the assumption of the theorem, $\tilde{\pi}(\mathbf{y})/\pi(\mathbf{y}) \xrightarrow{P} \infty$. We therefore may choose $\epsilon > 0, \gamma \in (0, 1)$ such that there must exist $n \in \mathbb{N}$ such that for every $N > n$,

$$P_N^* \left\{ \frac{\tilde{\pi}(\mathbf{y})}{\pi(\mathbf{y})} > \epsilon + 1 \right\} > \gamma.$$

Set $\kappa = \epsilon > 0$ and note that $\tilde{\pi}(\mathbf{y}) > \pi(\mathbf{y})$ eventually to yield the result. Suppose next that $\mathbf{u}_* \notin Q$. Then by Eq. (3.1), $\tilde{\pi}(\mathbf{y})/\pi(\mathbf{y}) \xrightarrow{P} 0$. Choose $\epsilon, \gamma \in (0, 1)$ such that there must exist $n \in \mathbb{N}$ such that for every

$N > n$:

$$\gamma < P_N^* \left\{ \frac{\tilde{\pi}(\mathbf{y})}{\pi(\mathbf{y})} < \varepsilon \right\} = P_N^* \left\{ 1 - \frac{\tilde{\pi}(\mathbf{y})}{\pi(\mathbf{y})} > 1 - \varepsilon \right\} = P_N^* \left\{ \left| \frac{\tilde{\pi}(\mathbf{y})}{\pi(\mathbf{y})} - 1 \right| > 1 - \varepsilon \right\},$$

where the last step uses that $\tilde{\pi}(\mathbf{y}) < \pi(\mathbf{y})$ eventually. Set $\kappa = 1 - \varepsilon > 0$ to yield the result. $\square$

The most obvious way to guarantee the conditions of Theorem 1 is for the model to satisfy a Bernstein-von Mises theorem (van der Vaart, 1998, Section 10.2). Kleijn and van der Vaart (2012) show that a Bernstein-von Mises-type result holds for a very broad class of misspecified models which suggests that Theorem 1 is widely applicable and its conclusions apply to many models used in practice.

Returning focus to the mixed models which are the subject of the present paper, Corollary 1 specializes Theorem 1 to mixed models of the form given in Eq. (2.1) with “true” parameter value θ_* .

Corollary 1. *Consider the model given by Eq. (2.1). Under Assumptions 2 to 7 in Appendix A and for θ_* defined therein, there exist $\kappa > 0$ and $\gamma \in (0, 1)$ such that*

$$\lim_{N \rightarrow \infty} P_N^* \left\{ \left| \frac{\tilde{\pi}(\mathbf{y}; \theta_*)}{\pi(\mathbf{y}; \theta_*)} - 1 \right| > \kappa \right\} > \gamma.$$

Proof. Restricting attention to $\theta = \theta_*$ reduces the problem to exactly that considered by Bilodeau, Stringer and Tang (2024a), and our Assumptions 2 to 4, 6 and 7 reduce to their Assumptions 1 – 5. In their Remark 5 they show that these assumptions imply that the Bernstein-von Mises theorem holds for $\pi(\mathbf{y}, \mathbf{u}; \theta_*)$; this in turn implies the conditions of Theorem 1. $\square$

While Corollary 1 only applies to the single parameter value θ_* and only states that the error cannot reach zero—as opposed to, say, diverging to ∞ —it is nonetheless sufficient to rule out inferences based on $\tilde{\pi}(\mathbf{y}; \theta)$ for most mixed models used in practice. In most cases the error of the approximation will depend on θ , therefore it is not guaranteed that the approximated integrated likelihood maintains its shape locally around the mode and consequently confidence intervals constructed using the local curvature or the likelihood drop may be unreliable.

3.2. Approximation error for adaptive quadrature

Likelihood approximations based on *adaptive* quadrature do converge. Lemma 1 quantifies the rate of convergence for adaptive quadrature approximations to the marginal likelihood in mixed models. This intermediate technical result is required to prove convergence of the approximate maximum likelihood estimator (Theorem 2). A similar result is assumed by Bianconcini (2014) although they do not specify the region of Θ in which the uniform convergence occurs, and a stronger result about uniform convergence of derivatives of the approximate log-likelihood is required by Ogden (2017). Our proof is self-contained and makes use of suitably upgraded versions of a number of technical lemmas recently provided by Bilodeau, Stringer and Tang (2024a). A slight loosening of the usual error rates compared to results obtained in Bianconcini (2014) and Ogden (2017) is required for uniformity of the approximation error to hold. Given that the uniformity is assumed and not shown in these previous works, it is possible that their rates are too optimistic for the mixed models considered at present.

We define $\zeta_i = n_i^{-\alpha}$ for $0 < \alpha < 1/4$, $n_{\min} = \min\{n_i : i = 1, \dots, m\}$, and $\zeta_N = n_{\min}^{-\alpha}$. We let the number of groups $m = n_{\min}^q$ for some $q > 0$, so that as $N \rightarrow \infty$, $m \rightarrow \infty$ as well. The radius ζ_N will define the shrinking region in the parameter space in which all our statements about uniform convergence

hold. The precise rate of shrinkage α is chosen to balance the concentration of the likelihood with the convergence to zero of the integration error. We also define a fixed neighbourhood of arbitrary radius $\delta > 0$, and a point $\theta_* \in \Theta$ around which the likelihood concentrates. This may be intuitively thought of as a “true” value of θ , and under weak conditions will be the point that maximizes the expected log-likelihood. We emphasize that at no point do we assume the model is correctly specified in the sense that P_N^* is recovered by $\pi(\theta; \mathbf{y})$ for any $\theta \in \Theta$.

Lemma 1 upgrades the main result of Bilodeau, Stringer and Tang (2024a) to hold uniformly at a slightly loosened rate.

Lemma 1. Fix $k \in \mathbb{N}$, let $\tilde{\pi}^{\alpha_Q}(\mathbf{y}; \theta)$ be the approximation of Eq. (2.5) with a quadrature rule satisfying $\mathcal{P}(k, d)$ and $m = n_{\min}^q$ for any $q > 0$. Then, under Assumptions 2 to 7 in Appendix A there exists $C > 0$ not depending on m such that

$$\lim_{n_{\min} \rightarrow \infty} P_N^* \left(\sup_{\theta \in \mathcal{B}_{\theta_*}(\zeta_N)} |\log \tilde{\pi}^{\alpha_Q}(\mathbf{y}; \theta) - \log \pi(\mathbf{y}; \theta)| < C \sum_{i=1}^m n_i^{-\lfloor (k+2)/3 \rfloor + \varepsilon} \right) = 1,$$

for every $\varepsilon > 0$.

Proof. See Appendix B. □

In general, the constant C depends on the size of the higher-order derivatives of the likelihood function in a neighbourhood around the data-generating parameter and an error term introduced by a truncation argument used in Bilodeau, Stringer and Tang (2024a). However, with Assumptions 2 to 7 we assume that the behaviour of the likelihood function is uniform across all groups, even as m increases, which results in an bound which is independent of i in Lemma 1. This uniformity does worsen the error bound by some sub-polynomial factor, i.e. it grows slower than any polynomial function in n_i , which results in the inclusion of the ε factor in Lemma 1.

3.3. Convergence of the approximate maximum marginal likelihood estimator

Let $\hat{\theta} = \arg \max_{\theta \in \Theta} \log \pi(\mathbf{y}; \theta)$ and $\tilde{\theta} = \arg \max_{\theta \in \Theta} \log \tilde{\pi}(\mathbf{y}; \theta)$. The “exact” MLE, $\hat{\theta}$, exists in theory and Jiang, Wand and Bhaskaran (2021) provide rigorous asymptotic convergence theory for it as an estimator of θ . If $\hat{\theta}$ could be computed in practice then it would be used for inference about θ . However, $\hat{\theta}$ is intractable because $\pi(\mathbf{y}; \theta)$ is. In applications inferences are based on $\tilde{\theta}$ and it is the statistical properties of this approximate maximum likelihood estimator that are of direct interest to practice. In Theorem 2 we quantify the additional error incurred by maximizing an approximate likelihood instead of the true likelihood. The proof technique is to use Lemma 1 to bound the distance between $\hat{\theta}$ and $\tilde{\theta}$ and then use this to express the asymptotic behaviour of $\tilde{\theta}$ in terms of that of both $\hat{\theta}$ and $\tilde{\pi}(\mathbf{y}; \theta)$ as provided by Jiang, Wand and Bhaskaran (2021) and Lemma 1, respectively.

Computing $\tilde{\theta}$ must generally be done using numerical optimization. While this is not the focus of the present analysis, we do require the mild assumption that the optimization algorithm return a finite answer with probability tending to 1 as more data are obtained. We formalize this in Assumption 1:

Assumption 1. $\lim_{n_{\min} \rightarrow \infty} P_N^* [\tilde{\theta} \in \mathcal{B}_{\theta_*}(\delta)] = 1$ for the δ chosen in Assumptions 2 to 7 and for $m = n_{\min}^q$ for some $q > 0$.

We note that Assumption 1 is so mild as to have been made implicitly in all previous work on this topic as it is required in order to invoke any regularity conditions on the likelihood evaluated at $\theta = \tilde{\theta}$. It is the only assumption that we do not formally verify in Section 4 for exponential family models. If it does not hold, then there is some non-zero probability that no inferences can be made at all as more data are obtained, and this would preclude further convergence analysis.

Assumptions 1, 2, 3, 6 and 7 are sufficient to show the much stronger result that $\tilde{\theta}$ is consistent for θ . This is formalized in Lemma 2:

Lemma 2. *Under Assumptions 1, 2, 3, 6 and 7,*

$$\lim_{n_{\min} \rightarrow \infty} P_N^* [\tilde{\theta} \in \mathcal{B}_{\theta_*}(\zeta_N)] = 1.$$

Proof. Note that

$$\tilde{\pi}^{\text{aQ}}(y_i; \theta) = |L_i^\theta| \sum_{z \in Q} \pi(y_i, L_i^\theta z + \hat{u}_i^\theta; \theta) \omega(z),$$

and by definition of maxima and the positivity of the likelihood function,

$$\begin{aligned} |L_i^\theta| \sum_{z \in Q} \pi(y_i, L_i^\theta z + \hat{u}_i^\theta; \theta) \omega(z) &\leq \max_{z \in Q} \omega(z) |L_i^\theta| |Q| \pi(y_i, \hat{u}_i^\theta; \theta), \\ |L_i^{\theta_*}| \sum_{z \in Q} \pi(y_i, L_i^{\theta_*} z + \hat{u}_i^{\theta_*}; \theta) \omega(z) &\geq \omega(0) |L_i^{\theta_*}| \pi(y_i, \hat{u}_i^{\theta_*}; \theta). \end{aligned}$$

Using the above inequalities, we show the difference of the logarithms of the approximate marginal likelihood for $\theta \in \mathcal{B}_{\theta_*}(\delta) \cap \mathcal{B}_{\theta_*}(\zeta_N)^C$ is negative, implying the maxima lies in $\mathcal{B}_{\theta_*}(\zeta_N)$. Fix $\theta \in \mathcal{B}_{\theta_*}(\delta) \cap \mathcal{B}_{\theta_*}(\zeta_N)^C$. Consider the following upper bound:

$$\begin{aligned} &\log(\tilde{\pi}^{\text{aQ}}(y_i; \theta)) - \log(\tilde{\pi}^{\text{aQ}}(y_i; \theta_*)) \\ &\leq \underbrace{\left\{ \log(|L_i^\theta|) - \log(|L_i^{\theta_*}|) \right\}}_{\mathcal{A}_1(\theta)} + \underbrace{\left\{ \log(|Q| \max_{z \in Q} \omega(z)) - \log(\omega(0)) \right\}}_{\mathcal{A}_2(\theta)} \\ &\quad + \underbrace{\left\{ \log(\pi(y_i, \hat{u}_i^\theta; \theta)) - \log(\pi(y_i, \hat{u}_i^{\theta_*}; \theta_*)) \right\}}_{\mathcal{A}_3(\theta)}. \end{aligned}$$

By Assumption 5 the pair $(\theta, \hat{u}_i^\theta) \in \mathcal{B}_{\theta_*}(\delta) \times \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)$. We now show $\mathcal{A}_3(\theta) < 0$ and that $\mathcal{A}_1(\theta)$ and $\mathcal{A}_2(\theta)$ are comparatively negligible in the limit. Observe that $\mathcal{A}_2(\theta)$ does not depend on N nor i . For $\mathcal{A}_1(\theta)$ we have for all $\varepsilon > 0$

$$\mathcal{A}_1(\theta) \leq \bar{\eta} p \log(n_i) n_i^\varepsilon / \underline{\eta}$$

by Assumption 3. For $\mathcal{A}_3(\theta)$:

$$\mathcal{A}_3(\theta) = \left\{ \log(\pi(y_i, \hat{u}_i^\theta; \theta)) - \log(\pi(y_i, \mathbf{u}_{*,i}; \theta_*)) + \log(\pi(y_i, \mathbf{u}_{*,i}; \theta_*)) - \log(\pi(y_i, \hat{u}_i^{\theta_*}; \theta_*)) \right\},$$

where we added and subtracted the term $\log(\pi(y_i, u_{*,i}; \theta_*))$. We have

$$\begin{aligned} & \frac{1}{n_i} \left\{ \log(\pi(y_i, u_{*,i}; \theta_*)) - \log(\pi(y_i, \hat{u}_i^{\theta_*}; \theta_*)) \right\} \\ &= -\frac{1}{n_i} (u_{*,i} - \hat{u}_i^{\theta_*})^\top \partial_u^2 \ell_i^\theta(u^\circ) (u_{*,i} - \hat{u}_i^{\theta_*}) \\ &\leq \bar{\eta} n_i^\varepsilon \|u_{*,i} - \hat{u}_i^{\theta_*}\|_2^2 \leq \beta' \bar{\eta} n_i^\varepsilon \zeta_N^2, \end{aligned}$$

where $u^\circ = (1 - t_1)u_{*,i} + t_1 \hat{u}_i^{\theta_*}$ for some $t_1 \in [0, 1]$, and the inequalities hold with probability tending to 1 by Assumptions 3 and 6. Further,

$$\begin{aligned} & \left\{ \log(\pi(y_i, \hat{u}_i^\theta; \theta)) - \log(\pi(y_i, u_{*,i}; \theta_*)) \right\} \\ &= (\theta - \theta_*, \hat{u}_i^\theta - u_{*,i})^\top \partial_{(\theta, u)} \log(\pi(y_i, u_{*,i}; \theta_*)) \\ &+ (\theta - \theta_*, \hat{u}_i^\theta - u_{*,i})^\top \partial_{(\theta, u)}^2 \log(\pi(y_i, u^\circ; \theta^\circ)) (\theta - \theta_*, \hat{u}_i^\theta - u_{*,i}), \end{aligned}$$

where $\theta^\circ = (1 - t_2)\theta_* + t_2\theta$ for some $t_2 \in [0, 1]$. For the first term, observe:

$$\begin{aligned} & \left| (\theta - \theta_*, \hat{u}_i^\theta - u_{*,i})^\top \partial_{(\theta, u)} \log(\pi(y_i, u_{*,i}; \theta_*)) \right| \\ &\leq \left\| (\theta - \theta_*, \hat{u}_i^\theta - u_{*,i}) \right\|_2 \left\| \partial_{(\theta, u)} \log(\pi(y_i, u_{*,i}; \theta_*)) \right\|_2 \\ &\leq \delta(pDn_i)^{1/2+\varepsilon}, \end{aligned}$$

from Assumption 2. For the second term by Assumption 3 for any $\varepsilon > 0$

$$(\theta - \theta_*, \hat{u}_i^\theta - u_{*,i})^\top \partial_{(\theta, u)}^2 \log(\pi(y_i, u^\circ; \theta^\circ)) (\theta - \theta_*, \hat{u}_i^\theta - u_{*,i}) \leq -\underline{\eta} \|\theta - \theta_*\|_2^2 \leq -\underline{\eta} n_i^{1-\varepsilon} \zeta_N^2,$$

where we recall that we have fixed $\theta \in \mathcal{B}_{\theta_*}(\delta) \cap \mathcal{B}_{\theta_*}(\zeta_N)^C$ and hence $\|\theta - \theta_*\|_2 \geq \zeta_N$.

Finally, recall that for every $i = 1, \dots, m$, $\zeta_N \geq n_i^{-\alpha}$ for some $0 < \alpha < 1/4$. In this case as ε is arbitrary we have $n_i^{1-\varepsilon} \zeta_N^2 > n_i^{1/2+\varepsilon'}$ for some $\varepsilon' > \varepsilon$, and hence for n_i large enough,

$$\mathcal{A}_1(\theta) + \mathcal{A}_2(\theta) + \mathcal{A}_3(\theta) < \frac{\bar{\eta} p n_i^\varepsilon \log n_i}{\underline{\eta}} + |\mathcal{A}_2(\theta)| + \delta(pDn_i)^{1/2+\varepsilon} - \underline{\eta} n_i^{1/2+\varepsilon'} < 0,$$

where $\mathcal{A}_2$ is constant in n_i and θ . Summing over $i = 1, \dots, m$, and noting that by our assumptions these bounds holds uniformly, we have, for $n_{\min}$ large enough,

$$\log(\tilde{\pi}^{\text{AQ}}(y; \theta)) - \log(\tilde{\pi}^{\text{AQ}}(y; \theta_*)) < 0$$

for arbitrary $\theta \in \mathcal{B}_{\theta_*}(\delta) \cap \mathcal{B}_{\theta_*}(\zeta_N)^C$. But by definition of $\tilde{\theta}$, we have $\log(\tilde{\pi}^{\text{AQ}}(y; \tilde{\theta})) \geq \log(\tilde{\pi}^{\text{AQ}}(y; \theta_*))$, and hence $\tilde{\theta} \in \mathcal{B}_{\theta_*}(\zeta_N) \cup \mathcal{B}_{\theta_*}(\delta)^C$. However by Assumption 1, $\tilde{\theta} \in \mathcal{B}_{\theta_*}(\delta)$, so we conclude that $\tilde{\theta} \in \mathcal{B}_{\theta_*}(\zeta_N)$. This completes the proof. $\square$

With Lemma 1 and Lemma 2 available we are in a position to fully characterize the statistical properties of θ in Theorem 2.

Theorem 2. Fix $k \in \mathbb{N}$. Then, under Assumptions 1 and 2 to 7, if there exists a sequence of vectors $r_N^* \rightarrow \infty$ such that as $n_{\min} \rightarrow \infty$,

$$r_N^* \cdot (\theta_* - \widehat{\theta}) \stackrel{d}{=} \mathbf{Z} + o_p(1),$$

for a random variable $\mathbf{Z}$ (where the multiplication is component-wise), then

$$r_N^* \cdot (\theta_* - \widetilde{\theta}) \stackrel{d}{=} \mathbf{Z} + o_p(1) + O_p \left\{ r_N^* n_{\min}^{-\lfloor (k+2)/3 \rfloor + 1) / 2 + \varepsilon} \right\},$$

for every $\varepsilon > 0$.

Proof. Define $\gamma^*(k) = n_{\min}^{-\lfloor (k+2)/3 \rfloor + \varepsilon}$. For notational convenience, define $\ell(\mathbf{y}; \theta) = \log \pi(\mathbf{y}; \theta)$ and $\widetilde{\ell}(\mathbf{y}; \theta) = \log \widetilde{\pi}(\mathbf{y}; \theta)$. By definition of $\widetilde{\theta}$,

$$\begin{aligned} \ell(\mathbf{y}; \widehat{\theta}) - \ell(\mathbf{y}; \widetilde{\theta}) &= \ell(\mathbf{y}; \widehat{\theta}) - \widetilde{\ell}(\mathbf{y}; \widehat{\theta}) + \widetilde{\ell}(\mathbf{y}; \widehat{\theta}) - \widetilde{\ell}(\mathbf{y}; \widetilde{\theta}) + \widetilde{\ell}(\mathbf{y}; \widetilde{\theta}) - \ell(\mathbf{y}; \widetilde{\theta}) \\ &\leq \ell(\mathbf{y}; \widehat{\theta}) - \widetilde{\ell}(\mathbf{y}; \widehat{\theta}) + \widetilde{\ell}(\mathbf{y}; \widetilde{\theta}) - \ell(\mathbf{y}; \widetilde{\theta}) \\ &\leq 2 \sup_{\theta \in \mathcal{B}_{\theta_*}(\zeta_N)} \left| \ell(\mathbf{y}; \theta) - \widetilde{\ell}(\mathbf{y}; \theta) \right|, \end{aligned}$$

where the last step follows by the assumption of the theorem that $\widehat{\theta} \xrightarrow{p} \theta_*$ and by Lemma 2. Thus, Lemma 1 implies that there exists $C > 0$ for which

$$\lim_{n_{\min} \rightarrow \infty} P_N^* \left(\ell(\mathbf{y}; \widehat{\theta}) - \ell(\mathbf{y}; \widetilde{\theta}) < C m \gamma^*(k) \right) = 1. \quad (3.2)$$

Further, by a first-order Taylor expansion there exists $\alpha \in [0, 1]$ such that if $\theta^\circ = \alpha \widehat{\theta} + (1 - \alpha) \widetilde{\theta}$,

$$\ell(\mathbf{y}; \widehat{\theta}) - \ell(\mathbf{y}; \widetilde{\theta}) = -\frac{1}{2} (\widehat{\theta} - \widetilde{\theta})^\top \partial_\theta^2 \ell(\mathbf{y}; \theta^\circ) (\widehat{\theta} - \widetilde{\theta}). \quad (3.3)$$

Apply Lemma 2 and again use the assumption that $\widehat{\theta} \xrightarrow{p} \theta_*$ to conclude that

$$\lim_{n_{\min} \rightarrow \infty} P_N^* \{ \theta^\circ \in \mathcal{B}_{\theta_*}(\zeta_N) \} = 1.$$

Therefore, by Assumption 3 and Eq. (3.3),

$$\ell(\mathbf{y}; \widehat{\theta}) - \ell(\mathbf{y}; \widetilde{\theta}) \geq \frac{1}{2} \|\widehat{\theta} - \widetilde{\theta}\|^2 \inf_{\theta \in \mathcal{B}_{\theta_*}(\zeta_N)} \lambda_p(-\partial_\theta^2 \ell(\mathbf{y}; \theta)),$$

so by Lemma 9 in Appendix C,

$$\lim_{n_{\min} \rightarrow \infty} P_N^* \left(\ell(\mathbf{y}; \widehat{\theta}) - \ell(\mathbf{y}; \widetilde{\theta}) \geq \frac{1}{2} \|\widehat{\theta} - \widetilde{\theta}\|^2 \cdot \underline{\eta}' \sum_{i=1}^m n_i^{1-\varepsilon} \right) = 1, \quad (3.4)$$

for some $\underline{\eta}' > 0$. Since

$$\frac{m \gamma^*(k)}{\sum_{i=1}^m n_i^{1-\varepsilon}} = \frac{\gamma^*(k)}{\frac{1}{m} \sum_{i=1}^m n_i^{1-\varepsilon}} \leq \frac{\gamma^*(k)}{n_{\min}^{1-\varepsilon}} = n_{\min}^{-\lfloor (k+2)/3 \rfloor - 1 + 2\varepsilon},$$

combining Eqs. (3.2) and (3.4) implies that there exists $C' > 0$ such that

$$\lim_{n_{\min} \rightarrow \infty} P_N^* \left\{ \|\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}\|_2 \leq C' n_{\min}^{(-\lfloor (k+2)/3 \rfloor - 1)/2 + \varepsilon} \right\} = 1. \quad (3.5)$$

By the triangle inequality and Eq. (3.5) we therefore have

$$r_N^* \cdot (\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_*) \stackrel{d}{=} \mathbf{Z} + o_p(1) + O_p \left\{ r_N^* n_{\min}^{(-\lfloor (k+2)/3 \rfloor - 1)/2 + \varepsilon} \right\},$$

for any $\varepsilon > 0$. □

The convergence rate, r_N^* , of the true MLE can be found in Jiang, Wand and Bhaskaran (2021) for GLMMs. Under the conditions that $m, n_{\min} \rightarrow \infty$ with $n_{\min}/m \rightarrow 0$, r_N^* has elements equal to $m^{1/2}$ or $(mn_{\min})^{1/2}$ depending on the exact structure of the linear predictor. The limiting random variable $\mathbf{Z}$ is Gaussian with a tractable variance matrix and Wald confidence intervals are formed in the usual manner with marginal variances obtained from a studentized pivot based on this limiting Gaussian distribution. We elaborate on the application of Theorem 2 to exponential family generalized linear mixed models in Section 4.

We emphasize that $\widetilde{\boldsymbol{\theta}}$ is a different estimator for $\boldsymbol{\theta}$ for each different k . Theorem 2 gives the relationship that is needed between m , $n_{\min}$ and k to yield asymptotically valid confidence intervals for $\boldsymbol{\theta}$ based on $\widetilde{\boldsymbol{\theta}}$, and hence provides guidance on which k —and hence which estimator of $\boldsymbol{\theta}$ —should be chosen. Specifically, k should be chosen large enough to ensure that the additional error term decreases as m and n increase. Since (a) these are asymptotic upper bounds depending on unknown constants and (b) in practice m and n are fixed for a particular set of data the usual practical advice is to choose k large enough such that inferences stop changing when k is further increased. Both Theorem 2 and the empirical results given in Subsection 4.3 support this strategy.

4. Generalized linear mixed models

4.1. Exponential family models

A generalized linear mixed model (GLMM) is:

$$\begin{aligned} y_{ij} \mid \mathbf{x}_{ij}, \mathbf{v}_{ij}, \mathbf{u}_i &\stackrel{\text{ind.}}{\sim} F(\eta_{ij}), \\ \eta_{ij} &= h(\mu_{ij}) = \beta_0 + \mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{v}_{ij}^T \mathbf{u}_i, \\ \mathbf{u}_i &\stackrel{\text{ind.}}{\sim} \text{Gaussian} \{ \mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\sigma}) \}. \end{aligned} \quad (4.1)$$

Here $h : \mathbb{R} \rightarrow \mathbb{R}$ is a known link function, $\boldsymbol{\Sigma}(\boldsymbol{\sigma})$ is a covariance matrix depending on unknown parameters $\boldsymbol{\sigma}$. The distribution F belongs to a natural exponential family with mean μ_{ij} and density

$$f(y_{ij} \mid \eta_{ij}) = \exp \left\{ \frac{y_{ij} \eta_{ij} - b(\eta_{ij})}{a(\boldsymbol{\phi})} + c(y; \boldsymbol{\phi}) \right\}, \quad (4.2)$$

for functions $a(\cdot)$, $b(\cdot)$, $c(\cdot)$ and any $y \in \mathcal{S}$ where $\mathcal{S} \subseteq \mathbb{R}$ is the sample space. For simplicity we treat the dispersion parameter $\boldsymbol{\phi}$ as fixed and known and assume $0 < a(\boldsymbol{\phi}) < \infty$. Observe that Eq. (4.1) is obtained from Eq. (2.1) where F is chosen to have density given by Eq. (4.2) and $G = \text{Gaussian} \{ \mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\sigma}) \}$.

Proposition 1 establishes that Theorems 1 and 2 and Lemma 1 apply to any GLMM that is non-degenerate and well specified; see Assumptions 8 and 9 in Appendix D. These conditions are very mild and all the “common” exponential family distributions are permitted including Gaussian, Binomial, Poisson, Gamma, and Negative Binomial.

Proposition 1. *If the generalized linear mixed model defined by Eq. (4.1) satisfies Assumptions 8 and 9 in Appendix D then it satisfies Assumptions 2 to 7 in Appendix A.*

Proof. See Appendix D in the Supplementary Materials. $\square$

4.2. Exact maximum likelihood estimator rates

Proposition 1 states that Theorem 2 applies to the GLMM defined by Eq. (4.1). In order to make quantitative use of Theorem 2 the convergence rates r_N^* of the components of the exact MLE must be specified. Jiang, Wand and Bhaskaran (2021) provide a thorough analysis of this topic for GLMMs. Our Theorem 2 complements their analysis to account for integration error, making explicit the error rates enjoyed by the approximate MLEs in GLMMs which are the quantities upon which inferences are based in practice.

The convergence rates for elements of $\hat{\theta}$ in a GLMM are different for variance components and for regression coefficients that share a covariate with a random effect compared to ones that do not. Let $\sigma = \text{vech}\{\Sigma(\sigma)\}$ and consider Eq. (4.1). For simplicity let $d = 1$ and take $v_{ij} = 1$, the random intercepts model. In the notation of the present paper Theorem 1 of Jiang, Wand and Bhaskaran (2021) implies that for fixed matrix $V \in \mathbb{R}^p$,

$$\sqrt{m} \begin{pmatrix} \hat{\beta}_0 - \beta_0^* \\ \sqrt{n}(\hat{\beta} - \beta^*) \\ \hat{\sigma} - \sigma^* \end{pmatrix} \xrightarrow{d} \text{Gaussian}(\mathbf{0}, V),$$

where $\hat{\theta} = (\hat{\beta}_0, \hat{\beta}^T, \hat{\sigma}^T)^T$ is the exact MLE and θ_* the corresponding “true value” of θ from Assumption 9. Corollary 2 states the error rate of $\hat{\theta}$, the quantity upon which inferences about θ are based in practice.

Corollary 2. *Assume that the covariates x_{ij} and v_{ij} are independent and identically distributed and satisfy Assumption 9 iv) with probability tending to 1 in limit as both m and $n_{\min} \rightarrow \infty$. Further assume that Assumption 8: i–iii and v, Assumption 9, and Assumption 1–3 from Jiang, Wand and Bhaskaran (2021) hold. Under these conditions the model given by Eq. (4.1) satisfies*

$$\sqrt{m} \begin{pmatrix} \tilde{\beta}_0 - \beta_0^* \\ \sqrt{n}(\tilde{\beta} - \beta^*) \\ \tilde{\sigma} - \sigma^* \end{pmatrix} \stackrel{d}{=} Z + o_p(1) + \begin{pmatrix} O_p(m^{1/2} n_{\min}^{(-\lfloor (k+2)/3 \rfloor - 1)/2 + \varepsilon}) \\ O_p(m^{1/2} n_{\min}^{(-\lfloor (k+2)/3 \rfloor)/2 + \varepsilon}) \\ O_p(m^{1/2} n_{\min}^{(-\lfloor (k+2)/3 \rfloor - 1)/2 + \varepsilon}) \end{pmatrix},$$

for all $\varepsilon > 0$ where $Z \sim N(0, V)$ for the positive definite matrix V given in Theorem 1 of Jiang, Wand and Bhaskaran (2021).

Proof. By Proposition 1, Assumptions 8 and 9 imply Assumptions 2 to 7. Further, Jiang, Wand and Bhaskaran (2021, Theorem 1) provides the asymptotics of $\hat{\theta}$ that are required by Theorem 2. $\square$

Remark 1. [Jiang, Wand and Bhaskaran \(2021\)](#) assume a random design for the covariates, and we have reflected this in the statement of the Corollary 2. This change from fixed design to random design requires a reformulation of the eigenvalue condition in Assumption 8 iv) as a probabilistic statement in Corollary 2.

Remark 2. For certain GLMs models (such as the logistic model) the cost of uniformity in the integration error can be sharpened to logarithmic factor instead of the sub-polynomial factor for general GLMs. For example, with more detailed accounting for the logistic model one can obtain a statement of the kind:

$$\sqrt{m} \begin{pmatrix} \tilde{\beta}_0 - \beta_0^* \\ \sqrt{n}(\tilde{\beta} - \beta^*) \\ \tilde{\sigma} - \sigma^* \end{pmatrix} \stackrel{d}{=} \mathbf{Z} + o_p(1) + \begin{pmatrix} O_p(m^{1/2} \log(n_{\min})^a n_{\min}^{(-\lfloor (k+2)/3 \rfloor - 1)/2}) \\ O_p(m^{1/2} \log(n_{\min})^a n_{\min}^{(-\lfloor (k+2)/3 \rfloor)/2}) \\ O_p(m^{1/2} \log(n_{\min})^a n_{\min}^{(-\lfloor (k+2)/3 \rfloor - 1)/2}) \end{pmatrix},$$

for some $a \in \mathbb{R}$.

[Jiang, Wand and Bhaskaran \(2021, Assumption 1 and 2\)](#) requires $n_{\min}/m \rightarrow 0$ as $m, n_{\min} \rightarrow \infty$, meaning that m should grow faster than $n_{\min}$. This is satisfied by $m = n_{\min}^q$ for any $q > 1$ meaning that m can grow arbitrarily faster than $n_{\min}$ and convergence of the exact MLE is still attained. Corollary 2 reveals that there is a limit to how fast m can grow compared to n when approximate integration is required to compute the MLE. Specifically, $q < r(k)$ or $q < r(k) + 1$ where $r(k) = \lfloor (k+2)/3 \rfloor$ is required for the integration error to be negligible for $\tilde{\beta}$ and $\tilde{\beta}_0, \tilde{\sigma}$ respectively. If m grows too fast compared to $n_{\min}$ the integration error will degrade the quality of the inferences about θ based on $\tilde{\theta}$. The solution is to increase k and hence $r(k)$, i.e. to use a more accurate integral approximation in the case that m is too large relative to $n_{\min}$. We illustrate this empirically in Subsection 4.3. We reiterate that the Laplace approximation is AQ with $k = 1$ so this discussion also applies to the question of when to use or not use the Laplace approximation to fit generalized linear mixed models.

4.3. Empirical error analysis

A practitioner will be faced with a fixed m and n but has control over k . The practical recommendation based on the theory presented in this paper is to choose k high enough that $\tilde{\theta}$ and approximate confidence intervals based on it do not change when k is increased or decreased. Here we present an empirical analysis which demonstrates the impact of Theorem 2 and its Corollary 2. The core idea is that if k is chosen large enough inferences based on $\tilde{\theta}$ should be indistinguishable from those based on the exact MLE. Specifically, Wald confidence intervals should attain close to nominal coverage on average. However, if k is chosen too low for a given m and n then the quality of the inferences should degrade as m is increased. Since the theory for the exact MLE predicts that the quality of inferences based on $\tilde{\theta}$ should improve on average as m is increased, this contradictory behaviour can be attributed to the increasing integral approximation error incurred as $m \rightarrow \infty$.

A reviewer points out that Theorem 2 does not make a statement about the quality of the Hessian of the approximate log-marginal likelihood as an approximation to the Hessian of the exact log-likelihood, and that this would be required to make a formal statement about the coverage of Wald intervals based on the former. Indeed, [Ogden \(2017\)](#) shows that an accurate Hessian approximation is sufficient for accurate Wald intervals in this context. However, the following simulations show empirical evidence of a setup in which any error in the Hessian is not large in comparison to that in the approximate

likelihood and the behaviour predicted by Theorem 2 is recovered for Wald confidence intervals based on the Hessian of the approximate log-marginal likelihood.

We construct simulations to investigate how this expected behaviour depends on m , n , and k . We simulate 1000 sets of data from the model

$$y_{ij} \mid x_{ij}, u_i \stackrel{\text{ind.}}{\sim} \text{Ber} \left\{ \log \left(\frac{\eta_{ij}}{1 - \eta_{ij}} \right) \right\}, \quad (4.3)$$

$$\eta_{ij} = (\beta_0 + u_i) + x_{ij}\beta_1, \quad (4.4)$$

$$u_i \stackrel{\text{ind.}}{\sim} \text{Gaussian} \left(0, \sigma^2 \right), \quad (4.5)$$

with $i = 1, \dots, m$ groups of size $j = 1, \dots, n$. The covariate x_{ij} was generated from $N(0, 1)$ and hence varied within and between groups. The parameters were $(\beta_0, \beta_1, \sigma) = (-4, 2, 2)$, leading to very imbalanced binary outcomes with $P(Y = 1)$ ranging from 0.18% for $x = 0$ and u at its 1% percentile to 58% for $x = 1$ and u at its 99% percentile, and $P(Y = 1) = 1.8\%$ and 12% for $x = 0$ and 1 and u at its mean of 0. The numbers of groups were very large at $m = 1000 \times (2^0, \dots, 2^4)$, and the group sizes small at $n = 2, 4, 6, 8, 10$. These were chosen so that m grows faster than n as required by Jiang, Wand and Bhaskaran (2021) and the integration error should dominate the sampling error for lower n and higher m if k is chosen too small. Overall, this is a simulation setup in which it should be challenging to make accurate approximate inferences.

Fig. 1 shows the results. When k is too low increasing n leads to approximate Wald confidence intervals that have worse coverage. The effect is less dramatic for larger n . However, once k is increased large enough the coverages remain nominal as m is increased indicating that the integration error is of a lower order than the sampling error. This supports the practical recommendation of simply increasing k until inferences stop changing as this is likely to be the point at which numerical error is less than the sampling error. Additional simulation results are shown in Appendix E.

5. Discussion and extensions

Take $m = 1$ and consider the following general latent variable model as a special case of Eq. (2.1):

$$\begin{aligned} y_i \mid x_i, u &\stackrel{\text{ind.}}{\sim} F(x_i, u; \beta, \phi), i = 1, \dots, n, \\ u &\sim G(\sigma). \end{aligned} \quad (5.1)$$

This covers a wide range of interesting models including penalized splines (Wood, 2016) and spatio-temporal models (Diggle et al., 2013). Theorem 2 applies with $m = 1$ and r_N^* depending only on n , to give the asymptotic behaviour of $\tilde{\theta}$ when that for $\hat{\theta}$ is available. However, in these interesting problems the asymptotics for $\hat{\theta}$ require the dimension of u to increase with n and this situation is not covered by our present analysis. Extension of Theorem 2 and the other results in this paper for Laplace-approximate marginal likelihood inference in the presence of such high-dimensional latent variables u appears feasible and the subject of ongoing work. The main technical challenge in proving Theorem 2 was to obtain the rate of local uniform convergence in probability of the likelihood approximation. Tang and Reid (2024) provide local uniform rates of convergence for Laplace approximate marginal likelihood in high dimensions, showing that if $\dim(u) = q$ then the approximation error is $O(q^3 \log(n)/n)$, implying that the likelihood approximation converges whenever $q = O(n^\alpha)$ for any $\alpha < 1/3$. Kauermann, Krivobokova and Fahrmeir (2009) shows that for spline models, $q = 2/9 < 1/3$ is sufficient for asymptotic convergence of the exact maximum likelihood estimator. Sanz-Alonso and Yang (2022) discuss

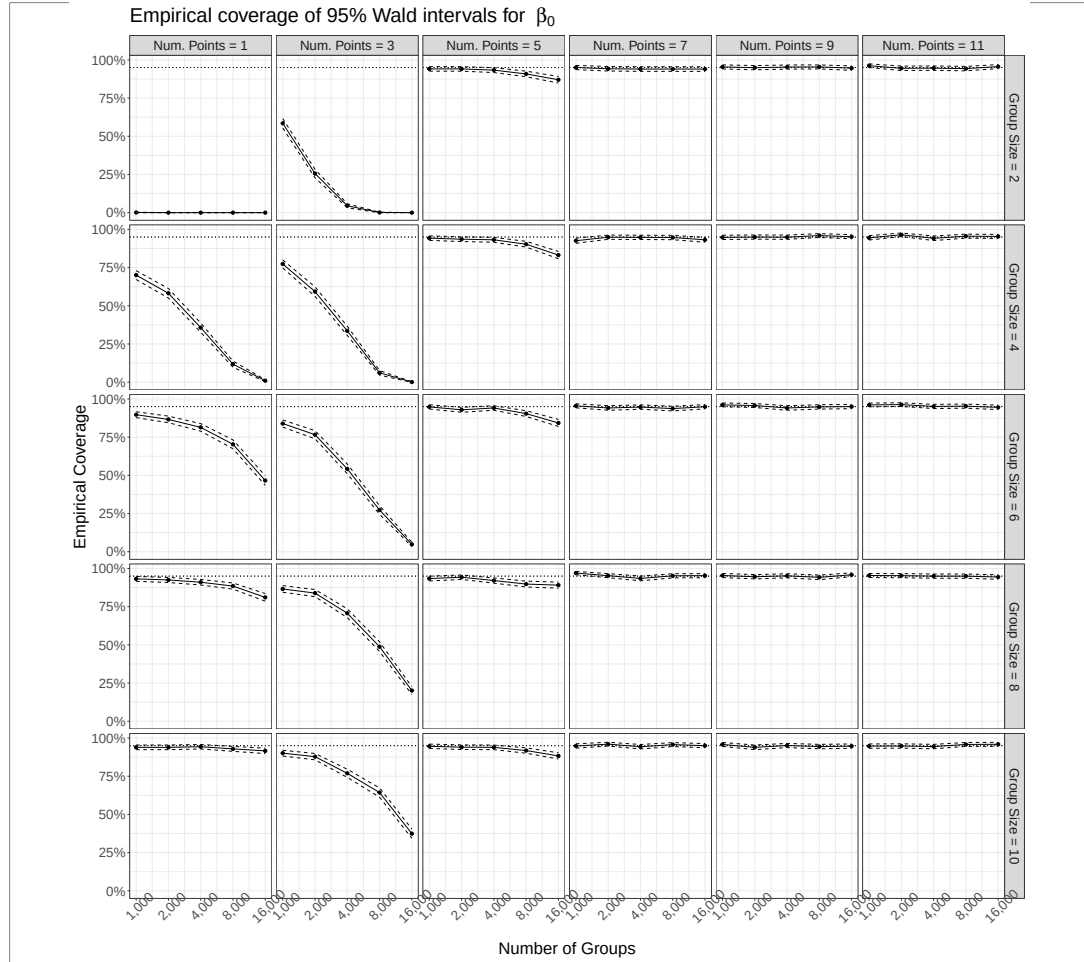


Figure 1. Empirical coverage of β_0 for the simulation of Subsection 4.3. A Bernoulli random intercept model (Eq. (4.3)) was fit 1000 times to each combination of numbers of groups (m , x-axis), group size (n , rows) and numbers of quadrature points (k , columns). The y-axis shows empirical coverages of 95% Wald confidence intervals centred at the approximate MLE, $\tilde{\theta}$, with standard errors computed using the diagonal of the inverse Hessian of the approximate marginal likelihood. Shown are empirical coverage proportions across 1000 simulations (•) along with Monte Carlo confidence intervals (---).

convergence rates and scaling for spatial Gaussian processes fit using finite element approximations which are considerably more complicated. Extension of Theorem 2 to the case where $\dim(\mathbf{u})$ increases with n via these novel results could lead to convergence results for Laplace-approximate maximum marginal likelihood estimators in these and related problems.

Appendix A: Assumptions

First, we inherit some standard notation from [Bilodeau, Stringer and Tang \(2024a\)](#). For a positive-definite $p \times p$ matrix A let $\lambda_1(A) \geq \dots \geq \lambda_p(A) > 0$ denote its ordered eigenvalues. For a generic

$m \times n$ matrix B we use $\|B\|_{op}$ to denote its maximal singular value, i.e. its operator norm. For any $f : \mathbb{R}^p \rightarrow \mathbb{R}$, $\alpha \subseteq \mathbb{N}^p$, and $\mathbf{x} \in \mathbb{R}^p$ we define

$$|\alpha| = \sum_{j=1}^p \alpha_j, \quad \alpha! = \prod_{j=1}^p \alpha_j!, \quad \mathbf{x}^\alpha = \mathbf{x}_\alpha = \prod_{j=1}^p x_j^{\alpha_j}, \text{ and}$$

$$\partial^\alpha f(\mathbf{x}) = \partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \cdots \partial x_p^{\alpha_p} f(\mathbf{x}) = \frac{\partial^{|\alpha|} f(\mathbf{x})}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \cdots \partial x_p^{\alpha_p}}.$$

For any vector $\mathbf{z}$ and radius $\delta > 0$ let $\mathcal{B}_z(\delta) = \{\mathbf{z}' : \|\mathbf{z} - \mathbf{z}'\|_2 < \delta\}$ where $\|\cdot\|_2$ denotes Euclidean norm.

In the Assumptions that follows let $m = n_{\min}^q$ for some $q > 0$. For any sequence of data-generating distributions P_N^* , we say Assumptions 2 to 7 hold if there exist $\delta > 0$, $0 < \alpha < 1/4$, $\theta_* \in \Theta$, and $\mathbf{u}_{*,i} \in \mathbb{R}^d, i = 1, \dots, m$ such that for $\zeta_N = n_{\min}^{-\alpha}$ as defined in Section 3 each of the following six statements holds for all values of $\epsilon > 0$.

Assumption 2. *There exist $t, D > 0$ such that for all $\alpha \subseteq \mathbb{N}^p$ with $0 \leq |\alpha| \leq t$,*

$$\lim_{n_{\min} \rightarrow \infty} P_N^* \left[\forall i=1, \dots, m \sup_{\theta \in \mathcal{B}_{\theta_*}(\delta)} \sup_{\mathbf{u} \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} |\partial_{\mathbf{u}}^\alpha \ell_i^\theta(\mathbf{u})| < n_i^{1+\epsilon} \cdot D \right] = 1.$$

Assumption 3. *There exist $0 < \underline{\eta} \leq \bar{\eta} < \infty$ such that for all $i \in [m]$*

$$\begin{aligned} \lim_{n_{\min} \rightarrow \infty} P_N^* \left[\forall i=1, \dots, m \ n_i^{1-\epsilon} \cdot \underline{\eta} \leq \inf_{\theta \in \mathcal{B}_{\theta_*}(\delta)} \inf_{\mathbf{u} \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} \lambda_d(-\partial_{(\theta, \mathbf{u})}^2 \ell_i^\theta(\mathbf{u})) \right. \\ \left. \leq \sup_{\theta \in \mathcal{B}_{\theta_*}(\delta)} \sup_{\mathbf{u} \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} \lambda_1(-\partial_{(\theta, \mathbf{u})}^2 \ell_i^\theta(\mathbf{u})) \leq n_i^{1+\epsilon} \cdot \bar{\eta} \right] = 1 \end{aligned}$$

Assumption 4. *There exists $b > 0$ such that*

$$\lim_{n_{\min} \rightarrow \infty} P_N^* \left[\forall i=1, \dots, m \sup_{\theta \in \mathcal{B}_{\theta_*}(\zeta_N)} \sup_{\mathbf{u} \in [\mathcal{B}_{\mathbf{u}_{*,i}}(\delta)]^c} \log \pi_i(\mathbf{y}_i \mid \mathbf{u}; \theta) - \log \pi_i(\mathbf{y}_i \mid \mathbf{u}_{*,i}; \theta) \leq -n_i^{1-\epsilon} b \right] = 1.$$

Assumption 5. *There exists $b' > 0$ such that*

$$\lim_{n_{\min} \rightarrow \infty} P_N^* \left[\forall i=1, \dots, m \sup_{\theta \in [\mathcal{B}_{\theta_*}(\delta)]^c} \sup_{\mathbf{u} \in [\mathcal{B}_{\mathbf{u}_{*,i}}(\delta)]^c} \log \pi_i(\mathbf{y}_i \mid \mathbf{u}; \theta) - \log \pi_i(\mathbf{y}_i \mid \mathbf{u}_{*,i}; \theta) \leq -n_i^{1-\epsilon} b' \right] = 1.$$

Assumption 6. *There exists $\beta > 0$ such that for every $\epsilon > 0$,*

$$\lim_{n_{\min} \rightarrow \infty} P_N^* \left[\forall i=1, \dots, m \sup_{\theta \in \mathcal{B}_{\theta_*}(\zeta_N)} \zeta_N^{-1} \left\| \hat{\mathbf{u}}_i^\theta - \mathbf{u}_{*,i} \right\|_2 \leq \beta n_i^\epsilon \right] = 1.$$

Furthermore, for every $\beta' > 0$ and for every function $G'(n_i)$ such that $\lim_{n \rightarrow \infty} G'(n_i) = \infty$:

$$\lim_{n_{\min} \rightarrow \infty} P_N^* \left[\forall i=1, \dots, m \ n_i^{1/2} G'(n_i)^{-1} \left\| \hat{\mathbf{u}}_i^\theta - \mathbf{u}_{*,i} \right\|_2 > \beta' n_i^\epsilon \right] = 0.$$

Assumption 7. *There exist $0 < c_1 < c_2 < \infty$ such that*

$$c_1 \leq \inf_{\theta \in \mathcal{B}_{\theta_*}(\zeta_N)} \inf_{\mathbf{u} \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} g(\mathbf{u}; \boldsymbol{\sigma}) \leq \sup_{\theta \in \mathcal{B}_{\theta_*}(\zeta_N)} \sup_{\mathbf{u} \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} g(\mathbf{u}; \boldsymbol{\sigma}) \leq c_2.$$

Furthermore, uniformly for all $i = 1, \dots, m$, for some $a \in \mathbb{R}$

$$\begin{aligned} \int_{\mathbb{R}^d} \|\partial_{\theta} \log \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta})\|_2 g(\mathbf{u}; \boldsymbol{\sigma}) d\mathbf{u} &= O(n_i^a), \\ \int_{\mathbb{R}^d} \|\partial_{\theta}^2 \log \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta})\|_{op} g(\mathbf{u}; \boldsymbol{\sigma}) d\mathbf{u} &= O(n_i^a), \end{aligned}$$

for all $\boldsymbol{\theta} \in \mathcal{B}_{\theta_*}(\zeta_N)$.

Remark 3. *We state our assumptions as limits in probability. In the proofs, for the sake of clarity and brevity, algebra which invokes these assumptions is often performed outside of a probability statement. This is understood to mean that the given statement holds for any event upon which the relevant assumption holds. The assumptions then state that the measure of the sets of events for which each statement holds tends to 1 as $n_{\min} \rightarrow \infty$.*

Appendix B: Proof of Lemma 1

Our approach is to show that the strengthening of the pointwise assumptions in Bilodeau, Stringer and Tang (2024a) to the uniform assumptions in Appendix A is sufficient to control the approximation error of the likelihood in a ball around the true parameter with rates slightly changed to reflect the additional ϵ factor in our uniform assumptions compared to their pointwise versions in Bilodeau, Stringer and Tang (2024a). For completeness we carefully verify that each step from Bilodeau, Stringer and Tang (2024a) can be appropriately upgraded to hold uniformly. In Proposition 1 we show that the stronger uniform assumptions holds for generalized linear mixed models.

To state intermediate results we require the following notation from Bilodeau, Stringer and Tang (2024a). For $b > 3$ let

$$\begin{aligned} \tau_{<b}^{(j)} &= \left\{ (t_3, \dots, t_{2k}) \in \mathbb{Z}_+^{2k-3} \mid \sum_{s=3}^{2k} t_s = j \text{ and } \sum_{s=3}^{2k} st_s \leq b-1 \right\}, \\ \tau_{=b}^{(j)} &= \left\{ (t_3, \dots, t_{2k}) \in \mathbb{Z}_+^{2k-3} \mid \sum_{s=3}^{2k} t_s = j \text{ and } \sum_{s=3}^{2k} st_s = b \right\}, \\ \tau_{\geq b}^{(j)} &= \left\{ (t_3, \dots, t_{2k}) \in \mathbb{Z}_+^{2k-3} \mid \sum_{s=3}^{2k} t_s = j \text{ and } \sum_{s=3}^{2k} st_s \geq b \right\}. \end{aligned}$$

Further, let $\tau(\mathbf{t}) = \sum_{s=3}^{2k} st_s$ for any $\mathbf{t} = (t_3, \dots, t_{2k}) \in \mathbb{Z}_+^{2k-3}$.

Fix m and $i \in [m]$. For all $\boldsymbol{\theta} \in \Theta$ and $N \in \mathbb{N}$, the following holds P_N^* -a.s.

$$\left| \frac{\pi_i(\mathbf{y}_i; \boldsymbol{\theta})}{\widetilde{\pi}_i^{\text{ao}}(\mathbf{y}_i; \boldsymbol{\theta})} - 1 \right| = \frac{|\pi_i(\mathbf{y}_i; \boldsymbol{\theta}) - \widetilde{\pi}_i^{\text{ao}}(\mathbf{y}_i; \boldsymbol{\theta})|}{\widetilde{\pi}_i^{\text{ao}}(\mathbf{y}_i; \boldsymbol{\theta})} = \frac{|\pi_i(\mathbf{y}_i; \boldsymbol{\theta}) - \widetilde{\pi}_i^{\text{ao}}(\mathbf{y}_i; \boldsymbol{\theta})|}{\pi_i(\mathbf{y}_i \mid \widehat{\mathbf{u}}_i^{\boldsymbol{\theta}}; \boldsymbol{\theta})} \frac{\pi_i(\mathbf{y}_i \mid \widehat{\mathbf{u}}_i^{\boldsymbol{\theta}}; \boldsymbol{\theta})}{\widetilde{\pi}_i^{\text{ao}}(\mathbf{y}_i; \boldsymbol{\theta})}. \quad (\text{B.1})$$

We prove the following intermediate results.

Lemma 3. Under assumptions Assumptions 2 to 4, 6 and 7, if $\mathfrak{R}(Q, \omega)$ is a quadrature rule satisfying $\mathcal{P}(k, d)$ then for all $\varepsilon > 0$ and for $m = n_{\min}^q$ for any $q > 0$:

$$\lim_{n_{\min} \rightarrow \infty} P_N^* \left(\forall_{i=1, \dots, m} \sup_{\theta \in \mathcal{B}_{\theta_*}(\zeta_N)} \frac{|\pi_i(\mathbf{y}_i; \theta) - \tilde{\pi}_i^{\mathfrak{A}_Q}(\mathbf{y}_i; \theta)|}{\pi_i(\mathbf{y}_i | \hat{\mathbf{u}}_i^\theta; \theta)} \leq D \frac{1}{n_i^{d/2 + \lfloor (k+2)/3 \rfloor - \varepsilon}} \right) = 1.$$

Lemma 4. Under assumptions Assumptions 2, 3, 6 and 7, if $\mathfrak{R}(Q, \omega)$ is a quadrature rule satisfying $\mathcal{P}(k, d)$ then for all $\varepsilon > 0$ and for $m = n_{\min}^q$ for any $q > 0$:

$$\lim_{n_{\min} \rightarrow \infty} P_N^* \left(\forall_{i=1, \dots, m} \sup_{\theta \in \mathcal{B}_{\theta_*}(\zeta_N)} \frac{\pi_i(\mathbf{y}_i | \hat{\mathbf{u}}_i^\theta; \theta)}{\tilde{\pi}_i^{\mathfrak{A}_Q}(\mathbf{y}_i; \theta)} \leq D n_i^{d/2 + \varepsilon} \right) = 1.$$

Consequently, for each $i \in [m]$, under Assumptions 2 to 4, 6 and 7 there exists a $C > 0$ such that

$$\lim_{n_{\min} \rightarrow \infty} P_N^* \left(\forall_{i=1, \dots, m} \sup_{\theta \in \mathcal{B}_{\theta_*}(\zeta_N)} \left| \frac{\tilde{\pi}_i^{\mathfrak{A}_Q}(\mathbf{y}_i; \theta)}{\pi_i(\mathbf{y}_i; \theta)} - 1 \right| < C n_i^{-\lfloor (k+2)/3 \rfloor + \varepsilon} \right) = 1.$$

for all $\varepsilon > 0$. Next, using $\log(1+x) \leq x$ for all $x > -1$ and $\log(1-x) \geq -2x$ for all $x \in [0, 3/4]$ gives that for each $i \in [m]$:

$$\lim_{n_{\min} \rightarrow \infty} P_N^* \left(\forall_{i=1, \dots, m} \sup_{\theta \in \mathcal{B}_{\theta_*}(\zeta_N)} |\log \tilde{\pi}_i^{\mathfrak{A}_Q}(\mathbf{y}_i; \theta) - \log \pi_i(\mathbf{y}_i; \theta)| < 2C n_i^{-\lfloor (k+2)/3 \rfloor + \varepsilon} \right) = 1.$$

Therefore,

$$\lim_{n_{\min} \rightarrow \infty} P_N^* \left(\sup_{\theta \in \mathcal{B}_{\theta_*}(\zeta_N)} |\log \tilde{\pi}^{\mathfrak{A}_Q}(\mathbf{y}; \theta) - \log \pi(\mathbf{y}; \theta)| < 2C \sum_{i=1}^m n_i^{-\lfloor (k+2)/3 \rfloor + \varepsilon} \right) = 1,$$

for all $\varepsilon > 0$.

B.1. Proof of Lemma 3

Fix arbitrary $\gamma > 0$ (to be tuned at the end as a function of d and k) and let $\gamma_i = \gamma \sqrt{(\log n_i)/n_i}$ for each $N \in \mathbb{N}$. First, expand the fraction of interest, giving

$$\begin{aligned} \frac{|\pi_i(\mathbf{y}_i; \theta) - \tilde{\pi}_i^{\mathfrak{A}_Q}(\mathbf{y}_i; \theta)|}{\pi_i(\mathbf{y}_i | \hat{\mathbf{u}}_i^\theta; \theta)} &= \frac{\left| \int \pi_i(\mathbf{y}_i, \mathbf{u}; \theta) d\mathbf{u} - |\mathbf{L}_i^\theta| \sum_{\mathbf{z} \in Q(d, k)} \pi_i(\mathbf{y}_i, \mathbf{L}_i^\theta \mathbf{z} + \hat{\mathbf{u}}_i^\theta; \theta) \omega_k(\mathbf{z}) \right|}{\pi_i(\mathbf{y}_i | \hat{\mathbf{u}}_i^\theta; \theta)} \\ &= g_i(\hat{\mathbf{u}}_i^\theta; \sigma) \left| \int \exp \left\{ \ell_i^\theta(\mathbf{u}) - \ell_i^\theta(\hat{\mathbf{u}}_i^\theta) \right\} d\mathbf{u} - |\mathbf{L}_i^\theta| \sum_{\mathbf{z} \in Q(d, k)} \omega_k(\mathbf{z}) \exp \left\{ \ell_i^\theta(\mathbf{L}_i^\theta \mathbf{z} + \hat{\mathbf{u}}_i^\theta) - \ell_i^\theta(\hat{\mathbf{u}}_i^\theta) \right\} \right|. \end{aligned}$$

After splitting the region of integration and applying the triangle inequality, this is upper bounded by

$$\begin{aligned} &\leq g_i(\widehat{\mathbf{u}}_i^\theta; \sigma) \left| \int_{\mathcal{B}_{\widehat{\mathbf{u}}_i^\theta}(\gamma_i)} \exp \left\{ \ell_i^\theta(\mathbf{u}) - \ell_i^\theta(\widehat{\mathbf{u}}_i^\theta) \right\} d\mathbf{u} - |\mathbf{L}_i^\theta| \sum_{\mathbf{z} \in \mathcal{Q}(d,k)} \omega_k(\mathbf{z}) \exp \left\{ \ell_i^\theta(\mathbf{L}_i^\theta \mathbf{z} + \widehat{\mathbf{u}}_i^\theta) - \ell_i^\theta(\widehat{\mathbf{u}}_i^\theta) \right\} \right| \\ &\quad + g_i(\widehat{\mathbf{u}}_i^\theta; \sigma) \int_{\left[\mathcal{B}_{\widehat{\mathbf{u}}_i^\theta}(\gamma_i) \right]^c} \exp \left\{ \ell_i^\theta(\mathbf{u}) - \ell_i^\theta(\widehat{\mathbf{u}}_i^\theta) \right\} d\mathbf{u}. \end{aligned} \quad (\text{B.2})$$

Now, define

$$\widehat{M}_i^\theta = \sup_{\alpha: |\alpha| \leq t} \sup_{\mathbf{u} \in \mathcal{B}_{\widehat{\mathbf{u}}_i^\theta}(\bar{\mathbf{z}} | \mathbf{L}_i^\theta | \vee \gamma_i)} \left| \partial^\alpha \ell_i^\theta(\mathbf{u}) \right|, \quad \bar{\eta}_i^\theta = \frac{\lambda_1(\mathbf{H}_i^\theta(\widehat{\mathbf{u}}_i^\theta))}{n}, \quad \text{and} \quad \underline{\eta}_i^\theta = \frac{\lambda_d(\mathbf{H}_i^\theta(\widehat{\mathbf{u}}_i^\theta))}{n},$$

where $\bar{\mathbf{z}} = \sup_{\mathbf{z} \in \mathcal{Q}} \|\mathbf{z}\|_2 < \infty$.

The main result we need to inherit from [Bilodeau, Stringer and Tang \(2024a\)](#) is the following.

Lemma 5. *For all $1 \leq k \leq \lfloor t/2 \rfloor$ and $\theta \in \mathcal{B}_{\theta_*}(\zeta_N)$, if $\mathfrak{R}(\mathcal{Q}, \omega)$ is a quadrature rule satisfying $\mathcal{P}(k, d)$ then there exists a constant $C_{k,d} > 0$ depending only on d and k such that for all $N \in \mathbb{N}$ it holds P_N^* -a.s. that*

$$\begin{aligned} &\left| \int_{\mathcal{B}_{\widehat{\mathbf{u}}_i^\theta}(\gamma_i)} \exp \left\{ \ell_i^\theta(\mathbf{u}) - \ell_i^\theta(\widehat{\mathbf{u}}_i^\theta) \right\} d\mathbf{u} - |\mathbf{L}_i^\theta| \sum_{\mathbf{z} \in \mathcal{Q}(d,k)} \omega_k(\mathbf{z}) \exp \left\{ \ell_i^\theta(\mathbf{L}_i^\theta \mathbf{z} + \widehat{\mathbf{u}}_i^\theta) - \ell_i^\theta(\widehat{\mathbf{u}}_i^\theta) \right\} \right| \\ &\leq C_{k,d} \left(\left(\frac{\bar{\eta}_i^\theta}{\underline{\eta}_i^\theta} \right)^{d/2} + 1 \right) n_i^{-d/2} \\ &\quad \times \left[\max_{j \in [\kappa]} \max_{\mathbf{t} \in \tau_{\geq 2k}^{(j)}} (\widehat{M}_i^\theta)^j (\underline{\eta}_i^\theta n_i)^{-\tau(\mathbf{t})/2} + \max_{\mathbf{t} \in \tau_{\geq 3(\kappa+1)+1\{2k=2 \pmod{3}\}}^{(\kappa+1)}} (\widehat{M}_i^\theta)^{\kappa+1} (\underline{\eta}_i^\theta n_i)^{-\tau(\mathbf{t})/2} \right. \\ &\quad \left. + (\underline{\eta}_i^\theta)^{-1/2} \max_{j \in \{0\} \cup [\kappa]} (\widehat{M}_i^\theta)^j n_i^{-\frac{\gamma^2 \underline{\eta}_i^\theta + 2}{4}} + \widehat{\Lambda}_n^\theta (\widehat{M}_i^\theta)^{\kappa+2} \max_{\mathbf{t} \in \tau_{\geq 3(\kappa+2)}^{(\kappa+2)}} (\underline{\eta}_i^\theta n_i)^{-\tau(\mathbf{t})/2} \right], \end{aligned} \quad (\text{B.3})$$

where

$$\widehat{\Lambda}_n^\theta = \left[\left(\frac{\bar{\eta}_i^\theta}{\underline{\eta}_i^\theta} \right)^{d/2} \exp \left\{ \widehat{M}_i^\theta \left(\frac{\log(n_i)}{n_i} \right)^{3/2} \right\} + \exp \left\{ \widehat{M}_i^\theta \max \left\{ (\underline{\eta}_i^\theta n_i)^{-3/2}, (\underline{\eta}_i^\theta n_i)^{-\kappa} \right\} \right\} \right]$$

and κ is the smallest integer such that $3(\kappa+1) \geq 2k$.

Proof of Lemma 5. For each $\theta \in \mathcal{B}_{\theta_*}(\zeta_N)$, this follows from Lemma 4 of [Bilodeau, Stringer and Tang \(2024a\)](#). All that remains to be checked is that the constant $C_{k,d}$ does not depend on θ which follows immediately from inspection of Appendix S.3 of [Bilodeau, Stringer and Tang \(2024a\)](#). $\square$

We then want to make use of the following, which provides the necessary convergence for each of the quantities used in Lemma 5.

Lemma 6. Under Assumptions 2, 3, 6 and 7 and for $m = n_{\min}^q$ for any $q > 0$ the following hold for every $\varepsilon > 0$:

- i) $\lim_{n_{\min} \rightarrow \infty} P_N^* \left(\forall_{i=1, \dots, m} \underline{\eta} n_i^{-\varepsilon} \leq \inf_{\theta \in \mathcal{B}_{\theta_*}(\zeta_N)} \eta_i^\theta \leq \sup_{\theta \in \mathcal{B}_{\theta_*}(\zeta_N)} \bar{\eta}_i^\theta \leq \bar{\eta} n_i^\varepsilon \right) = 1.$
- ii) $\lim_{n_{\min} \rightarrow \infty} P_N^* \left(\forall_{i=1, \dots, m} \sup_{\theta \in \mathcal{B}_{\theta_*}(\zeta_N)} \widehat{M}_i^\theta \leq n_i^{1+\varepsilon} \cdot D \right) = 1.$
- iii) $\lim_{n_{\min} \rightarrow \infty} P_N^* \left(\forall_{i=1, \dots, m} \sup_{\theta \in \mathcal{B}_{\theta_*}(\zeta_N)} g_i(\widehat{\mathbf{u}}_i^\theta; \sigma) \leq c_2 \right) = 1.$

Proof of Lemma 6. i) By Assumption 6 for every $\varepsilon > 0$:

$$\lim_{n_{\min} \rightarrow \infty} P_N^* \left[\forall_{i=1, \dots, m} \sup_{\theta \in \mathcal{B}_{\theta_*}(\zeta_N)} \zeta_N^{-1} \left\| \widehat{\mathbf{u}}_i^\theta - \mathbf{u}_* \right\|_2 \leq \beta n_i^\varepsilon \right] = 1.$$

Choose $\varepsilon < \alpha$ so $\lim_{n_{\min} \rightarrow \infty} P_N^* \left(\forall_{i=1, \dots, m} \forall \theta \in \mathcal{B}_{\theta_*}(\zeta_N) \quad \widehat{\mathbf{u}}_i^\theta \in \mathcal{B}_{\mathbf{u}_*, i}(\delta) \right) = 1$ and hence

$$\begin{aligned} & \lim_{n_{\min} \rightarrow \infty} P_N^* \left(\forall_{i=1, \dots, m} \sup_{\theta \in \mathcal{B}_{\theta_*}(\zeta_N)} \bar{\eta}_i^\theta \leq \bar{\eta} n_i^\varepsilon \right) \\ &= \lim_{n_{\min} \rightarrow \infty} P_N^* \left(\forall_{i=1, \dots, m} \sup_{\theta \in \mathcal{B}_{\theta_*}(\zeta_N)} \frac{\lambda_1(-\partial_{\mathbf{u}}^2 \ell_i^\theta(\widehat{\mathbf{u}}_i^\theta))}{n} \leq \bar{\eta} n_i^\varepsilon \right) \\ &\geq \lim_{n_{\min} \rightarrow \infty} P_N^* \left(\forall_{i=1, \dots, m} \forall \theta \in \mathcal{B}_{\theta_*}(\zeta_N) \quad \widehat{\mathbf{u}}_i^\theta \in \mathcal{B}_{\mathbf{u}_*, i}(\delta), \sup_{\theta \in \mathcal{B}_{\theta_*}(\zeta_N)} \sup_{\mathbf{u} \in \mathcal{B}_{\mathbf{u}_*, i}(\delta)} \frac{\lambda_1(-\partial_{\mathbf{u}}^2 \ell_i^\theta(\mathbf{u}))}{n_i} \leq \bar{\eta} n_i^\varepsilon \right) \\ &= 1, \end{aligned}$$

where the last step uses Assumption 3. The inequality for $\underline{\eta}_n$ is analogous.

ii) First, recall that $|\mathbf{L}_i^\theta| \leq (\underline{\eta}_i^\theta n_i)^{-d/2+\varepsilon}$ for all $\theta \in \mathcal{B}_{\theta_*}(\zeta_N)$ and $\varepsilon > 0$. By i), we have

$$\lim_{n_{\min} \rightarrow \infty} P_N^* \left(\forall_{i=1, \dots, m} \sup_{\theta \in \mathcal{B}_{\theta_*}(\zeta_N)} (\underline{\eta}_i^\theta n_i)^{-d/2+\varepsilon} \leq (\underline{\eta} n_i)^{-d/2+\varepsilon} \right) = 1.$$

That is, for all $\varepsilon' > 0$,

$$\lim_{n_{\min} \rightarrow \infty} P_N^* \left(\forall_{i=1, \dots, m} \sup_{\theta \in \mathcal{B}_{\theta_*}(\zeta_N)} \bar{\mathbf{z}} |\mathbf{L}_i^\theta| \vee \gamma_i < \varepsilon' \right) = 1,$$

so by Assumption 6 again,

$$\lim_{n_{\min} \rightarrow \infty} P_N^* \left(\forall_{i=1, \dots, m} \forall \theta \in \mathcal{B}_{\theta_*}(\zeta_N) \quad \mathcal{B}_{\widehat{\mathbf{u}}_i^\theta}(\bar{\mathbf{z}} |\mathbf{L}_i^\theta| \vee \gamma_i) \subseteq \mathcal{B}_{\mathbf{u}_*, i}(\delta) \right) = 1.$$

Thus,

$$\begin{aligned}
& \lim_{n_{\min} \rightarrow \infty} P_N^* \left(\forall_{i=1, \dots, m} \sup_{\theta \in \mathcal{B}_{\theta_*}(\zeta_N)} \widehat{M}_i^\theta \leq n_i^{1+\varepsilon} \cdot D \right) \\
&= \lim_{n_{\min} \rightarrow \infty} P_N^* \left(\forall_{i=1, \dots, m} \sup_{\theta \in \mathcal{B}_{\theta_*}(\zeta_N)} \sup_{\alpha: |\alpha| \leq t} \sup_{\mathbf{u} \in \mathcal{B}_{\widehat{\mathbf{u}}_i^\theta}(\bar{\mathbf{z}} | \mathbf{L}_i^\theta | \vee \gamma_i)} \left| \partial^\alpha \ell_i^\theta(\mathbf{u}) \right| \leq n_i^{1+\varepsilon} \cdot D \right) \\
&\geq \lim_{n_{\min} \rightarrow \infty} P_N^* \left(\forall_{i=1, \dots, m} \forall \theta \in \mathcal{B}_{\theta_*}(\zeta_N) \quad \mathcal{B}_{\widehat{\mathbf{u}}_i^\theta}(\bar{\mathbf{z}} | \mathbf{L}_i^\theta | \vee \gamma_i) \subseteq \mathcal{B}_{\mathbf{u}_{*,i}}(\delta), \right. \\
&\quad \left. \sup_{\theta \in \mathcal{B}_{\theta_*}(\zeta_N)} \sup_{\alpha: |\alpha| \leq t} \sup_{\mathbf{u} \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} \left| \partial^\alpha \ell_i^\theta(\mathbf{u}) \right| \leq n_i^{1+\varepsilon} \cdot D \right) = 1,
\end{aligned}$$

for all $\varepsilon > 0$, where the last step uses Assumption 2.

iii) This follows directly from Assumptions 6 and 7. $\square$

We now control the contribution of the integral of the normalized likelihood outside of a fixed ball.

Lemma 7. *Under Assumptions 4 and 7, for every $\varepsilon > 0$ and for $m = n_{\min}^q$ for any $q > 0$*

$$\lim_{n_{\min} \rightarrow \infty} P_N^* \left[\forall_{i=1, \dots, m} \sup_{\theta \in \mathcal{B}_{\theta_*}(\zeta_N)} \int_{[\mathcal{B}_{\mathbf{u}_{*,i}}(\delta)]^c} \exp\{\ell_i^\theta(\mathbf{u}) - \ell_i^\theta(\widehat{\mathbf{u}}_i^\theta)\} d\mathbf{u} \leq \frac{e^{-n_i^{1-\varepsilon} b}}{c_1} \right] = 1,$$

Proof of Lemma 7. Fix $\theta \in \mathcal{B}_{\theta_*}(\zeta_N)$. By the proof of Lemma 5 in Bilodeau, Stringer and Tang (2024a),

$$\begin{aligned}
& \int_{[\mathcal{B}_{\mathbf{u}_{*,i}}(\delta)]^c} \exp\{\ell_i^\theta(\mathbf{u}) - \ell_i^\theta(\widehat{\mathbf{u}}_i^\theta)\} d\mathbf{u} \\
&\leq \frac{1}{g(\mathbf{u}_{*,i}; \boldsymbol{\sigma})} \sup_{\mathbf{u} \in [\mathcal{B}_{\mathbf{u}_{*,i}}(\delta)]^c} \exp\{\log \pi_i(\mathbf{y}_i | \mathbf{u}; \boldsymbol{\theta}) - \log \pi_i(\mathbf{y}_i | \mathbf{u}_{*,i}; \boldsymbol{\theta})\}.
\end{aligned}$$

The result then follows by applying Assumptions 4 and 7. $\square$

Lemma 8. *Under Assumptions 3, 4, 6 and 7, there exists a constant $D > 0$ such that for every $\varepsilon > 0$ and for $m = n_{\min}^q$ for any $q > 0$:*

$$\lim_{n_{\min} \rightarrow \infty} P_N^* \left[\forall_{i=1, \dots, m} \sup_{\theta \in \mathcal{B}_{\theta_*}(\zeta_N)} \int_{[\mathcal{B}_{\widehat{\mathbf{u}}_i^\theta}(\gamma_i)]^c \cap \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} \exp\{\ell_i^\theta(\mathbf{u}) - \ell_i^\theta(\widehat{\mathbf{u}}_i^\theta)\} d\mathbf{u} \leq D \frac{1}{n_i^{\gamma^2 \eta / 4 + d/2 - \varepsilon}} \right] = 1.$$

Proof of Lemma 8. Fix $\theta \in \mathcal{B}_{\theta_*}(\zeta_N)$ and for any value of $\varepsilon > 0$ let

$$\underline{\eta}_i^{*,\theta}(\varepsilon) = \inf_{\mathbf{u} \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} \lambda_d(-\partial_{\mathbf{u}}^2 \ell_i^\theta(\mathbf{u})) / n_i^{1-\varepsilon}.$$

By the proof of Lemma 6 in [Bilodeau, Stringer and Tang \(2024a\)](#),

$$\int_{[\mathcal{B}_{\mathbf{u}_i^\theta}(\gamma_i)]^c \cap \mathcal{B}_{\mathbf{u}_*}(\delta)} \exp\{\ell_i^\theta(\mathbf{u}) - \ell_i^\theta(\widehat{\mathbf{u}}_i^\theta)\} d\mathbf{u} \leq \left(\frac{2\pi e}{\underline{\eta}_i^{*,\theta}}\right)^{d/2} n_i^{-\gamma^2 \underline{\eta}_i^{*,\theta}/4 - d/2 + \varepsilon}.$$

By Assumption 6, $\lim_{n_i \rightarrow \infty} P_N^* \left(\forall i=1, \dots, m \ \forall \theta \in \mathcal{B}_{\theta_*}(\zeta_N) \ \widehat{\mathbf{u}}_i^\theta \in \mathcal{B}_{\mathbf{u}_*,i}(\delta) \right) = 1$, and by Assumption 3,

$$\lim_{n_{\min} \rightarrow \infty} P_N^* \left[\forall i=1, \dots, m \ \inf_{\theta \in \mathcal{B}_{\theta_*}(\zeta_N)} \underline{\eta}_i^{*,\theta}(\varepsilon) \geq \underline{\eta} \right] = 1,$$

giving the statement of the lemma. □

B.2. Proof of Lemma 4

Fix $\theta \in \mathcal{B}_{\theta_*}(\zeta_N)$. Adapting Section S.2.3 of [Bilodeau, Stringer and Tang \(2024a\)](#) it holds that if $\widehat{\mathbf{u}}_i^\theta \in \mathcal{B}_{\mathbf{u}_*}(\delta/2)$, $\underline{\eta}_i^\theta \geq \underline{\eta}$, $\overline{\eta}_i^* \leq \overline{\eta}$, and $\overline{\eta}_i^\theta \leq \overline{\eta}$,

$$\begin{aligned} & \frac{\widehat{\pi}_i^{\text{AQ}}(\mathbf{y}_i; \theta)}{\pi_i(\mathbf{y}_i \mid \widehat{\mathbf{u}}_i^\theta; \theta)} \\ & \geq c_1 (\overline{\eta} n_i)^{-d/2 - \varepsilon} \left[(2\pi)^{d/2} - \frac{1}{6} d^3 (\underline{\eta} n_i)^{-3/2 + \varepsilon} \overline{\mathbf{z}} \max_{\mathbf{z} \in Q, (i_1, i_2, i_3) \in [d]} \left| \partial^{i_1 i_2 i_3} \ell_i^\theta(\mathbf{V}_i^{\mathbf{L}_i^\theta \mathbf{z} + \widehat{\mathbf{u}}_i^\theta}) \right| \cdot \sum_{\mathbf{z} \in Q} |\omega(\mathbf{z})| \right], \end{aligned}$$

where $\mathbf{V}_i^{\mathbf{L}_i^\theta \mathbf{z} + \widehat{\mathbf{u}}_i^\theta} = \tau_{\mathbf{z}}(\mathbf{L}_i^\theta \mathbf{z} + \widehat{\mathbf{u}}_i^\theta) + (1 - \tau_{\mathbf{z}})\widehat{\mathbf{u}}_i^\theta$ for some $\tau_{\mathbf{z}} \in [0, 1]$ and for all $\varepsilon > 0$. Next, if $\overline{\eta}_i^\theta \leq \overline{\eta}$,

$$|\mathbf{L}_i^\theta| = \sqrt{|\mathbf{H}_i^\theta|^{-1}} = |\mathbf{H}_i^\theta|^{-1/2} \geq \left[\lambda_1(\mathbf{H}_i^\theta) \right]^{-d/2 - \varepsilon} = (\overline{\eta}_i^\theta n_i)^{-d/2 - \varepsilon} \geq (\overline{\eta} n_i)^{-d/2 - \varepsilon}.$$

Thus, by Lemma 6 part i) and Assumption 2,

$$\lim_{n_{\min} \rightarrow \infty} P_N^* \left[\forall i=1, \dots, m \ \sup_{\theta \in \mathcal{B}_{\theta_*}(\zeta_N)} \max_{\mathbf{z} \in Q, (i_1, i_2, i_3) \in [d]} \left| \partial^{i_1 i_2 i_3} \ell_i^\theta(\mathbf{V}_i^{\mathbf{L}_i^\theta \mathbf{z} + \widehat{\mathbf{u}}_i^\theta}) \right| \leq n_i^{1+\varepsilon} \cdot D \right] = 1.$$

That is, for all $\varepsilon' > 0$,

$$\begin{aligned} & \lim_{n_{\min} \rightarrow \infty} P_N^* \left[\forall i=1, \dots, m \ \sup_{\theta \in \mathcal{B}_{\theta_*}(\zeta_N)} \frac{1}{6} d^3 (\underline{\eta} n_i)^{-3/2 + \varepsilon} \overline{\mathbf{z}} \cdot \right. \\ & \quad \left. \max_{\mathbf{z} \in Q, (i_1, i_2, i_3) \in [d]} \left| \partial^{i_1 i_2 i_3} \ell_i^\theta(\mathbf{V}_i^{\mathbf{L}_i^\theta \mathbf{z} + \widehat{\mathbf{u}}_i^\theta}) \right| \cdot \sum_{\mathbf{z} \in Q} |\omega(\mathbf{z})| \leq \varepsilon' \right] = 1. \end{aligned}$$

The statement of the lemma then follows from Assumptions 3, 6 and 7, as well as Lemma 6. □

Acknowledgments

We are grateful for helpful comments from Helen Ogden, Glen McGee, Jeffrey Negrea, Art Owen, and the Editor and two reviewers.

Funding

Blair Bilodeau was supported by an NSERC Canada Graduate Scholarship and the Vector Institute. Alex Stringer is supported by NSERC Discovery Grant RGPIN-2023-03331.

Supplementary Material

Supplementary materials for: “Asymptotics of numerical integration for two-level mixed models”

The supplementary materials [Bilodeau, Stringer and Tang \(2024b\)](#) contains all proofs to the results not included in the paper as well as some auxiliary results.

References

- BATES, D., MÄCHLER, M., BOLKER, B. and WALKER, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67** 1–48.
- BHASKARAN, A. and WAND, M. P. (2023). Dispersion parameter extension of precise generalized linear mixed model asymptotics. *Statist. Probab. Lett.* **193** 109691.
- BIANCONCINI, S. (2014). Asymptotic properties of adaptive maximum likelihood estimators in latent variable models. *Bernoulli* **20** 1507–1531.
- BILODEAU, B., STRINGER, A. and TANG, Y. (2024a). Stochastic convergence rates and applications of adaptive quadrature in Bayesian inference. *J. Amer. Statist. Assoc.* **119** 690–700.
- BILODEAU, B., STRINGER, A. and TANG, Y. (2024b). Supplement to “Asymptotics of numerical integration for two-level mixed models”.
- BRESLOW, N. E. and CLAYTON, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* **88** 9–25.
- DAVIS, P. J. and RABINOWITZ, P. (1984). *Methods of Numerical Integration*. Academic Press.
- DIGGLE, P., MORAGA, P., ROWLINGSON, B. and TAYLOR, B. (2013). Spatial and spatio-temporal log-Gaussian Cox processes: extending the geostatistical paradigm. *Statist. Sci.* **28** 542–563.
- JIANG, J., WAND, M. P. and BHASKARAN, A. (2021). Usable and precise asymptotics for generalized linear mixed model analysis and design. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 55–82.
- JIN, S. and ANDERSSON, B. (2020). A note on the accuracy of adaptive Gauss-Hermite quadrature. *Biometrika* **107** 737–744.
- KAUERMANN, J. G., KRIVOBOKOVA, T. and FAHRMEIR, L. (2009). Some asymptotic results on generalized penalized spline smoothing. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 487–503.
- KLEIJN, B. J. K. and VAN DER VAART, A. W. (2012). The Bernstein von-Mises theorem under misspecification. *Electron. J. Stat.* **6** 354–381.
- LESAFFRE, E. and SPIESSENS, B. (2001). On the effect of the number of quadrature points in a logistic random-effects model: an example. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **50** 325–335.
- LIU, Q. and PIERCE, D. A. (1994). A note on Gauss-Hermite quadrature. *Biometrika* **81** 624–629.
- MAESTRINI, L., BHASKARAN, A. and WAND, M. P. (2024). Second term improvement to generalized linear mixed model asymptotics. *Biometrika* **111** 1077–1084.
- NAYLOR, J. C. and SMITH, A. F. M. (1982). Applications of a method for the efficient computation of posterior distributions. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **31** 214–225.

-
- NIE, L. (2007). Convergence rate of MLE in generalized linear and nonlinear mixed-effects models: theory and applications. *J. Statist. Plann. Inference* **137** 1787–1804.
- OGDEN, H. (2017). On asymptotic validity of naive inference with an approximate likelihood. *Biometrika* **104** 153–164.
- OWEN, A. B. (2019). *Monte Carlo Book: The Quasi-Monte Carlo Parts*. Unpublished manuscript.
- PINHEIRO, J. C. and BATES, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed effects models. *J. Comput. Graph. Statist.* **4** 12–35.
- RIGOLLET, P. and HÜTTER, J.-C. (2023). High-dimensional statistics. *arXiv preprint arXiv:2310.19244*.
- RIZOPOULOS, D. (2020). GLMMadaptive: generalized linear mixed models using adaptive Gaussian quadrature R package version 0.7-15.
- SANZ-ALONSO, D. and YANG, R. (2022). Finite element representations of Gaussian processes: Balancing numerical and statistical accuracy. *SIAM/ASA J. Uncertain. Quantif* **10** 1323–1349.
- SAUMARD, A. and WELLNER, J. A. (2014). Log-concavity and strong log-concavity: a review. *Stat. Surv.* **8** 45–114.
- TANG, Y. and REID, N. (2024). Laplace and saddlepoint approximations in high dimensions. *Bernoulli (in print)*.
- VAN DER VAART, A. (1998). *Asymptotic Statistics*. Cambridge University Press.
- VONESH, E. F. (1996). A note on the use of Laplace’s approximation for nonlinear mixed effect models. *Biometrika* **83** 447–452.
- WOOD, S. (2016). Smoothing parameter and model selection for general smooth models. *J. Amer. Statist. Assoc.* **111** 1548–1575.
-

Supplementary materials for: “Asymptotics of numerical integration for two-level mixed models”

BLAIR BILODEAU^{1,a}, ALEX STRINGER^{2,b}, YANBO TANG^{*3,c},

¹University of Toronto, ^ablair.bilodeau@mail.utoronto.ca

²University of Waterloo, ^balex.stringer@uwaterloo.ca

³Imperial College London, ^cyanbo.tang@imperial.ac.uk

Appendix C: Supporting results for the proof of Theorem 2

The proof of Theorem 2 requires the following additional lemma:

Lemma 9. For $0 < \underline{\eta}'$, and for all $\varepsilon > 0$ and for $m = n_{\min}^q$ for any $q > 0$

$$\lim_{n_{\min} \rightarrow \infty} P_N^* \left[\underline{\eta}' \sum_{i=1}^m n_i^{1-\varepsilon} \leq \inf_{\theta \in \mathcal{B}_{\theta_*}(\zeta_N)} \lambda_p(-\partial_{\theta}^2 \ell(\mathbf{y}; \theta)) \right] = 1.$$

Proof. Recall that $\ell(\mathbf{y}; \theta) = \log \pi(\mathbf{y}; \theta) = \log \int \pi(\mathbf{y}; \mathbf{u}, \theta) d\mathbf{u} = \sum_{i=1}^m \log \int \pi(\mathbf{y}_i; \mathbf{u}_i, \theta) d\mathbf{u}_i$. Using Leibniz rule for exchanging integration with differentiation we have the following for all $\theta \in \mathcal{B}_{\theta_*}(\zeta_N)$:

$$\begin{aligned} \lambda_p(-\partial_{\theta}^2 \ell(\mathbf{y}; \theta)) &= \lambda_p \left(-\partial_{\theta}^2 \sum_{i=1}^m \log \int_{\mathbf{u}_i \in \mathbb{R}^d} \pi(\mathbf{y}_i; \mathbf{u}_i, \theta) d\mathbf{u}_i \right), \\ &\geq \sum_{i=1}^m \lambda_p \left(-\partial_{\theta}^2 \log \int_{\mathbf{u}_i \in \mathbb{R}^d} \pi(\mathbf{y}_i; \mathbf{u}_i, \theta) d\mathbf{u}_i \right), \end{aligned}$$

by Weyl’s inequality. The remainder of the proof is dedicated to showing that

$$\lambda_p \left(-\partial_{\theta}^2 \log \int_{\mathbf{u}_i \in \mathbb{R}^d} \pi(\mathbf{y}_i; \mathbf{u}_i, \theta) d\mathbf{u}_i \right) \geq C n_i^{1-\varepsilon},$$

for a constant $C > 0$ independent of i , from which the statement of the Lemma follows.

Note that we may write:

$$\begin{aligned}
& \log \int_{\mathbf{u}_i \in \mathbb{R}^d} \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta}) d\mathbf{u}_i \\
&= \underbrace{\log \left(\frac{\int_{\mathbf{u}_i \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta}) d\mathbf{u}_i}{\int_{\mathcal{B}_{\boldsymbol{\theta}_*}(\delta)} \int_{\mathbf{u}_i \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta}) d\mathbf{u}_i d\boldsymbol{\theta}} \right)}_A + \underbrace{\log \left(\frac{\int_{\mathbf{u}_i \in \mathbb{R}^d} \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta}) d\mathbf{u}_i}{\int_{\mathbf{u}_i \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta}) d\mathbf{u}_i} \right)}_B \\
&+ \underbrace{\log \left(\int_{\mathcal{B}_{\boldsymbol{\theta}_*}(\delta)} \int_{\mathbf{u}_i \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta}) d\mathbf{u}_i d\boldsymbol{\theta} \right)}_C.
\end{aligned} \tag{C.1}$$

The C term in Eq. (C.1) is 0 when differentiated with respect to $\boldsymbol{\theta}$, so it can be ignored. The A term in Eq. (C.1) is the marginal of a truncated posterior distribution of $\boldsymbol{\theta}$ with uniform prior on $\boldsymbol{\theta}$ on $\mathcal{B}_{\boldsymbol{\theta}_*}(\delta)$, truncated to $\mathcal{B}_{\boldsymbol{\theta}_*}(\delta) \times \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)$. By Assumption 3, the smallest eigenvalue of the negative hessian of the log-density of this truncated posterior of $(\boldsymbol{\theta}, \mathbf{u})$ is uniformly lower bounded by $\underline{\eta} n_i^{1-\varepsilon}$, implying it is a strongly log-concave distribution. As strong log-concavity is preserved by marginalization, it follows that the negative hessian of the first term is also lower bounded by $\underline{\eta} n_i^{1-\varepsilon}$. See [Saumard and Wellner \(2014\)](#) Proposition 2.24 c) and Theorems 3.8 for the statement and proof of these results.

We now show that the largest singular value of the matrix B in Eq. (C.1) is asymptotically negligible. When differentiated with respect to $\boldsymbol{\theta}$, B equals:

$$\begin{aligned}
& -\partial_{\boldsymbol{\theta}}^2 \log \left(\frac{\int_{\mathbf{u}_i \in \mathbb{R}^d} \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta}) d\mathbf{u}_i}{\int_{\mathbf{u}_i \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta}) d\mathbf{u}_i} \right) \\
&= -\partial_{\boldsymbol{\theta}}^2 \log \left(\int_{\mathbf{u}_i \in \mathbb{R}^d} \exp\{\log \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta})\} d\mathbf{u}_i \right) + \partial_{\boldsymbol{\theta}}^2 \log \left(\int_{\mathbf{u}_i \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} \exp\{\log \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta})\} d\mathbf{u}_i \right) \\
&= \underbrace{\frac{\int_{\mathbf{u}_i \in \mathbb{R}^d} -\partial_{\boldsymbol{\theta}}^2 \log \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta}) \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta}) d\mathbf{u}_i}{\int_{\mathbf{u}_i \in \mathbb{R}^d} \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta}) d\mathbf{u}_i} - \frac{\int_{\mathbf{u}_i \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} -\partial_{\boldsymbol{\theta}}^2 \log \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta}) \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta}) d\mathbf{u}_i}{\int_{\mathbf{u}_i \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta}) d\mathbf{u}_i}}_D \\
&+ \left(\frac{\int_{\mathbf{u}_i \in \mathbb{R}^d} \partial_{\boldsymbol{\theta}} \log \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta}) \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta}) d\mathbf{u}_i}{\int_{\mathbf{u}_i \in \mathbb{R}^d} \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta}) d\mathbf{u}_i} \right) \left(\frac{\int_{\mathbf{u}_i \in \mathbb{R}^d} \partial_{\boldsymbol{\theta}}^{\top} \log \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta}) \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta}) d\mathbf{u}_i}{\int_{\mathbf{u}_i \in \mathbb{R}^d} \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta}) d\mathbf{u}_i} \right) \\
&- \left(\frac{\int_{\mathbf{u}_i \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} \partial_{\boldsymbol{\theta}} \log \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta}) \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta}) d\mathbf{u}_i}{\int_{\mathbf{u}_i \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta}) d\mathbf{u}_i} \right) \left(\frac{\int_{\mathbf{u}_i \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} \partial_{\boldsymbol{\theta}}^{\top} \log \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta}) \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta}) d\mathbf{u}_i}{\int_{\mathbf{u}_i \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta}) d\mathbf{u}_i} \right),
\end{aligned} \tag{C.2}$$

these terms can also be interpreted as the conditional expectation under the posterior measure, and of the truncated posterior measures of $\theta|\mathbf{u}$. Consider D ,

$$\begin{aligned}
& \frac{\int_{\mathbf{u}_i \in \mathbb{R}^d} -\partial_{\theta}^2 \log \pi(\mathbf{y}_i; \mathbf{u}_i, \theta) \pi(\mathbf{y}_i; \mathbf{u}_i, \theta) d\mathbf{u}_i}{\int_{\mathbf{u}_i \in \mathbb{R}^d} \pi(\mathbf{y}_i; \mathbf{u}_i, \theta) d\mathbf{u}_i} - \frac{\int_{\mathbf{u}_i \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} -\partial_{\theta}^2 \log \pi(\mathbf{y}_i; \mathbf{u}_i, \theta) \pi(\mathbf{y}_i; \mathbf{u}_i, \theta) d\mathbf{u}_i}{\int_{\mathbf{u}_i \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} \pi(\mathbf{y}_i; \mathbf{u}_i, \theta) d\mathbf{u}_i} \\
&= \underbrace{\frac{\int_{\mathbf{u}_i \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)^c} -\partial_{\theta}^2 \log \pi(\mathbf{y}_i; \mathbf{u}_i, \theta) \pi(\mathbf{y}_i; \mathbf{u}_i, \theta) d\mathbf{u}_i}{\int_{\mathbf{u}_i \in \mathbb{R}^d} \pi(\mathbf{y}_i; \mathbf{u}_i, \theta) d\mathbf{u}_i}}_{D.1} \\
&+ \underbrace{\left(\frac{\int_{\mathbf{u}_i \in \mathbb{R}^p} \pi(\mathbf{y}_i; \mathbf{u}_i, \theta) d\mathbf{u}_i}{\int_{\mathbf{u}_i \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} \pi(\mathbf{y}_i; \mathbf{u}_i, \theta) d\mathbf{u}_i} - 1 \right) \frac{\int_{\mathbf{u}_i \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} -\partial_{\theta}^2 \log \pi(\mathbf{y}_i; \mathbf{u}_i, \theta) \pi(\mathbf{y}_i; \mathbf{u}_i, \theta) d\mathbf{u}_i}{\int_{\mathbf{u}_i \in \mathbb{R}^p} \pi(\mathbf{y}_i; \mathbf{u}_i, \theta) d\mathbf{u}_i}}_{D.2}. \tag{C.3}
\end{aligned}$$

Take $\delta' > 0$ to be a constant to be specify later, multiply and divide $D.1$ by $\exp\{-\log \pi(\mathbf{y}_i; \mathbf{u}_{*}, \theta)\}$ and take it's operator norm, resulting in

$$\begin{aligned}
& \frac{\left\| \int_{\mathbf{u}_i \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)^c} -\partial_{\theta}^2 \log \pi(\mathbf{y}_i; \mathbf{u}_i, \theta) \exp\{\log \pi(\mathbf{y}_i; \mathbf{u}_i, \theta) - \log \pi(\mathbf{y}_i; \mathbf{u}_{*}, \theta)\} g(\mathbf{u}_i) d\mathbf{u}_i \right\|_{op}}{\int_{\mathbf{u}_i \in \mathbb{R}^d} \exp\{\log \pi(\mathbf{y}_i; \mathbf{u}_i, \theta) - \log \pi(\mathbf{y}_i; \mathbf{u}_{*}, \theta)\} g(\mathbf{u}_i) d\mathbf{u}_i} \\
&\leq \frac{\exp(-n_i^{1-\varepsilon} b) \int_{\mathbf{u}_i \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)^c} \left\| -\partial_{\theta}^2 \log \pi(\mathbf{y}_i; \mathbf{u}_i, \theta) g(\mathbf{u}_i) \right\|_{op} d\mathbf{u}_i}{\int_{\mathbf{u}_i \in \mathcal{B}_{\mathbf{u}_{*}}(\delta'_{n_i})} \exp\{\log \pi(\mathbf{y}_i; \mathbf{u}_i, \theta) - \log \pi(\mathbf{y}_i; \mathbf{u}_{*}, \theta)\} g(\mathbf{u}_i) d\mathbf{u}_i} \\
&\leq \frac{\exp(-n_i^{1-\varepsilon} b) \int_{\mathbf{u}_i \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)^c} \left\| -\partial_{\theta}^2 \log \pi(\mathbf{y}_i; \mathbf{u}_i, \theta) g(\mathbf{u}_i) \right\|_{op} d\mathbf{u}_i}{\int_{\mathbf{u}_i \in \mathcal{B}_{\mathbf{u}_{*}}(\delta'_{n_i})} \exp\{\log \pi(\mathbf{y}_i; \mathbf{u}_i, \theta) - \log \pi(\mathbf{y}_i; \mathbf{u}_{*}, \theta)\} g(\mathbf{u}_i) d\mathbf{u}_i}, \tag{C.4}
\end{aligned}$$

for every $\varepsilon > 0$ and a constant $b > 0$ independent of i by Assumption 4 and $\delta_{n_i} = n_i^{\alpha'}$ for some value of $0 < \alpha'$ to be specified later. We now upper bound the numerator and denominator of this fraction separately. The numerator of $D.1$ is upper bounded by

$$\begin{aligned}
& \int_{\mathbf{u}_i \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)^c} \left\| -\partial_{\theta}^2 \log \pi(\mathbf{y}_i; \mathbf{u}_i, \theta) \right\|_{op} g(\mathbf{u}_i) d\mathbf{u}_i \leq \int_{\mathbf{u}_i \in \mathbb{R}^d} \left\| -\partial_{\theta}^2 \log \pi(\mathbf{y}_i; \mathbf{u}_i, \theta) \right\|_{op} g(\mathbf{u}_i) d\mathbf{u}_i \\
&= O(n_i^k)
\end{aligned}$$

by Assumption 7, where the $O(\cdot)$ term is uniform in i . We now bound the denominator of $D.1$ by writing it as:

$$\exp \left\{ \underbrace{\log \pi(\mathbf{y}_i; \mathbf{u}_i, \theta) - \log \pi(\mathbf{y}_i; \widehat{\mathbf{u}}_i^{\theta}, \theta)}_E + \underbrace{\log \pi(\mathbf{y}_i; \widehat{\mathbf{u}}_i^{\theta}, \theta) - \log \pi(\mathbf{y}_i; \mathbf{u}_{*}, \theta)}_F \right\}.$$

The F term:

$$\log \pi(\mathbf{y}_i; \widehat{\mathbf{u}}_i^\theta, \boldsymbol{\theta}) - \log \pi(\mathbf{y}_i; \mathbf{u}_*, \boldsymbol{\theta}) > 0.$$

While, by a second order Taylor expansion, the E term:

$$\begin{aligned} \log \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta}) - \log \pi(\mathbf{y}_i; \widehat{\mathbf{u}}_i^\theta, \boldsymbol{\theta}) &= (\mathbf{u}_i - \widehat{\mathbf{u}}_i^\theta)^\top \{ \partial_\theta^2 \log \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta}^*) \} (\mathbf{u}_i - \widehat{\mathbf{u}}_i^\theta) \\ &\geq -(\delta'_{n_i})^2 n_i^{1+\varepsilon} \bar{\eta} = -n_i^{1+\varepsilon-\alpha'} \bar{\eta}, \end{aligned} \quad (\text{C.5})$$

for $\bar{\eta}$ independent of i and where

$$\left\{ \partial_\theta^2 \log \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta}^*) \right\} = \int_0^1 (1-t) \partial_\theta^2 \log \pi(\mathbf{y}_i; \widehat{\mathbf{u}}_i^\theta + t\mathbf{u}_i, \boldsymbol{\theta}) dt,$$

by Assumption 3, as for any positive definite matrix $A(t)$ indexed by a scalar random variable t , it is the case that by the variational representation of the maximal eigenvalue and the linearity of expectations that:

$$\begin{aligned} \lambda_1 \left(\int A(t) dt \right) &= \max_{\|x\|=1} x^\top \left(\int A(t) dt \right) x \\ &= \max_{\|x\|=1} \int x^\top A(t) x dt \\ &\leq \int \max_{\|x\|=1} x^\top A(t) x dt = \int \lambda_1(A(t)) dt. \end{aligned} \quad (\text{C.6})$$

Combining the bounds on E and F , we have that the denominator of $D.1$ is lower bounded by:

$$\begin{aligned} &\int_{\mathbf{u}_i \in \mathcal{B}_{\mathbf{u}_*}(\delta'_{n_i})} \exp\{\log \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta}) - \log \pi(\mathbf{y}_i; \mathbf{u}_*, \boldsymbol{\theta})\} g(\mathbf{u}_i) d\mathbf{u}_i \\ &\geq -2(\delta')^2 n_i^{1+\varepsilon-\alpha'} \bar{\eta} \int_{\mathbf{u}_i \in \mathcal{B}_{\mathbf{u}_*}(\delta'_{n_i})} g(\mathbf{u}_i) d\mathbf{u}_i \\ &= \exp\{-2(\delta')^2 n_i^{1+\varepsilon-\alpha'} \bar{\eta}\} \mathbb{P}[\mathbf{u} \in \mathcal{B}_{\mathbf{u}_*}(\delta'_{n_i})]. \end{aligned}$$

We can lower bound the probability in the above line by:

$$\begin{aligned} \mathbb{P}[\mathbf{u} \in \mathcal{B}_{\mathbf{u}_*}(\delta'_{n_i})] &= \int_{\mathcal{B}_{\mathbf{u}_*}(\delta'_{n_i})} \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}(\boldsymbol{\sigma})|^{d/2}} \exp\left(-\frac{1}{2} \mathbf{u}^\top \boldsymbol{\Sigma}(\boldsymbol{\sigma}) \mathbf{u}\right) d\mathbf{u} \\ &\geq \int_{\mathcal{B}_{\mathbf{u}_*}(\delta'_{n_i})} \frac{1}{(2\pi)^{d/2} (\|\boldsymbol{\Sigma}(\boldsymbol{\sigma})\|_{op})^{d/2}} \exp\left(-\frac{1}{2 \|\boldsymbol{\Sigma}(\boldsymbol{\sigma})\|_{op}} \left\{ \|\mathbf{u} - \mathbf{u}_{*,i}\|_2^2 + \|\mathbf{u}_{*,i}\|_2^2 \right\}\right) d\mathbf{u} \\ &\geq \frac{1}{(2\pi)^{d/2} (\|\boldsymbol{\Sigma}(\boldsymbol{\sigma})\|_{op})^{d/2}} \exp\left(-\frac{\{\delta_{n_i}^2 + \|\boldsymbol{\Sigma}(\boldsymbol{\sigma})\|_{op} \log(m)\}}{2\lambda_d(\boldsymbol{\Sigma}(\boldsymbol{\sigma}))}\right) \int_{\mathcal{B}_{\mathbf{u}_*}(\delta'_{n_i})} d\mathbf{u} \\ &\geq \frac{(\delta'_{n_i})^2}{(2\pi)^{d/2} (\|\boldsymbol{\Sigma}(\boldsymbol{\sigma})\|_{op})^{d/2}} \exp\left(-\frac{\{\delta_{n_i}^2 + \|\boldsymbol{\Sigma}(\boldsymbol{\sigma})\|_{op} \log(m)\}}{2\lambda_d(\boldsymbol{\Sigma}(\boldsymbol{\sigma}))}\right) \\ &\geq \exp(C(\delta'_{n_i})^2) = \exp(Cn_i^{-2\alpha'}) \end{aligned}$$

for some $C > 0$, as $\max_{i=1,\dots,m} \|\mathbf{u}_{*,i}\|_2^2 = O(\log(m))$ with probability tending to 1 by Theorem 1.14 in [Rigollet and Hütter \(2023\)](#). Combining the bounds on the numerator and denominator of $D.1$:

$$\begin{aligned} & \frac{\left\| \int_{\mathbf{u}_i \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)^C} -\partial_{\boldsymbol{\theta}}^2 \log \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta}) \exp\{\log \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta}) - \log \pi(\mathbf{y}_i; \mathbf{u}_*, \boldsymbol{\theta})\} g(\mathbf{u}_i) d\mathbf{u}_i \right\|_{op}}{\int_{\mathbf{u}_i \in \mathbb{R}^d} \exp\{\log \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta}) - \log \pi(\mathbf{y}_i; \mathbf{u}_*, \boldsymbol{\theta})\} g(\mathbf{u}_i) d\mathbf{u}_i} \\ & \leq O(n_i^k) \exp\{-n_i^{1-\varepsilon} b + 2n_i^{1+\varepsilon-\alpha'} \bar{\eta}\} \exp(Cn_i^{2\alpha'}) = o(1), \end{aligned}$$

by taking $\varepsilon < \alpha' < 1/2 - \varepsilon/2$, uniformly in i . Now consider $D.2$ in Eq. (C.3), note that for any fixed value of $\boldsymbol{\theta} \in \mathcal{B}_{\boldsymbol{\theta}_*}(\zeta_N)$

$$\left| \frac{\int_{\mathbf{u}_i \in \mathbb{R}^d} \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta}) d\mathbf{u}_i}{\int_{\mathbf{u}_i \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta}) d\mathbf{u}_i} - 1 \right| \leq \frac{\int_{\mathbf{u}_i \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)^C} \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta}) d\mathbf{u}_i}{\int_{\mathbf{u}_i \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta}) d\mathbf{u}_i} = O(\exp(-n_i^{1-\varepsilon})),$$

by the same arguments used to bound D.1. Secondly,

$$\frac{\int_{\mathbf{u}_i \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} \left\| -\partial_{\boldsymbol{\theta}}^2 \log \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta}) \right\|_{op} \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta}) d\mathbf{u}_i}{\int_{\mathbf{u}_i \in \mathbb{R}^p} \pi(\mathbf{y}_i; \mathbf{u}_i, \boldsymbol{\theta}) d\mathbf{u}_i} \leq n_i^k \bar{\eta},$$

thus the term $D.2 = O(\exp(-n_i^{1-\varepsilon}))$, which decays exponentially and is therefore asymptotically negligible, thereby showing that the entirety of D is negligible.

As for the other terms in Eq. (C.2) which are not shown, nearly the same proof strategy applies. The only additional details required are that the prior expectation of the first derivative is finite and the Frobenius norm upper-bounds the largest eigenvalue of a matrix. Therefore the B term in Eq. (C.1) is uniformly negligible showing the desired result. $\square$

Appendix D: Proof of Proposition 1

D.1. Regularity conditions

Proposition 1 holds for all well-specified GLMMs for which the response distribution is non-degenerate. Assumptions 8 and 9 formalize what is meant by these notions. In the Assumptions which follow, assume that $m = n_{\min}^q$ for some $q > 0$ and that n_i are increasing at the same rate.

Assumption 8 (Non-Degenerate). *The GLMM is non-degenerate if there exist $\delta > 0$ and $\boldsymbol{\theta}_* \in \Theta$ such that each of the following six statements are true with $m = n_{\min}^q$ and $n_{\min} \rightarrow \infty$.*

(i) *The natural parameter space,*

$$\Xi = \left\{ \eta \in \mathbb{R} : \int_S \exp \left\{ \frac{y\eta}{a(\boldsymbol{\phi})} + c(y; \boldsymbol{\phi}) \right\} dy < \infty \right\},$$

is an open subset of $\mathbb{R}$ containing $\eta = 0$.

(ii) *For every $C > 0$, $\inf_{|\eta| < C \log(n_{\min})^{1-\varepsilon}} |b''(\eta)| > n_{\min}^{-\varepsilon}$, for all $\varepsilon > 0$.*

(iii) There exist $0 < \underline{\xi}_1 \leq \bar{\xi}_1 < \infty$ such that

$$0 < \underline{\xi}_1 \leq \inf_{\theta \in \mathcal{B}_{\theta_*}(\delta)} \lambda_d \{ \Sigma(\sigma) \} \leq \sup_{\theta \in \mathcal{B}_{\theta_*}(\delta)} \lambda_1 \{ \Sigma(\sigma) \} \leq \bar{\xi}_1 < \infty.$$

Further, there exists $\bar{\xi}'_1$ for all $\alpha \subseteq \mathbb{N}^s$ with $|\alpha| \in \{1, 2\}$,

$$\sup_{\theta \in \mathcal{B}_{\theta_*}(\delta)} \lambda_1 \left\{ \frac{\partial^{|\alpha|}}{\partial \sigma^\alpha} \Sigma(\sigma) \right\} \leq \bar{\xi}'_1 < \infty.$$

(iv) There exist $0 < \underline{\xi}_2 \leq \bar{\xi}_2 < \infty$ such that for each $i \in [m]$,

$$0 < \underline{\xi}_2 n_i \leq \lambda_d \{ \mathbf{D}_i^T \mathbf{D}_i \} \leq \lambda_1 \{ \mathbf{D}_i^T \mathbf{D}_i \} \leq \bar{\xi}_2 n_i < \infty,$$

where $\mathbf{D}_i = [\mathbf{X}_i : \mathbf{V}_i]$ with $\mathbf{X}_i = (\mathbf{x}_{i1}^T, \dots, \mathbf{x}_{in_i}^T)^T$ and $\mathbf{V}_i = (\mathbf{v}_{i1}^T, \dots, \mathbf{v}_{in_i}^T)^T$. Furthermore for some $\epsilon > 0$, $\max_{i,j} \|\mathbf{x}_{ij}\|_2 = O(\log(n_{\min})^{1-\epsilon})$ and $\max_{i,j} \|\mathbf{v}_i\|_2 = O(\log(n_{\min})^{(1-\epsilon)/2})$.

(v) The derivatives $b'(\eta)$ and $b''(\eta)$ are $o(\exp(\eta^{2-\epsilon}))$ and for all $C > 0$

$$\sup_{|\eta| < C \log(n)^{1-\epsilon}} |b^{(k)}(\eta)| \leq n^\epsilon$$

for all $\epsilon > 0$ and $k \in \mathbb{N}$.

(vi) For all $C > 0$ and all $\epsilon > 0$, $\sup_{|\eta| < C \log(n)^{1-\epsilon}} \mathbb{E}[|c(Y; \phi)|^k] = O(n^\epsilon)$, for all integers k such that $k < q + 1$, where the expectation is taken with respect to the distribution of Y with natural parameter η .

Assumption 9 (Well-Specified and Consistency). The GLMM is well-specified if there exist $\theta_* \in \Theta$ and $\mathbf{u}_{*,i} \in \mathcal{U}$ for each $i \in [m]$ such that for each $j \in [n_i]$, $y_{ij} \mid \mathbf{x}_{ij}, \mathbf{v}_{ij} \stackrel{\text{ind}}{\sim} F(\mathbf{x}_{ij}^T \boldsymbol{\beta}_* + \mathbf{v}_{ij}^T \mathbf{u}_{*,i})$. Furthermore assume that the marginal maximum likelihood estimator for θ_* is consistent at a rate of m^ϵ for some $\epsilon > 0$.

Remark 4. In exponential models considered, the derivatives of the log-likelihood are functions of the natural parameter and the restriction of $|\eta| < C \log(n)^{1-\epsilon}$ in Assumption 8 (ii), (V) and (vi) is used to control the size of these derivatives.

Remark 5. Assumption 8 i) and iv) are standard assumptions, while Assumption 8 v) and vi) are satisfied for most commonly used exponential family such as the Gaussian, Inverse Gaussian, Gamma, Poisson and Binomial distributions. Assumption 8 v) requires that growth rates of the mean and variance of the random variable as a function of the natural parameter η are slower than $\exp(\eta^2)$ which is true of all models listed above. Assumption 8 vi) can be shown to hold for the Gamma distribution through Stirling approximation, for the inverse normal this condition is implied by the fact that the inverse k -th moments exist for all k and for the other listed distributions this can be shown through direct calculation. Example calculations for the Poisson are given in Example 1.

Example 1. We provide sample calculations for Poisson regression to show that Assumption 8 v) and vi) are satisfied. This case is one of the more difficult to check as the mean function is increasing

exponentially in the natural parameter, contrary to most other cases of GLMs used in practice. For Assumption 8 v), note that for the Poisson $b(\eta) = \exp(\eta)$:

$$\sup_{|\eta| < C \log(n)^{1-\epsilon}} b^{(k)}(\eta) = \sup_{|\eta| < C \log(n)^{1-\epsilon}} \exp(\eta) \leq \exp(C \log(n)^{1-\epsilon}) = o(n^\epsilon),$$

for all $\epsilon > 0$. While for the Poisson distribution $c(Y) = \log(Y!) \leq Y^{1+\epsilon'}$ for any arbitrary $\epsilon' > 0$ by Stirling's approximation, therefore for $k \in \mathbb{N}$:

$$\sup_{|\eta| < C \log(n)^{1-\epsilon}} \mathbb{E}[(Y^{1+\epsilon'})^k] \leq \sup_{|\eta| < C \log(n)^{1-\epsilon}} \mathbb{E}[Y^{k+1}] \leq k \exp(C(k+1) \log(n)^{1-\epsilon}) = o(n^\epsilon),$$

for any $\epsilon > 0$ showing the desired result.

Remark 6. The rate of consistency of the marginal maximum likelihood estimator is shown to be at the very least $m^{1/2}$ in Jiang, Wand and Bhaskaran (2021), therefore Assumption 9 can be easily satisfied.

D.2. Proof

We now state a Lemma which will help us bound the natural parameter of the observations uniformly to control key quantities which will appear in the proof.

Lemma 10. Under Assumption 8 (iii) and (iv), for all $\theta \in \mathcal{B}_{\theta_*}(\delta)$ and $\mathbf{u} \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)$ there exist some constants $\bar{\xi}_2^* < \infty$ and $\epsilon > 0$:

$$\lim_{n_{\min} \rightarrow \infty} P_N^* \left[\max_{i,j} \left| \eta_{ij}^\theta(\mathbf{u}) \right| \leq \bar{\xi}_2^* \log(n_{\min})^{1-\epsilon} \right] = 1. \quad (\text{D.1})$$

Proof. For $\theta \in \mathcal{B}_{\theta_*}(\delta)$ and $\mathbf{u} \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)$, the following holds for some unit vector $\mathbf{z}$ and $\mathbf{z}'$

$$\begin{aligned} \left| \eta_{ij}^\theta(\mathbf{u}) \right| &= \left| \mathbf{x}_{ij}^\top (\boldsymbol{\beta}_* + \delta \mathbf{z}) + \mathbf{v}_{ij}^\top (\mathbf{u}_{*,i} + \delta \mathbf{z}') \right| \\ &\leq \|\mathbf{x}_{ij}\|_2 \|\boldsymbol{\beta}_*\|_2 + \delta (\|\mathbf{x}_{ij}\|_2 + \|\mathbf{v}_{ij}\|_2) + \|\mathbf{v}_{ij}\|_2 \|\mathbf{u}_{*,i}\|_2, \end{aligned}$$

where the first two terms are uniformly $O(\log(n)^{1-\epsilon})$ and $O(\log(n)^{\frac{1-\epsilon}{2}})$ by Assumption 8 (iv), therefore it remains to bound the final term uniformly in probability. Note that $\mathbf{u}_{*,i}$ follows a zero mean Gaussian whose covariance matrix has eigenvalues bounded by $\bar{\xi}_1$ by Assumption 8 (iii), therefore for each i , $\|\mathbf{u}_{*,i}\|_2^2$ is stochastically dominated by a $\bar{\xi}_1 \chi_{d,i}^2$ Chi-squared random variable. By Lemma D.2 in Tang and Reid (2024), $\max_{i=1,\dots,m} (\chi_{d,i}^2)^{1/2} = O(\log(m)^{1/2}) = O(\log(n_{\min})^{1/2})$ with probability $O(1/m)$. Noting that $m \rightarrow \infty$ and that by Assumption $\max_{i,j} \|\mathbf{v}_i\|_2 = O(\log(n)^{\frac{1-\epsilon}{2}})$ shows the desired result. $\square$

We now control the probability that certain averages used in the subsequent proofs will be close to their expected value jointly as $m \rightarrow \infty$ and $n \rightarrow \infty$; recall that $m = n_{\min}^q$ for $q > 0$. The ϵ factor is needed here for the uniform control of the likelihood derivatives as the number of groups increases.

Lemma 11. Under Assumption 8, there exist constants $B_1, B_2, B_3 > 0$ and a polynomial function $G(x)$ where $\lim_{x \rightarrow \infty} G(x) = \infty$ such that for all $\epsilon > 0$:

- $\lim_{n_{\min} \rightarrow \infty} P_N^* \left(\forall_{i=1, \dots, m} \frac{1}{n_i} \sum_{j=1}^{n_i} |y_{ij}| > B_1 n_{\min}^\epsilon \right) \rightarrow 0,$
- $\lim_{n_{\min} \rightarrow \infty} P_N^* \left(\forall_{i=1, \dots, m} \forall_{k=1, \dots, d} \left| \frac{1}{n_i} \sum_{j=1}^{n_i} \{y_{ij} \mathbf{v}_{jk} - \mathbb{E} y_{ij} \mathbf{v}_{jk}\} \right| > \frac{B_2}{G(n_{\min})} \right) \rightarrow 0.$
- $\lim_{n_{\min} \rightarrow \infty} P_N^* \left(\forall_{i=1, \dots, m} \forall_{k=1, \dots, d} \left| \frac{1}{\sqrt{n_i}} \sum_{j=1}^{n_i} \{y_{ij} \mathbf{v}_{jk} - \mathbb{E} y_{ij} \mathbf{v}_{jk}\} \right| > B_2 n_{\min}^\epsilon \right) \rightarrow 0.$
- $\lim_{n_{\min} \rightarrow \infty} P_N^* \left(\forall_{i=1, \dots, m} \frac{1}{n_i} \sum_{j=1}^{n_i} |c(y_{ij}; \phi)| > B_3 n_{\min}^\epsilon \right) \rightarrow 0.$

Proof. These four inequalities will be proven with Markov's inequality and a union bound. For each i fixed and for any $t > 0$,

$$P_N^* \left(\frac{1}{n_i} \sum_{j=1}^{n_i} |y_{ij}| > t \right) \leq P_N^* \left(\sqrt{\frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}^2} > t \right) = P_N^* \left(\frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}^2 > t^2 \right),$$

as $\sum_{j=1}^{n_i} |y_{ij}| \leq \sqrt{n_i} \sqrt{\sum_{j=1}^{n_i} y_{ij}^2}$. First consider the centred version of these sums, for some $t' > 0$ and any even integer a ,

$$\begin{aligned} P_N^* \left(\frac{1}{n_i} \sum_{j=1}^{n_i} (y_{ij}^2 - \mathbb{E} y_{ij}^2) > t' \right) &= P_N^* \left(\left\{ \frac{1}{n_i} \sum_{j=1}^{n_i} (y_{ij}^2 - \mathbb{E} y_{ij}^2) \right\}^a > (t')^a \right) \\ &\leq \frac{\mathbb{E} \left(\left\{ \sum_{j=1}^{n_i} (y_{ij}^2 - \mathbb{E} y_{ij}^2) \right\}^a \right)}{n_i^a (t')^a} \leq \frac{D_a n_{\min}^\epsilon}{n_i^{a/2} (t')^a}, \end{aligned}$$

for some constant which only depends on a as only terms with even parity will be non-zero in above expectation by independence the observation for fixed i and from Lemma 10 and Assumption 8 v) the moments of y_{ij}^2 are uniformly bounded by any polynomial of $n_{\min}$. Next, it is also the case that

$$\frac{1}{n_i} \sum_{j=1}^{n_i} \mathbb{E} y_{ij}^2 = O(n_{\min}^\epsilon)$$

uniformly in i by Lemma 10 and Assumption 8 v). Combining these bounds and letting $t' = 1$,

$$P_N^* \left(\forall_{i=1, \dots, m} \frac{1}{n_i} \sum_{j=1}^{n_i} |y_{ij}| > B_1 \right) \leq P_N^* \left(\forall_{i=1, \dots, m} \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}^2 > 1 + \epsilon \right) \leq \frac{m D_a}{n_{\min}^{a/2 - \epsilon}} \rightarrow 0,$$

by the Assumption that $m = n_{\min}^q$ for some $q > 0$, $m = o(n_{\min}^{a/2 - \epsilon})$ for some even value of a .

For the proof of the second statement, we use Lemma 10 and Assumption 8 v) to obtain

$$P_N^* \left(\forall_{i=1, \dots, m} \forall_{k=1, \dots, d} \left| \frac{1}{n_i} \sum_{j=1}^{n_i} \{y_{ij} \mathbf{v}_{jk} - \mathbb{E} y_{ij} \mathbf{v}_{jk}\} \right| > t \right) \leq \frac{d m D_a}{n_{\min}^{a/2 - \epsilon} t^a},$$

from similar arguments as above. Letting $t = o(n_{\min}^{-q'})$ for $q' < a/2 - q - \epsilon$ gives the desired result. The third statement follows by simple substitution and noting that by Assumption 8 v) a in the above bound can be made arbitrarily large and ϵ can be made arbitrarily small. For any $\epsilon' > 0$,

$$P_N^* \left(\forall i=1, \dots, m \quad \forall k=1, \dots, d \quad \left| \frac{1}{\sqrt{n_i}} \sum_{j=1}^{n_i} \{y_{ij} \mathbf{v}_{jk} - \mathbb{E} y_{ij} \mathbf{v}_{jk}\} \right| > n^{\epsilon'} \right) \leq \frac{dm D_a n_{\min}^{\epsilon}}{n_{\min}^{a\epsilon'}} \rightarrow 0,$$

for a such that $a\epsilon' > q + 1$. The final statement can be similarly shown with the same steps as in proof of the first statement and Assumption 8 vi) and Lemma 10. $\square$

Lemma 12. *For any GLMM satisfying Assumption 8, there exist $D > 0$ such that for any $\epsilon > 0$,*

$$\lim_{n_{\min} \rightarrow \infty} P_N^* \left[\forall i=1, \dots, m \quad \sup_{\theta \in \mathcal{B}_{\theta_*}(\delta)} \sup_{\mathbf{u} \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} |\ell_i^{\theta}(\mathbf{u})| < n_i^{1+\epsilon} \cdot D \right] = 1.$$

Proof. Fix $i \in [m]$. The joint log-likelihood is, up to constants,

$$\begin{aligned} \ell_i^{\theta}(\mathbf{u}) &= \sum_{j=1}^{n_i} \log f(y_{ij} | \eta_{ij}^{\theta}(\mathbf{u})) g(\mathbf{u} | \sigma), \\ &= \frac{1}{a(\phi)} \sum_{j=1}^{n_i} \left\{ y_{ij} \eta_{ij}^{\theta}(\mathbf{u}) - b(\eta_{ij}^{\theta}(\mathbf{u})) \right\} - \frac{1}{2} \log |\Sigma(\sigma)| - \frac{1}{2} \mathbf{u}^T \Sigma^{-1}(\sigma) \mathbf{u}, \end{aligned}$$

where $\eta_{ij}^{\theta}(\mathbf{u}) = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{v}_{ij}^T \mathbf{u}$. For all $\theta \in \mathcal{B}_{\theta_*}(\delta)$ and $\mathbf{u} \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)$,

$$\begin{aligned} \frac{1}{n_i} |\ell_i^{\theta}(\mathbf{u})| &\leq \frac{1}{2n_i} \left\{ \log |\Sigma(\sigma)| + \mathbf{u}^T \Sigma^{-1}(\sigma) \mathbf{u} \right\} \\ &\quad + \frac{1}{a(\phi)n_i} \sum_{j=1}^{n_i} \left\{ |y_{ij} \eta_{ij}^{\theta}(\mathbf{u})| + |b(\eta_{ij}^{\theta}(\mathbf{u}))| \right\} \\ &\quad + \frac{1}{n_i} \sum_{j=1}^{n_i} |c(y_{ij}; \phi)|. \end{aligned}$$

We bound each term individually. For the first term,

$$\begin{aligned} &\sup_{\theta \in \mathcal{B}_{\theta_*}(\delta)} \sup_{\mathbf{u} \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} \frac{1}{2} \left\{ \log |\Sigma(\sigma)| + \mathbf{u}^T \Sigma^{-1}(\sigma) \mathbf{u} \right\} \\ &\leq \frac{1}{2} \left(d \log \bar{\xi}_1 + \bar{\xi}_1^{-1} \sup_{\mathbf{u} \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} \|\mathbf{u}\|_2^2 \right) \leq \frac{1}{2} \left(d \log \bar{\xi}_1 + \bar{\xi}_1^{-1} (\|\mathbf{u}_*\|_2 + \delta)^2 \right), \end{aligned}$$

by Assumption 8(iv). Note that $\|\mathbf{u}_{*,i}\|^2 = O_p(d)$ as $\mathbf{u}_*$ follows a multivariate Gaussian distribution with bounded eigenvalues. Therefore:

$$\frac{1}{2n_i} \left\{ \log |\Sigma(\sigma)| + \mathbf{u}^T \Sigma^{-1}(\sigma) \mathbf{u} \right\} = O_p \left(\frac{1}{n_i} \right).$$

For the second term, by Lemma 10,

$$\begin{aligned} & \sup_{\theta \in \mathcal{B}_{\theta_*}(\delta)} \sup_{\mathbf{u} \in \mathcal{B}_{\mathbf{u}_*, i}(\delta)} \frac{1}{a(\boldsymbol{\phi})n_i} \sum_{j=1}^{n_i} \left\{ \left| y_{ij} \eta_{ij}^\theta(\mathbf{u}) \right| + \left| b(\eta_{ij}^\theta(\mathbf{u})) \right| \right\} \\ & \leq \frac{1}{a(\boldsymbol{\phi})} \left\{ \bar{\xi}_2^* n_{\min}^\varepsilon \frac{1}{n_i} \sum_{j=1}^{n_i} |y_{ij}| + \frac{1}{n_i} \sup_{|\eta| < \bar{\xi}_2^* \log(n_{\min})^{1-\varepsilon}} |b(\eta)| \right\} \\ & \leq \frac{\bar{\xi}_2^*}{a(\boldsymbol{\phi})} (B_1 n_{\min}^\varepsilon + O(n_{\min}^\varepsilon)) = O(n_{\min}^\varepsilon), \end{aligned}$$

with probability tending to 1 uniformly in the index i by Lemmas 10 and 11 and Assumption 8 v).

For the third term,

$$\frac{1}{n_i} \sum_{j=1}^{n_i} |c(y_{ij}; \boldsymbol{\phi})| \leq n_{\min}^\varepsilon B_3$$

with probability tending to 1 uniformly in the index i by Lemma 11.

Combining these bounds, we have that for all i

$$\lim_{n_i \rightarrow \infty} \frac{1}{n_i} |\ell_i^\theta(\mathbf{u})| = O(n_{\min}^\varepsilon)$$

with probability 1 for arbitrarily small $\varepsilon > 0$, which implies the result. $\square$

We now proceed with the main proof of Proposition 1.

Fix $i \in [m]$. The derivatives of $\ell_i^\theta(\mathbf{u})$ are

$$\partial_{\mathbf{u}}^\alpha \ell_i^\theta(\mathbf{u}) = \begin{cases} \frac{1}{a(\boldsymbol{\phi})} \sum_{j=1}^{n_i} \left\{ y_{ij} - b'(\eta_{ij}^\theta(\mathbf{u})) \right\} \mathbf{v}_{ij1}^{\alpha_1} \cdots \mathbf{v}_{ijd}^{\alpha_d} + \sum_{t=1}^d \left(\boldsymbol{\Sigma}(\boldsymbol{\sigma})_{\alpha_t}^{-1} \right)^{\alpha_t} \mathbf{u} & |\alpha| = 1, \\ -\frac{1}{a(\boldsymbol{\phi})} \sum_{j=1}^{n_i} b^{(|\alpha|)}(\eta_{ij}^\theta(\mathbf{u})) \mathbf{v}_{ij1}^{\alpha_1} \cdots \mathbf{v}_{ijd}^{\alpha_d} + \sum_{t=1}^d \sum_{s=1}^d \left(\boldsymbol{\Sigma}(\boldsymbol{\sigma})_{\alpha_t \alpha_s}^{-1} \right)^{\alpha_t \alpha_s} & |\alpha| = 2, \\ -\frac{1}{a(\boldsymbol{\phi})} \sum_{j=1}^{n_i} b^{(|\alpha|)}(\eta_{ij}^\theta(\mathbf{u})) \mathbf{v}_{ij1}^{\alpha_1} \cdots \mathbf{v}_{ijd}^{\alpha_d} & |\alpha| \geq 3. \end{cases}$$

Assumption 2.

The $|\alpha| = 0$ case follows directly from Lemma 12. Suppose $|\alpha| = 1$. Then, by Assumption 8(ii) and Lemma 10,

$$\sup_{\theta \in \mathcal{B}_{\theta_*}(\delta)} \sup_{\mathbf{u} \in \mathcal{B}_{\mathbf{u}_*, i}(\delta)} \partial_{\mathbf{u}}^\alpha \ell_i^\theta(\mathbf{u}) \leq \frac{M}{a(\boldsymbol{\phi})} \left\{ \sum_{j=1}^{n_i} |y_{ij}| + \sup_{|\eta| < \bar{\xi}_2^* \log(n_{\min})^{1-\varepsilon}} |b'(\eta)| \right\},$$

from which it follows that for all i

$$\lim_{n_i \rightarrow \infty} \frac{1}{n_i} \sup_{\theta \in \mathcal{B}_{\theta_*}(\delta)} \sup_{\mathbf{u} \in \mathcal{B}_{\mathbf{u}_*, i}(\delta)} \partial_{\mathbf{u}}^\alpha \ell_i^\theta(\mathbf{u}) \leq \frac{M}{a(\boldsymbol{\phi})} \left\{ B_1 n_{\min}^\varepsilon + \sup_{|\eta| < \bar{\xi}_2^* \log(n_{\min})^{1-\varepsilon}} |b'(\eta)| \right\} = O(n_{\min}^\varepsilon),$$

for all $\varepsilon > 0$ with probability tending to 1 by Lemmas 10 and 11.

If $|\alpha| = 2$ then

$$\begin{aligned} & \frac{1}{n_i} \sup_{\theta \in \mathcal{B}_{\theta_*}(\delta)} \sup_{\mathbf{u} \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} \partial_{\mathbf{u}}^{\alpha} \ell_i^{\theta}(\mathbf{u}) \\ & \leq \frac{M^{|\alpha|}}{a(\boldsymbol{\phi})} \left\{ \sup_{|\eta| < \bar{\xi}_2^* \log(n_{\min})^{1-\varepsilon}} \left| b^{(|\alpha|)}(\eta) \right| \right\} + \sup_{\theta \in \mathcal{B}_{\theta_*}(\delta)} \sup_{\mathbf{u} \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} \left| \sum_{t=1}^d \left(\boldsymbol{\Sigma}(\sigma)_{\alpha_t}^{-1} \right)^{\alpha_t \cdot} \mathbf{u} \right| \end{aligned}$$

and if $|\alpha| \geq 3$ then

$$\begin{aligned} & \frac{1}{n_i} \sup_{\theta \in \mathcal{B}_{\theta_*}(\delta)} \sup_{\mathbf{u} \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} \partial_{\mathbf{u}}^{\alpha} \ell_i^{\theta}(\mathbf{u}) \\ & \leq \frac{M^{|\alpha|}}{a(\boldsymbol{\phi})} \left\{ \sup_{|\eta| < \bar{\xi}_2^* \log(n_{\min})^{1-\varepsilon}} \left| b^{(|\alpha|)}(\eta) \right| \right\} + \sup_{\theta \in \mathcal{B}_{\theta_*}(\delta)} \sup_{\mathbf{u} \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} \left| \sum_{t=1}^d \sum_{s=1}^d \left(\boldsymbol{\Sigma}(\sigma)_{\alpha_t \alpha_s}^{-1} \right)^{\alpha_t \alpha_s} \right| \\ & = O(n^{\varepsilon}) + \sup_{\theta \in \mathcal{B}_{\theta_*}(\delta)} \sup_{\mathbf{u} \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} \left| \sum_{t=1}^d \sum_{s=1}^d \left(\boldsymbol{\Sigma}(\sigma)_{\alpha_t \alpha_s}^{-1} \right)^{\alpha_t \alpha_s} \right| \end{aligned}$$

for all $\varepsilon > 0$ by Lemma 10. Furthermore we have that, by Assumption 8 (iv),

$$\sup_{\theta \in \mathcal{B}_{\theta_*}(\delta)} \sup_{\mathbf{u} \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} \left| \sum_{t=1}^d \left(\boldsymbol{\Sigma}(\sigma)_{\alpha_t}^{-1} \right)^{\alpha_t \cdot} \mathbf{u} \right| < \bar{\xi}_{-1}^{-1} \delta < \infty$$

and $\sum_{t=1}^d \sum_{s=1}^d \left(\boldsymbol{\Sigma}(\sigma)_{\alpha_t \alpha_s}^{-1} \right)^{\alpha_t \alpha_s} < \bar{\xi}_{-1}^{-1} < \infty$. Assumption 2 now follows by noting that these bounds are independent of i .

Assumption 3

The Hessian is

$$\begin{aligned} -\partial_{(\theta, \mathbf{u})}^2 \ell_i^{\theta}(\mathbf{u}_i) &= \mathbf{D}_i^{\top} \mathbf{W}_i^{\theta}(\mathbf{u}_i) \mathbf{D}_i + \partial_{(\theta, \mathbf{u})}^2 \left\{ \frac{1}{2} \log |\boldsymbol{\Sigma}(\sigma)| + \frac{1}{2} \mathbf{u}_i^{\top} \boldsymbol{\Sigma}^{-1}(\sigma) \mathbf{u}_i \right\}, \\ &\equiv \mathbf{H}_i^{\theta}(\mathbf{u}_i) + \mathbf{G}_i^{\theta}(\mathbf{u}_i), \end{aligned}$$

where $\mathbf{D}_i = [\mathbf{X}_i : \mathbf{V}_i]$ as defined in Assumption 8 and $\mathbf{W}_i^{\theta}(\mathbf{u}_i) = \text{diag} \{ b''(\eta_{i1}^{\theta}(\mathbf{u}_i)), \dots, b''(\eta_{i n_i}^{\theta}(\mathbf{u}_i)) \}$. We use the elementary fact that if A and B are $p \times p$ real symmetric matrices then $\lambda_p(A) + \lambda_p(B) \leq \lambda_p(A+B) \leq \lambda_1(A+B) \leq \lambda_1(A) + \lambda_1(B)$. Note that this also implies that the eigenvalues of a block matrix are upper and lower bounded by the eigenvalues of the blocks.

First, for all $\theta \in \mathcal{B}_{\theta_*}(\delta)$,

$$\begin{aligned} \sup_{\theta \in \mathcal{B}_{\theta_*}(\delta)} \sup_{\mathbf{u}_i \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} \lambda_1(\mathbf{H}_i^\theta(\mathbf{u}_i)) &= \sup_{\theta \in \mathcal{B}_{\theta_*}(\delta)} \sup_{\mathbf{u}_i \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} \left\| \mathbf{W}_i^\theta(\mathbf{u}_i)^{1/2} \mathbf{D}_i \right\|_2^2, \\ &\leq \sup_{\theta \in \mathcal{B}_{\theta_*}(\delta)} \sup_{\mathbf{u}_i \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} \left\| \mathbf{W}_i^\theta(\mathbf{u}_i)^{1/2} \right\|_2^2 \left\| \mathbf{D}_i \right\|_2^2 \\ &\leq n_i \sup_{|\eta| \leq \bar{\xi}_2^* \log(n_{\min})^{1-\varepsilon}} |b''(\eta)| M^2, \end{aligned}$$

where we have used Assumption 8(ii) and Lemma 10. That is,

$$\frac{1}{n_i} \sup_{\theta \in \mathcal{B}_{\theta_*}(\delta)} \sup_{\mathbf{u}_i \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} \lambda_1(\mathbf{H}_i^\theta(\mathbf{u}_i)) \leq O(n_{\min}^\varepsilon),$$

with probability tending to 1 by Lemma 10. Further, for all $\theta \in \mathcal{B}_{\theta_*}(\delta)$,

$$\begin{aligned} \inf_{\theta \in \mathcal{B}_{\theta_*}(\delta)} \inf_{\mathbf{u}_i \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} \lambda_d(\mathbf{H}_i^\theta(\mathbf{u}_i)) &= \inf_{\theta \in \mathcal{B}_{\theta_*}(\delta)} \inf_{\mathbf{u}_i \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} \inf_{\mathbf{z}: \|\mathbf{z}\|_2=1} \mathbf{z}^\top \mathbf{D}_i^\top \mathbf{W}_i^\theta(\mathbf{u}_i) \mathbf{D}_i \mathbf{z} \\ &= \inf_{\theta \in \mathcal{B}_{\theta_*}(\delta)} \inf_{\mathbf{u}_i \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} \inf_{\mathbf{z}: \|\mathbf{z}\|_2=1} \left\| \mathbf{W}_i^\theta(\mathbf{u}_i)^{1/2} \mathbf{D}_i \mathbf{z} \right\|_2^2 \\ &\geq \inf_{|\eta| \leq \bar{\xi}_2^* \log(n_{\min})^{1-\varepsilon}} |b''(\eta)| \inf_{\mathbf{z}: \|\mathbf{z}\|_2=1} \left\| \mathbf{D}_i \mathbf{z} \right\|_2^2 \\ &= \inf_{|\eta| \leq \bar{\xi}_2^* \log(n_{\min})^{1-\varepsilon}} |b''(\eta)| \lambda_1(\mathbf{D}_i^\top \mathbf{D}_i) \\ &\geq \underline{\xi}_2 n_i \inf_{|\eta| \leq \bar{\xi}_2^* \log(n_{\min})^{1-\varepsilon}} |b''(\eta)|, \end{aligned}$$

where the last step follows from Assumption 8(v). Since $\inf_{|\eta| \leq \bar{\xi}_2^* \log(n_{\min})^{1-\varepsilon}} |b''(\eta)| > n_{\min}^{-\varepsilon}$ for any $\varepsilon > 0$ by Assumption 8(iii), we conclude that

$$\frac{1}{n_i} \inf_{\theta \in \mathcal{B}_{\theta_*}(\delta)} \inf_{\mathbf{u}_i \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} \lambda_d(\mathbf{H}_i^\theta(\mathbf{u}_i)) > n_{\min}^{-\varepsilon}.$$

For $\mathbf{G}_i^\theta(\mathbf{u}_i)$, its second order derivative is made of the following three pieces which we now evaluate:

$$\begin{aligned} -\partial_{\mathbf{u}_i \mathbf{u}_i^\top}^2 \left\{ \frac{1}{2} \log |\boldsymbol{\Sigma}(\boldsymbol{\sigma})| + \frac{1}{2} \mathbf{u}_i^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\sigma}) \mathbf{u}_i \right\} &= \boldsymbol{\Sigma}(\boldsymbol{\sigma})^{-1}, \\ -\partial_{(\theta_j) \mathbf{u}_i^\top}^2 \left\{ \frac{1}{2} \log |\boldsymbol{\Sigma}(\boldsymbol{\sigma})| + \frac{1}{2} \mathbf{u}_i^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\sigma}) \mathbf{u}_i \right\} &= -\boldsymbol{\Sigma}(\boldsymbol{\sigma})^{-1} \left\{ \partial_{\boldsymbol{\sigma}_{(j-q+l)}} \boldsymbol{\Sigma}(\boldsymbol{\sigma}) \right\} \boldsymbol{\Sigma}(\boldsymbol{\sigma})^{-1} \mathbf{u}_i \\ -\partial_{\theta_j \theta_l}^2 \left\{ \frac{1}{2} \log |\boldsymbol{\Sigma}(\boldsymbol{\sigma})| + \frac{1}{2} \mathbf{u}_i^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\sigma}) \mathbf{u}_i \right\} \\ &= \frac{1}{2} \text{Tr} \left\{ \boldsymbol{\Sigma}(\boldsymbol{\sigma})^{-1} \partial_{\boldsymbol{\sigma}_{(j-q+l)} \boldsymbol{\sigma}_{(l-q+l)}}^2 \boldsymbol{\Sigma}(\boldsymbol{\sigma}) \right\} - \frac{1}{2} \mathbf{u}_i^\top \left[\boldsymbol{\Sigma}(\boldsymbol{\sigma})^{-1} \left\{ \partial_{\boldsymbol{\sigma}_{(j-q+l)} \boldsymbol{\sigma}_{(l-q+l)}}^2 \boldsymbol{\Sigma}(\boldsymbol{\sigma}) \right\} \boldsymbol{\Sigma}(\boldsymbol{\sigma})^{-1} \right] \mathbf{u}_i, \end{aligned}$$

for $j, l = q + l + 1, \dots, q + l + s$; the respective terms are 0 otherwise. This matrix does not depend on n_i , and Assumption 8 (iv) is sufficient to upper and lower bound its spectrum. Note that for the first term,

$$0 < \bar{\xi}_1^{-1} \leq \inf_{\theta \in \mathcal{B}_{\theta_*}(\delta)} \lambda_d \left\{ \Sigma(\sigma)^{-1} \right\} \leq \sup_{\theta \in \mathcal{B}_{\theta_*}(\delta)} \lambda_1 \left\{ \Sigma(\sigma)^{-1} \right\} \leq \underline{\xi}_1^{-1} < \infty.$$

We can bound the second term by

$$\left\| \Sigma(\sigma)^{-1} \left\{ \partial_{\sigma(j-q+l)} \Sigma(\sigma) \right\} \Sigma(\sigma)^{-1} \mathbf{u}_i \right\|_2 \leq \left\| \Sigma(\sigma)^{-1} \right\|_2^2 \|\mathbf{u}_i\|_2 \left\| \partial_{\sigma(j-q+l)} \Sigma(\sigma) \right\|_2,$$

where $\|A\|_2 \equiv \|A\|_{op} = \lambda_1(A)^{1/2}$. It follows that

$$\sup_{\theta \in \mathcal{B}_{\theta_*}(\delta)} \sup_{\mathbf{u}_i \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} \left\| -\Sigma(\sigma)^{-1} \left\{ \partial_{\sigma(j-q+l)} \Sigma(\sigma) \right\} \Sigma(\sigma)^{-1} \mathbf{u}_i \right\|_2 = O_p(\sqrt{\log(m)}),$$

uniformly in i as the maximum of $\|\mathbf{u}_i\|$ is of order $O_p(\sqrt{\log(m)})$ by Theorem 1.14 in [Rigollet and Hütter \(2023\)](#) and Assumption 8 iii). Further, note that for $A \in \mathbb{R}^{d \times d}$, $\|A\|_2 \leq d \|A\|_+$ where $\|A\|_+ = \max_{1 \leq i, j \leq d} |A_{ij}|$. Since the dimension d is fixed, it suffices to bound the third block element-wise. For the trace term, we apply Holder's inequality and the fact that $\text{Tr}(A) = \lambda_1(A) + \dots + \lambda_d(A)$ to write

$$\begin{aligned} \text{Tr} \left\{ \Sigma(\sigma)^{-1} \partial_{\sigma(j-q+l)\sigma(l-q+l)}^2 \Sigma(\sigma) \right\} &\leq \left\| \Sigma(\sigma)^{-1} \right\|_{op} \text{Tr} \left\{ \partial_{\sigma(j-q+l)\sigma(l-q+l)}^2 \Sigma(\sigma) \right\}, \\ &\leq d \lambda_1 \left\{ \Sigma(\sigma)^{-1} \right\}^{1/2} \lambda_1 \left\{ \partial_{\sigma(j-q+l)\sigma(l-q+l)}^2 \Sigma(\sigma) \right\}, \\ &= d \lambda_p \left\{ \Sigma(\sigma)^{-1} \right\}^{-1/2} \lambda_1 \left\{ \partial_{\sigma(j-q+l)\sigma(l-q+l)}^2 \Sigma(\sigma) \right\}, \\ &\leq d \bar{\xi}_1^{-1/2} \bar{\xi}_1'. \end{aligned}$$

We conclude that

$$\sup_{\theta \in \mathcal{B}_{\theta_*}(\delta)} \sup_{\mathbf{u}_i \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} \text{Tr} \left\{ \Sigma(\sigma)^{-1} \partial_{\sigma(j-q+l)\sigma(l-q+l)}^2 \Sigma(\sigma) \right\} \leq d \bar{\xi}_1^{-1/2} \bar{\xi}_1' < \infty$$

by Assumption 8 (iv).

Using $\lambda_1(AB) \leq \lambda_1(A)\lambda_1(B)$ for any real, symmetric, positive definite A and B , we also have

$$\begin{aligned} \mathbf{u}_i^\top \left[\Sigma(\sigma)^{-1} \left\{ \partial_{\sigma(j-q+l)\sigma(l-q+l)}^2 \Sigma(\sigma) \right\} \Sigma(\sigma)^{-1} \right] \mathbf{u}_i &\leq \lambda_1 \left\{ \Sigma(\sigma)^{-1} \left\{ \partial_{\sigma(j-q+l)\sigma(l-q+l)}^2 \Sigma(\sigma) \right\} \Sigma(\sigma)^{-1} \right\}, \\ &\leq \lambda_1 \left\{ \Sigma(\sigma)^{-1} \right\}^2 \lambda_1 \left\{ \partial_{\sigma(j-q+l)\sigma(l-q+l)}^2 \Sigma(\sigma) \right\}, \\ &\leq \bar{\xi}_1^{-2} \bar{\xi}_1' \|\mathbf{u}_i\|_2^2, \end{aligned}$$

and so by Assumption 8 (iv) we conclude that

$$\begin{aligned} &\sup_{\theta \in \mathcal{B}_{\theta_*}(\delta)} \sup_{\mathbf{u}_i \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} \mathbf{u}_i^\top \left[\Sigma(\sigma)^{-1} \left\{ \partial_{\sigma(j-q+l)\sigma(l-q+l)}^2 \Sigma(\sigma) \right\} \Sigma(\sigma)^{-1} \right] \mathbf{u}_i \\ &\leq \bar{\xi}_1^{-2} \bar{\xi}_1' \|\mathbf{u}_i + \delta Z\|_2^2 = O_p(\log(m)), \end{aligned}$$

uniformly in i , where z is some unit vector and the last statement follows again from the fact that $\mathbf{u}_i$ follows a zero mean Gaussian distribution. We have therefore shown that

$$\lambda_1 \{ \mathbf{G}_i^\theta(\mathbf{u}_i) \} = O_p(\log(m)).$$

Combined with the fact that $\mathbf{G}_i^\theta(\mathbf{u})$ does not depend on n_i , we conclude that

$$\lambda_1 \{ n_i^{-1} \mathbf{G}_i^\theta(\mathbf{u}_i) \} = O_p(n_i^{-1} \log(m)),$$

and hence

$$\lambda_d \{ n_i^{-1} \mathbf{G}_i^\theta(\mathbf{u}_i) \} = O_p(n_i^{-1} \log(m))$$

as well for all $\varepsilon > 0$. Noting that these bounds are either independent of or uniform in i , Assumption 3 now follows from combining these bounds.

Assumption 4 We first prove the following two lemmas:

Lemma 13. *For all $i = 1, \dots, m$, $-\log \pi_i(\mathbf{y}_i | \mathbf{u}; \theta)$ are almost surely globally convex as a function of $\mathbf{u} \in \mathcal{U} \subseteq \mathbb{R}^d$ for each $\theta \in \mathcal{B}_{\theta_*}(\delta)$.*

Proof. We have

$$\pi(\mathbf{y}_i, \mathbf{u}_i; \theta) = \prod_{j=1}^{n_i} f(y_{ij} | \mathbf{u}_i; \beta) g(\mathbf{u}_i; \sigma).$$

Observe that $-\sum_{j=1}^{n_i} \log f(y_{ij} | \eta)$ is almost surely globally convex in $\eta \in \Xi$ (see Assumption 8 (i)), which follows immediately from the identity $\mathbf{H}_i^\theta(\mathbf{u}) = \mathbf{D}_i^\top \mathbf{W}_i^\theta(\mathbf{u}) \mathbf{D}_i$ shown in the verification of Assumption 3 and from Assumption 8 (iii). Note as well that η is linear, and hence convex, in $\mathbf{u}$. Further, note that $g(\mathbf{u} | \sigma)$ is a Gaussian density and hence $-\log g(\mathbf{u} | \sigma)$ is convex in $\mathbf{u}$. The result follows because sums and compositions of convex functions are convex. $\square$

Lemma 14. *Let $\bar{\mathbf{u}}_i^\theta = \arg \max_{\mathbf{u} \in [\mathcal{B}_{\mathbf{u}_{*,i}}(\delta)]^c} \log \pi_i(\mathbf{y}_i | \mathbf{u}; \theta)$ for arbitrary $\theta \in \mathcal{B}_{\theta_*}(\delta)$. Then*

$$\lim_{n_{\min} \rightarrow \infty} P_N^* \left(\max_{i=1, \dots, m} \left\| \bar{\mathbf{u}}_i^\theta - \mathbf{u}_{*,i} \right\|_2 = \delta \right) = 1.$$

Proof. By Assumption 6,

$$\lim_{n_{\min} \rightarrow \infty} P_N^* \left[\forall i=1, \dots, m \quad \widehat{\mathbf{u}}_i^\theta \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta) \right] = 1,$$

which shows that $\bar{\mathbf{u}}_i^\theta \neq \widehat{\mathbf{u}}_i^\theta$ for large enough N . Define

$$\mathbf{v}_\delta = \mathbf{u}_{*,i} + \frac{\bar{\mathbf{u}}_i^\theta - \mathbf{u}_{*,i}}{\left\| \bar{\mathbf{u}}_i^\theta - \mathbf{u}_{*,i} \right\|_2} \delta,$$

and note that $\left\| \mathbf{v}_\delta - \mathbf{u}_{*,i} \right\|_2 = \delta$. But it follows from Lemma 13 that

$$\log \pi_i(\mathbf{y}_i | \mathbf{u}_{*,i}; \theta) - \log \pi_i(\mathbf{y}_i | \mathbf{v}_\delta; \theta) \leq \log \pi_i(\mathbf{y}_i | \mathbf{u}_{*,i}; \theta) - \log \pi_i(\mathbf{y}_i | \bar{\mathbf{u}}_i^\theta; \theta),$$

almost surely, and hence $\log \pi_i(\mathbf{y}_i \mid \mathbf{v}_\delta; \boldsymbol{\theta}) \geq \log \pi_i(\mathbf{y}_i \mid \bar{\mathbf{u}}_i^\theta; \boldsymbol{\theta})$, which in turn implies that $\mathbf{v}_\delta = \bar{\mathbf{u}}_i^\theta$ by definition of $\bar{\mathbf{u}}_i^\theta$. The proof is completed by noting that this argument holds for every i uniformly. $\square$

We proceed with the main verification of Assumption 4. Fix a value of $\boldsymbol{\theta} \in \mathcal{B}_{\boldsymbol{\theta}_*}(\zeta_N)$. Taylor expansion yields:

$$\begin{aligned} & \frac{1}{n_i} \{ \log \pi_i(\mathbf{y}_i \mid \mathbf{u}_i; \boldsymbol{\theta}) - \log \pi_i(\mathbf{y}_i \mid \mathbf{u}_{*,i}; \boldsymbol{\theta}) \} \\ &= \frac{1}{n_i} \{ (\mathbf{u}_i - \mathbf{u}_{*,i})^\top \partial_{\mathbf{u}} \log \pi_i(\mathbf{y}_i \mid \mathbf{u}_{*,i}; \boldsymbol{\theta}) + (\mathbf{u}_i - \mathbf{u}_{*,i})^\top \partial_{\mathbf{u}} \partial_{\mathbf{u}}^\top \log \pi_i(\mathbf{y}_i \mid \mathbf{u}^\circ; \boldsymbol{\theta}) (\mathbf{u}_i - \mathbf{u}_{*,i}) \} \\ &= \frac{1}{n_i} \{ (\mathbf{u}_i - \mathbf{u}_{*,i})^\top \partial_{\mathbf{u}} \log \pi_i(\mathbf{y}_i \mid \mathbf{u}_{*,i}; \boldsymbol{\theta}_*) + (\mathbf{u}_i - \mathbf{u}_{*,i})^\top \partial_{\mathbf{u}} \partial_{\boldsymbol{\theta}}^\top \log \pi_i(\mathbf{y}_i \mid \mathbf{u}_{*,i}; \boldsymbol{\theta}^\circ) (\boldsymbol{\theta} - \boldsymbol{\theta}_*) \} \\ &+ \frac{1}{n_i} (\mathbf{u}_i - \mathbf{u}_{*,i})^\top \partial_{\mathbf{u}} \partial_{\mathbf{u}}^\top \log \pi_i(\mathbf{y}_i \mid \mathbf{u}^\circ; \boldsymbol{\theta}) (\mathbf{u}_i - \mathbf{u}_{*,i}), \end{aligned}$$

where $\boldsymbol{\theta}^\circ = t_1 \boldsymbol{\theta} + (1 - t_1) \boldsymbol{\theta}_*$ and $\mathbf{u}^\circ = t_2 \mathbf{u}_i + (1 - t_2) \mathbf{u}_{*,i}$ for some $t_1, t_2 \in [0, 1]$. Our approach is to use this expansion to argue that

$$\lim_{n_i \rightarrow \infty} \frac{1}{n_i} \{ \log \pi_i(\mathbf{y}_i \mid \bar{\mathbf{u}}_i^\theta; \boldsymbol{\theta}) - \log \pi_i(\mathbf{y}_i \mid \mathbf{u}_{*,i}; \boldsymbol{\theta}) \} < 0,$$

which implies the result.

To bound the first term, observe that

$$\begin{aligned} & \max_{i=1, \dots, m} \left| (\bar{\mathbf{u}}_i^\theta - \mathbf{u}_{*,i})^\top \left\{ \frac{\partial_{\mathbf{u}} \log \pi_i(\mathbf{y}_i \mid \mathbf{u}_{*,i}; \boldsymbol{\theta}_*)}{n_i} \right\} \right| \\ &= \max_{i=1, \dots, m} \left\| \bar{\mathbf{u}}_i^\theta - \mathbf{u}_{*,i} \right\| \left\{ \max_{k=1, \dots, d} \frac{\sqrt{d} \left| \frac{1}{n_i} \sum_{j=1}^{n_i} \{y_{ij} \mathbf{v}_{jk} - \mathbb{E} y_{ij} \mathbf{v}_{jk}\} \right|}{n_i} + \frac{\| \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_*) \mathbf{u}_{*,i} \|}{n_i} \right\} \\ &\leq \delta \left\{ O_p \left(\frac{1}{G(n_{\min})} \right) + O_p \left(\frac{\sqrt{\log(m)}}{n_{\min}} \right) \right\}, \end{aligned}$$

by Lemmas 11 and 14 and the fact that the maximum of the norm of m Gaussian vectors is $O_p(\sqrt{\log(m)})$ by Theorem 1.14 in Rigollet and Hütter (2023). The third term satisfies

$$(\bar{\mathbf{u}}_i^\theta - \mathbf{u}_{*,i})^\top \frac{\partial_{\mathbf{u}} \partial_{\mathbf{u}}^\top \log \pi_i(\mathbf{y}_i \mid \mathbf{u}^\circ; \boldsymbol{\theta})}{n_i} (\bar{\mathbf{u}}_i^\theta - \mathbf{u}_{*,i}) \leq - \left\| \bar{\mathbf{u}}_i^\theta - \mathbf{u}_{*,i} \right\|_2^2 \underline{\eta} n_{\min}^{-\varepsilon}$$

by Assumption 3 for every $\varepsilon > 0$. We therefore conclude

$$\lim_{n_i \rightarrow \infty} \frac{1}{n_i} (\bar{\mathbf{u}}_i^\theta - \mathbf{u}_{*,i})^\top \partial_{\mathbf{u}} \partial_{\mathbf{u}}^\top \log \pi_i(\mathbf{y}_i \mid \mathbf{u}^\circ; \boldsymbol{\theta}) (\bar{\mathbf{u}}_i^\theta - \mathbf{u}_{*,i}) \leq -\delta^2 \underline{\eta} n_{\min}^{-\varepsilon},$$

by Lemma 14.

As for the second term: note that $n_i^{-1} \mathbf{H}_i^\theta(\bar{\mathbf{u}}_i^\theta) \geq C n_{\min}^\varepsilon$ for any $\varepsilon > 0$ and for some $C > 0$, and therefore any of its sub-matrix must have singular values growing at most at the same order. Since $\boldsymbol{\theta} \in \mathcal{B}_{\boldsymbol{\theta}_*}(\zeta_N)$

and $\zeta_N \downarrow 0$ at a polynomial rate, we have

$$\begin{aligned} & \frac{1}{n_i} |(\mathbf{u}_i - \mathbf{u}_{*,i})^\top \partial_{\mathbf{u}} \partial_{\boldsymbol{\theta}}^\top \log \pi_i(\mathbf{y}_i | \mathbf{u}_{*,i}; \boldsymbol{\theta}^\circ)(\boldsymbol{\theta} - \boldsymbol{\theta}_*)| \\ & \leq \delta \zeta_N \left\| \frac{1}{n_i} \partial_{\mathbf{u}} \partial_{\boldsymbol{\theta}}^\top \log \pi_i(\mathbf{y}_i | \mathbf{u}_{*,i}; \boldsymbol{\theta}^\circ) \right\|_{op} = O(n^{\varepsilon - \alpha}), \end{aligned}$$

for some $\alpha > 0$ by Assumption 6, therefore as ε is arbitrary, this terms decreases at polynomial rate. Assumption 4 follows by combining these three bounds.

Assumption 5

There exists a $\delta' < \delta$ such that $[\mathcal{B}_{(\boldsymbol{\theta}_*, \mathbf{u}_{*,i})}(\delta')]^c \subset [\mathcal{B}_{\boldsymbol{\theta}_*}(\delta)]^c \times [\mathcal{B}_{\mathbf{u}_{*,i}}(\delta)]^c$, where $\times$ is the direct product operator. We show the stronger assumption that there exists $b > 0$ such that for all $i \in [m]$,

$$\lim_{n_{\min} \rightarrow \infty} P_N^* \left[\forall_{i=1, \dots, m} \sup_{(\boldsymbol{\theta}, \mathbf{u}) \in [\mathcal{B}_{(\boldsymbol{\theta}_*, \mathbf{u}_{*,i})}(\delta')]^c} \log \pi_i(\mathbf{y}_i | \mathbf{u}; \boldsymbol{\theta}) - \log \pi_i(\mathbf{y}_i | \mathbf{u}_{*,i}; \boldsymbol{\theta}_*) \leq -n_i \cdot b \right] = 1.$$

By Assumption 6 and Assumption 9, we have that

$$\lim_{n_{\min} \rightarrow \infty} P_N^* \left\{ \forall_{i=1, \dots, m} (\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{u}}_i^\theta) \in \mathcal{B}_{(\boldsymbol{\theta}_*, \mathbf{u}_{*,i})}(\delta') \right\} = 1, \quad (\text{D.2})$$

where $\widehat{\boldsymbol{\theta}}$ is the marginal maximum likelihood estimator.

By arguments analogous to the proof of Lemma 14 (see the verification of Assumption 4), it can be shown that

$$\lim_{n_{\min} \rightarrow \infty} P_N^* \left(\forall_{i=1, \dots, m} \left\| (\bar{\boldsymbol{\theta}}, \bar{\mathbf{u}}_i^\theta) - (\boldsymbol{\theta}_*, \mathbf{u}_{*,i}) \right\| = \delta' \right) = 1, \quad (\text{D.3})$$

where

$$(\bar{\boldsymbol{\theta}}, \bar{\mathbf{u}}_i^\theta) = \arg \max_{(\boldsymbol{\theta}, \mathbf{u}) \in [\mathcal{B}_{(\boldsymbol{\theta}_*, \mathbf{u}_{*,i})}(\delta')]^c} \log \pi_i(\mathbf{y}_i | \mathbf{u}; \boldsymbol{\theta}).$$

It follows that:

$$\begin{aligned} & \sup_{(\boldsymbol{\theta}_*, \mathbf{u}_*) \in [\mathcal{B}_{(\boldsymbol{\theta}_*, \mathbf{u}_*)}(\delta')]^c} \frac{1}{n_i} \left\{ \log \pi_i(\mathbf{y}_i | \mathbf{u}; \boldsymbol{\theta}) - \log \pi_i(\mathbf{y}_i | \mathbf{u}_{*,i}; \boldsymbol{\theta}_*) \right\} \\ & = \frac{1}{n_i} \left\{ \log \pi_i(\mathbf{y}_i | \bar{\mathbf{u}}_i^\theta; \bar{\boldsymbol{\theta}}) - \log \pi_i(\mathbf{y}_i | \mathbf{u}_{*,i}; \boldsymbol{\theta}_*) \right\} \\ & = n_i^{-1} \left\{ \log \pi_i(\mathbf{y}_i | \bar{\mathbf{u}}_i^\theta; \bar{\boldsymbol{\theta}}) - \log \pi_i(\mathbf{y}_i | \widehat{\mathbf{u}}_i^\theta; \widehat{\boldsymbol{\theta}}) \right\} + n_i^{-1} \left\{ \log \pi_i(\mathbf{y}_i | \widehat{\mathbf{u}}_i^\theta; \widehat{\boldsymbol{\theta}}) - \log \pi_i(\mathbf{y}_i | \mathbf{u}_{*,i}; \boldsymbol{\theta}_*) \right\}. \end{aligned}$$

By Assumption 6,

$$n_i^{-1} \left\{ \log \pi_i(\mathbf{y}_i | \widehat{\mathbf{u}}_i^\theta; \widehat{\boldsymbol{\theta}}) - \log \pi_i(\mathbf{y}_i | \mathbf{u}_{*,i}; \boldsymbol{\theta}_*) \right\} = O(n_{\min}^\varepsilon),$$

almost surely at a uniform rate. By Eqs. (D.2) and (D.3), there exists a $\delta'' > 0$ such that

$$\lim_{n_i \rightarrow \infty} P_N^* \left(\forall_{i=1, \dots, m} \left\| (\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{u}}_i^\theta) - (\bar{\boldsymbol{\theta}}, \bar{\mathbf{u}}_i^\theta) \right\| > \delta'' \right) = 1.$$

By Assumption 3 and second-order Taylor expansion around $(\widehat{\mathbf{u}}_i^\theta, \widehat{\boldsymbol{\theta}})$ followed by a first order Taylor expansion on the score around $(\mathbf{u}_{*,i}, \boldsymbol{\theta}_*)$ we have:

$$\begin{aligned}
& n_i^{-1} \{ \log \pi_i(y_i | \bar{\mathbf{u}}_i^\theta; \bar{\boldsymbol{\theta}}) - \log \pi_i(y_i | \widehat{\mathbf{u}}_i^\theta; \widehat{\boldsymbol{\theta}}) \} \\
&= n_i^{-1} (\bar{\mathbf{u}}_i^\theta - \widehat{\mathbf{u}}_i^\theta, \bar{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}})^\top \{ \partial_{\boldsymbol{\theta}} \log \pi_i(y_i | \widehat{\mathbf{u}}_i^\theta; \widehat{\boldsymbol{\theta}}) \} \\
&+ n_i^{-1} (\bar{\mathbf{u}}_i^\theta - \widehat{\mathbf{u}}_i^\theta, \bar{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}})^\top \{ \partial_{\boldsymbol{\theta}}^2 \log \pi_i(y_i | \mathbf{u}^\star; \boldsymbol{\theta}^\star) \} (\bar{\mathbf{u}}_i^\theta - \widehat{\mathbf{u}}_i^\theta, \bar{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}) / 2 \\
&= n_i^{-1} (\bar{\mathbf{u}}_i^\theta - \widehat{\mathbf{u}}_i^\theta, \bar{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}})^\top \{ \partial_{\boldsymbol{\theta}} \log \pi_i(y_i | \mathbf{u}_{*,i}; \boldsymbol{\theta}_*) \} \\
&+ n_i^{-1} (\bar{\mathbf{u}}_i^\theta - \widehat{\mathbf{u}}_i^\theta, \bar{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}})^\top \{ \partial_{\boldsymbol{\theta}} \log \pi_i(y_i | \mathbf{u}'; \boldsymbol{\theta}') \} (\widehat{\mathbf{u}}_i^\theta - \mathbf{u}_{*,i}, \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*) \\
&+ n_i^{-1} (\bar{\mathbf{u}}_i^\theta - \widehat{\mathbf{u}}_i^\theta, \bar{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}})^\top \{ \partial_{\boldsymbol{\theta}}^2 \log \pi_i(y_i | \mathbf{u}^\star; \boldsymbol{\theta}^\star) \} (\bar{\mathbf{u}}_i^\theta - \widehat{\mathbf{u}}_i^\theta, \bar{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}) / 2
\end{aligned} \tag{D.4}$$

Where in the above, the score evaluated at the data generating value tends to 0 uniformly by Lemma 10 and

$$\begin{aligned}
\partial_{\boldsymbol{\theta}}^2 \log \pi_i(y_i | \mathbf{u}'; \boldsymbol{\theta}') &= \int_0^1 (1-t) \partial_{\boldsymbol{\theta}}^2 \log \pi_i(y_i; \widehat{\mathbf{u}}_i^\theta + t\mathbf{u}_*, \widehat{\boldsymbol{\theta}} + t\boldsymbol{\theta}_*) dt \\
\partial_{\boldsymbol{\theta}}^2 \log \pi_i(y_i | \mathbf{u}^\star; \boldsymbol{\theta}^\star) &= \int_0^1 (1-t) \partial_{\boldsymbol{\theta}}^2 \log \pi_i(y_i; \widehat{\mathbf{u}}_i^\theta + t\bar{\mathbf{u}}_i^\theta, \widehat{\boldsymbol{\theta}} + t\bar{\boldsymbol{\theta}}) dt.
\end{aligned}$$

by the same argument as in Eq. (C.6) the maximal eigenvalue of these matrices are lower bounded by $n_i^{1-\varepsilon} \underline{\eta}$ for any $\varepsilon > 0$. Noting that $(\widehat{\mathbf{u}}_i^\theta - \mathbf{u}_{*,i}, \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*) = o_p(1)$ uniformly in the index i by Assumptions 6 and 9, only the final term in Eq. (D.4) is non-negligible. Therefore:

$$Eq. (D.4) \leq o_p(1) + \left\| \bar{\mathbf{u}}_i^\theta - \widehat{\mathbf{u}}_i^\theta, \bar{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}} \right\|_2 \lambda_1 \{ n_i^{-1} \partial_{\boldsymbol{\theta}}^2 \log \pi_i(y_i | \mathbf{u}^\star; \boldsymbol{\theta}^\star) \} \leq -\delta^2 \underline{\eta} n_{\min}^{-\varepsilon},$$

with probability uniformly tending to 1. Assumption 5 holds by taking $b = \delta^2 \underline{\eta} / 2$ and noting that the upper bound does not depend on i .

Assumption 6

Fix $\boldsymbol{\theta} \in \mathcal{B}_{\boldsymbol{\theta}_*}(\zeta_N)$. For any $\boldsymbol{\theta}$ and $i \in [m]$, the conditional mode $\widehat{\mathbf{u}}_i^\theta$ satisfies

$$\partial_{\mathbf{u}} \log \pi(y_i; \widehat{\mathbf{u}}_i^\theta, \boldsymbol{\theta}) + \partial_{\mathbf{u}} g(\widehat{\mathbf{u}}_i^\theta; \boldsymbol{\theta}) = \{y_i - \mathbf{b}'_i(\boldsymbol{\theta}, \widehat{\mathbf{u}}_i^\theta)\} \mathbf{V}_i + \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}) \widehat{\mathbf{u}}_i^\theta = \mathbf{0},$$

where $\mathbf{b}'_i(\boldsymbol{\theta}, \mathbf{u}) = (b'(\eta_{i1}(\boldsymbol{\theta}, \mathbf{u})), \dots, b'(\eta_{in_i}(\boldsymbol{\theta}, \mathbf{u})))^\top$. A first-order Taylor expansion of the first term of this equation about $(\boldsymbol{\theta}_*, \mathbf{u}_{*,i})$ evaluated at $(\boldsymbol{\theta}, \widehat{\mathbf{u}}_i^\theta)$ gives

$$\begin{aligned}
& \{y_i - \mathbf{b}'_i(\boldsymbol{\theta}_*, \mathbf{u}_{*,i})\} \mathbf{V}_i - \mathbf{X}_i^\top \mathbf{W}_i^{\theta^\circ}(\mathbf{u}^\circ) \mathbf{V}_i (\boldsymbol{\beta} - \boldsymbol{\beta}_*) \\
& - \left\{ \mathbf{V}_i^\top \mathbf{W}_i^{\theta^\circ}(\mathbf{u}^\circ) \mathbf{V}_i + \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}) \right\} (\widehat{\mathbf{u}}_i^\theta - \mathbf{u}_{*,i}) - \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}) \mathbf{u}_{*,i} = \mathbf{0},
\end{aligned}$$

where $\mathbf{W}_i^\theta(\mathbf{u}) = \text{diag} \left\{ b''(\eta_{i1}^{\tilde{\theta}_1}(\tilde{\mathbf{u}}_1)), \dots, b''(\eta_{i n_i}^{\tilde{\theta}_{n_i}}(\tilde{\mathbf{u}}_{n_i})) \right\}$, $\tilde{\theta}_i = (1 - t_i)\theta + \theta_*$, $\tilde{\mathbf{u}}_j = (1 - t_j)\mathbf{u} + \mathbf{u}_*$ for $t_i, t_j \in [0, 1]$ and $\mathbf{u}^\circ = (\tilde{\mathbf{u}}_1, \dots, \tilde{\mathbf{u}}_{n_i})$, $\theta^\circ = (\tilde{\theta}_1, \dots, \tilde{\theta}_{n_i})$. Rearranging, we obtain

$$\begin{aligned} (\hat{\mathbf{u}}_i^\theta - \mathbf{u}_{*,i}) &= \left\{ \mathbf{V}_i^\top \mathbf{W}_i^{\theta^\circ}(\mathbf{u}^\circ) \mathbf{V}_i + \Sigma^{-1}(\theta) \right\}^{-1} \{ \mathbf{y}_i - \mathbf{b}'_i(\theta_*, \mathbf{u}_{*,i}) \} \mathbf{V}_i \\ &\quad - \left\{ \mathbf{V}_i^\top \mathbf{W}_i^{\theta^\circ}(\mathbf{u}^\circ) \mathbf{V}_i + \Sigma^{-1}(\theta) \right\}^{-1} \mathbf{X}_i^\top \mathbf{W}_i^\theta(\mathbf{u}^\circ) \mathbf{V}_i (\boldsymbol{\beta} - \boldsymbol{\beta}_*) \\ &\quad + \left\{ \mathbf{V}_i^\top \mathbf{W}_i^{\theta^\circ}(\mathbf{u}^\circ) \mathbf{V}_i + \Sigma^{-1}(\theta) \right\}^{-1} \Sigma^{-1}(\theta) \mathbf{u}_{*,i} \end{aligned}$$

and hence,

$$\begin{aligned} \zeta_N^{-1} \left\| \hat{\mathbf{u}}_i^\theta - \mathbf{u}_{*,i} \right\|_2 &\leq \left\| \left\{ \mathbf{V}_i^\top \mathbf{W}_i^{\theta^\circ}(\mathbf{u}^\circ) \mathbf{V}_i + \Sigma^{-1}(\theta) \right\}^{-1} \right\|_2 \\ &\quad \times \left\{ \zeta_N^{-1} \left\| \{ \mathbf{y}_i - \mathbf{b}'_i(\theta_*, \mathbf{u}_{*,i}) \} \mathbf{V}_i \right\|_2 + \zeta_N^{-1} \left\| \Sigma^{-1}(\theta) \mathbf{u}_{*,i} \right\|_2 + \zeta_N^{-1} \|\boldsymbol{\beta} - \boldsymbol{\beta}_*\|_2 \left\| \mathbf{X}_i^\top \mathbf{W}_i^\theta(\mathbf{u}^\circ) \mathbf{V}_i \right\|_2 \right\}, \end{aligned} \quad (\text{D.5})$$

where we recall that $\zeta_N = (\min_{i=1, \dots, m} n_i)^{-\alpha}$ with $0 < \alpha < 1/4$. Eq. (D.5) will be used to show both statements in the Assumption.

For the first statement, we show that the two terms converges in probability to zero and the second term is bounded in probability uniformly by any polynomial function in both the index i and for any $\theta \in \mathcal{B}_{\theta_*}(\zeta_N)$. Recall from the proof of Assumption 3 that we have

$$\frac{1}{n_i} \sup_{\theta \in \mathcal{B}_{\theta_*}(\delta)} \sup_{\mathbf{u} \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} \lambda_1(\mathbf{D}_i^\top \mathbf{W}_i^\theta(\mathbf{u}) \mathbf{D}_i) < \bar{\eta} n_{\min}^\varepsilon.$$

and

$$\frac{1}{n_i} \inf_{\theta \in \mathcal{B}_{\theta_*}(\delta)} \inf_{\mathbf{u} \in \mathcal{B}_{\mathbf{u}_{*,i}}(\delta)} \lambda_d(\mathbf{D}_i^\top \mathbf{W}_i^\theta(\mathbf{u}) \mathbf{D}_i) > \underline{\eta} n_{\min}^{-\varepsilon}.$$

for some any $\varepsilon > 0$, where $\mathbf{D}_i = [\mathbf{X}_i : \mathbf{V}_i]$. We therefore have

$$\sup_{\theta \in \mathcal{B}_{\theta_*}(\zeta_N)} \left\| \mathbf{X}_i^\top \mathbf{W}_i^\theta(\mathbf{u}^\circ) \mathbf{V}_i \right\|_2 < \bar{\eta} n_{\min}^\varepsilon$$

and

$$\sup_{\theta \in \mathcal{B}_{\theta_*}(\zeta_N)} \left\| \left\{ \mathbf{V}_i^\top \mathbf{W}_i^{\theta^\circ}(\mathbf{u}^\circ) \mathbf{V}_i + \Sigma^{-1}(\theta) \right\}^{-1} \right\|_2 < \left(\underline{\eta} n^{1-\varepsilon} + \underline{\xi}_1 \right)^{-1} < \underline{\eta}^{-1} n_{\min}^{-(1-\varepsilon)}.$$

By definition we have $\sup_{\theta \in \mathcal{B}_{\theta_*}(\gamma_m)} \|\boldsymbol{\beta} - \boldsymbol{\beta}_*\|_2 \leq \zeta_N$. Combining these, we have that the third term,

$$\begin{aligned} &\sup_{\theta \in \mathcal{B}_{\theta_*}(\zeta_N)} \zeta_N^{-1} \|\boldsymbol{\beta} - \boldsymbol{\beta}_*\|_2 \left\| \mathbf{X}_i^\top \mathbf{W}_i^\theta(\mathbf{u}^\circ) \mathbf{V}_i \right\|_2 \left\| \left\{ \mathbf{V}_i^\top \mathbf{W}_i^{\theta^\circ}(\mathbf{u}^\circ) \mathbf{V}_i + \Sigma^{-1}(\theta) \right\}^{-1} \right\|_2 \\ &\leq \zeta_N \zeta_N^{-1} \underline{\eta}^{-1} \bar{\eta} n_{\min}^{1+\varepsilon} n_{\min}^{-(1-\varepsilon)} C_2^{-1} = O(n_{\min}^\varepsilon), \end{aligned}$$

for any $\varepsilon > 0$. By Lemma 11,

$$\max_{i=1,\dots,m} \mathbb{P} \left(n_i^{-1} \|\{y_i - \mathbf{b}'_i(\boldsymbol{\theta}_*, \mathbf{u}_{*,i})\} V_i\|_2 \leq \frac{\sqrt{d} B_2}{G(n_{\min})} \right) \geq 1 - o(1).$$

We then have:

$$\begin{aligned} & \zeta_N^{-1} \|\{y_i - \mathbf{b}'_i(\boldsymbol{\theta}_*, \mathbf{u}_{*,i})\} V_i\|_2 \left\| \left\{ V_i^\top \mathbf{W}^{\boldsymbol{\theta}^\circ}(\mathbf{u}^\circ) V_i + \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}) \right\}^{-1} \right\|_2, \\ & \leq \left[n_i^{-1} \|\{y_i - \mathbf{b}'_i(\boldsymbol{\theta}_*, \mathbf{u}_{*,i})\} V_i\|_2 \right] \cdot \left[\zeta_N^{-1} n_i n_i^{-(1+\varepsilon)} C_2^{-1} \right], \\ & \leq \frac{\sqrt{d} n_{\min}^{\varepsilon-\alpha} B_2}{G(n_{\min})} \rightarrow 0, \end{aligned}$$

with probability tending to 1 uniformly in i . with $0 < \alpha < 1/4$ by construction. For the second term, by a similar argument:

$$\left\| \left\{ V_i^\top \mathbf{W}^{\boldsymbol{\theta}^\circ}(\mathbf{u}^\circ) V_i + \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}) \right\}^{-1} \right\|_2 \left\| \zeta_N^{-1} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}) \mathbf{u}_{*,i} \right\|_2 \leq n_i^{-1+\varepsilon} C_2^{-1} \zeta_N^{-1} \bar{\xi}_1 \|\mathbf{u}_{*,i}\|_2 \rightarrow 0, \quad (\text{D.6})$$

by Assumption 8 iii) and the fact that $\max_{i=1,\dots,m} \|\mathbf{u}_{*,i}\|_2 = O_P(\sqrt{\log(m)})$ by Theorem 1.14 in [Rigollet and Hütter \(2023\)](#). This proves the first statement in Assumption 6.

For the second statement, we evaluate the same Taylor expansion at $(\boldsymbol{\theta}_*, \widehat{\mathbf{u}}_i^{\boldsymbol{\theta}^*})$ to obtain

$$\frac{n_i^{1/2}}{G'(n_i)} \|\widehat{\mathbf{u}}_i^{\boldsymbol{\theta}^*} - \mathbf{u}_{*,i}\|_2 \leq \frac{n_i^{1/2}}{G'(n_i)} \|\{y_i - \mathbf{b}'_i(\boldsymbol{\theta}_*, \mathbf{u}_{*,i})\} V_i\|_2 \left\| \left\{ V_i^\top \mathbf{W}^{\boldsymbol{\theta}^*}(\mathbf{u}^\circ) V_i + \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}) \right\}^{-1} \right\|_2,$$

where now $\mathbf{u}^\circ = t \widehat{\mathbf{u}}_i^{\boldsymbol{\theta}^*} + (1-t) \mathbf{u}_{*,i}$ for $t \in [0, 1]$ and we recall that G' is any nonzero function with $\lim_{n \rightarrow \infty} G(n) = \infty$. We have

$$\begin{aligned} & \frac{n_i^{1/2}}{G'(n_i)} \|\{y_i - \mathbf{b}'_i(\boldsymbol{\theta}_*, \mathbf{u}_{*,i})\} V_i\|_2 \left\| \left\{ V_i^\top \mathbf{W}^{\boldsymbol{\theta}^*}(\mathbf{u}^\circ) V_i + \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}) \right\}^{-1} \right\|_2, \\ & \leq \frac{1}{G'(n_i)} \left[n_i^{-1/2} \|\{y_i - \mathbf{b}'_i(\boldsymbol{\theta}_*, \mathbf{u}_{*,i})\} V_i\|_2 \right] \left[n_i \left\| \left\{ V_i^\top \mathbf{W}^{\boldsymbol{\theta}^*}(\mathbf{u}^\circ) V_i + \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}) \right\}^{-1} \right\|_2 \right], \\ & \leq \frac{1}{G'(n_i)} \left[n_i^{-1/2} \|\{y_i - \mathbf{b}'_i(\boldsymbol{\theta}_*, \mathbf{u}_{*,i})\} V_i\|_2 \right] C_2^{-1} n_i^\varepsilon = O(n_i^\varepsilon), \end{aligned}$$

by Lemma 11. This proves the second statement in Assumption 6.

Assumption 7

The first statement of Assumption 7 follows directly from the fact that our prior is Gaussian with full support.

For the second and third statement for all $\theta \in \mathcal{B}_{\theta_*}(\zeta_N)$, in GLMMs the score function is:

$$\begin{aligned} & \int_{\mathbf{u} \in \mathbb{R}^d} \|\partial_{\theta} \log \pi(\mathbf{y}_i; \mathbf{u}, \theta)\|_2 \phi(\mathbf{u}; \mathbf{0}, \Sigma(\sigma)) d\mathbf{u} \\ & \leq n_i^{1/2} \|X_i\|_{op} \int_{\mathbf{u} \in \mathbb{R}^d} \left\{ \frac{\|\mathbf{y}_i - \mathbf{b}'_i(\theta_*, \mathbf{u}_{*,i})\|_2}{n_i^{1/2}} + \frac{\|\mathbf{b}'_i(\theta_*, \mathbf{u}_{*,i}) - \mathbf{b}'_i(\theta, \mathbf{u})\|_2}{n_i^{1/2}} \right\} \phi(\mathbf{u}; \mathbf{0}, \Sigma(\sigma)) d\mathbf{u} \\ & \leq \bar{\xi}_2 n_i^{1+\varepsilon} \left\{ o_p(n_{\min}^{\varepsilon}) + \int_{\mathbf{u} \in \mathbb{R}^d} \frac{\|\mathbf{b}'_i(\theta_*, \mathbf{u}_{*,i}) - \mathbf{b}'_i(\theta, \mathbf{u})\|_2}{n_i^{1/2}} \phi(\mathbf{u}; \mathbf{0}, \Sigma(\sigma)) d\mathbf{u} \right\}, \end{aligned}$$

uniformly in i by Lemma 11 and by Assumption 8 iv) $\|X_i\|_{op} \leq \bar{\xi}_2 n_i^{1/2+\varepsilon}$. The term within the integrand for every $\mathbf{u}$

$$\begin{aligned} \frac{\|\mathbf{b}'_i(\theta_*, \mathbf{u}_{*,i}) - \mathbf{b}'_i(\theta, \mathbf{u})\|_2}{n_i^{1/2}} &= \|(\theta_* - \theta, \mathbf{u}_{*,i} - \mathbf{u}) \circ \mathbf{b}''_i(\tilde{\theta}, \tilde{\mathbf{u}})\|_2 \\ &\leq \|(\theta_* - \theta, \mathbf{u}_{*,i} - \mathbf{u})\|_2 \|\mathbf{b}''_i(\tilde{\theta}, \tilde{\mathbf{u}})\|_2 \\ &\leq \{\zeta_N + \|\mathbf{u}_{*,i}\|_2 + \|\mathbf{u}\|_2\} \\ &\quad \times o\left(\exp\left(-\log(n_{\min})^{1-\varepsilon} \left\{\|\mathbf{u}\|_2^{2-\varepsilon} + 2\|\mathbf{u}_{*,i}\|_2^{2-\varepsilon}\right\}\right)\right) \leq C < \infty \end{aligned}$$

for some constant $C > 0$, where $\circ$ denotes the element wise product and $\tilde{\theta}, \tilde{\mathbf{u}}$ are vectors whose components satisfies $\tilde{\theta}_j = t_j \theta + (1 - t_j) \theta_*$ and $\tilde{\mathbf{u}}_j = t'_j \mathbf{u} + (1 - t'_j) \mathbf{u}_{*,i}$ for $t_j, t'_j \in [0, 1]$ and $j = 1, \dots, n_i$. The last inequality holds as the natural parameter is linear in $\mathbf{u}$ and the largest $\|\mathbf{V}_i\|_2$ is $O(\log(n_{\min})^{\frac{1-\varepsilon}{2}})$ and by Assumption 8 v). The expectation of this term is uniformly bounded as $\max_{i=1, \dots, m} \|\mathbf{u}_{*,i}\|_2 = O_p(\sqrt{\log(m)})$, and therefore this integrand is uniformly $o(\exp(\|\mathbf{u}\|_2^{2-\varepsilon}))$ and therefore integrable. This holds as the natural parameter is linear in $\mathbf{u}$ and by Assumption 8 v).

As for the final statement, for $\mathbf{W}_i^{\theta}(\mathbf{u})$ as defined in the proof of Assumption 6

$$\begin{aligned} & \int_{\mathbf{u} \in \mathbb{R}^d} \left\| \partial_{\theta}^2 \log \pi(\mathbf{y}_i; \mathbf{u}, \theta) \right\|_{op} \phi(\mathbf{u}; \mathbf{0}, \Sigma(\sigma)) d\mathbf{u} \\ & \leq \int_{\mathbf{u} \in \mathbb{R}^d} \|X_i^T \mathbf{W}_i^{\theta}(\mathbf{u}) X_i\|_{op} \phi(\mathbf{u}; \mathbf{0}, \Sigma(\sigma)) d\mathbf{u} \\ & \leq \int_{\mathbf{u} \in \mathbb{R}^d} \max\{\mathbf{W}_i^{\theta}(\mathbf{u})\} \|X_i^T X_i\|_{op} \phi(\mathbf{u}; \mathbf{0}, \Sigma(\sigma)) d\mathbf{u}, \end{aligned}$$

where $\max\{\mathbf{W}_i^{\theta}(\mathbf{u})\}$ is the maximum entry of the matrix and $\|X_i^T X_i\|_{op} \leq \bar{\xi}_2 n_i^{1+\varepsilon}$ for all $\varepsilon > 0$. The maximum variance of an observation can then be show to grow slower than $\exp(-\|\mathbf{u}\|_2^{2-\varepsilon})$ by Assumption 8 v) through similar steps as before, showing the desired result. $\square$

Appendix E: Additional simulation results

This section gives additional results for the simulation study of Subsection 4.3. We report bias, coverage, and root-mean square error (RMSE) for β_0 and β_1 .

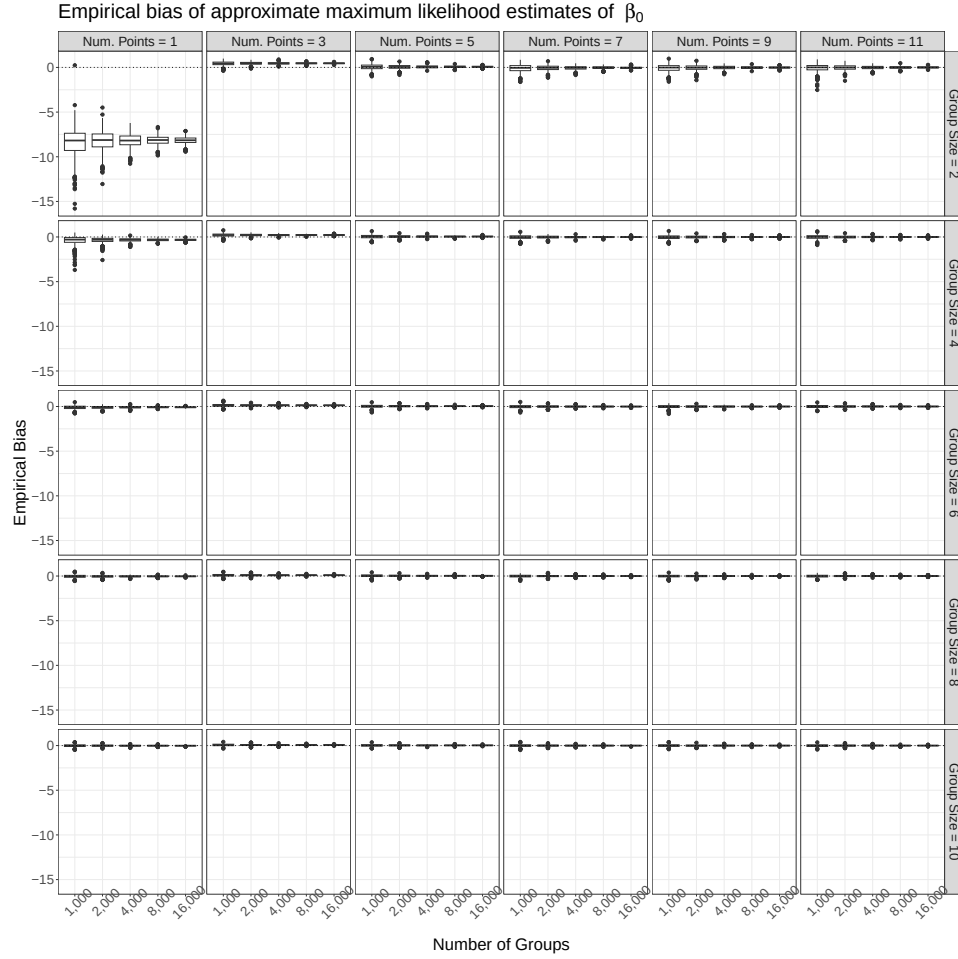


Figure 2. Empirical bias of β_0 for the simulation of Subsection 4.3. A Bernoulli random intercept model (Eq. (4.3)) was fit 1000 times to each combination of numbers of groups (m , x-axis), group size (n , rows) and numbers of quadrature points (k , columns). The y-axis shows the empirical bias of the approximate MLE, $\tilde{\theta}$ as an estimate of θ . Shown are empirical bias across 1000 simulations ($\bullet$) along with Monte Carlo confidence intervals ($--$).

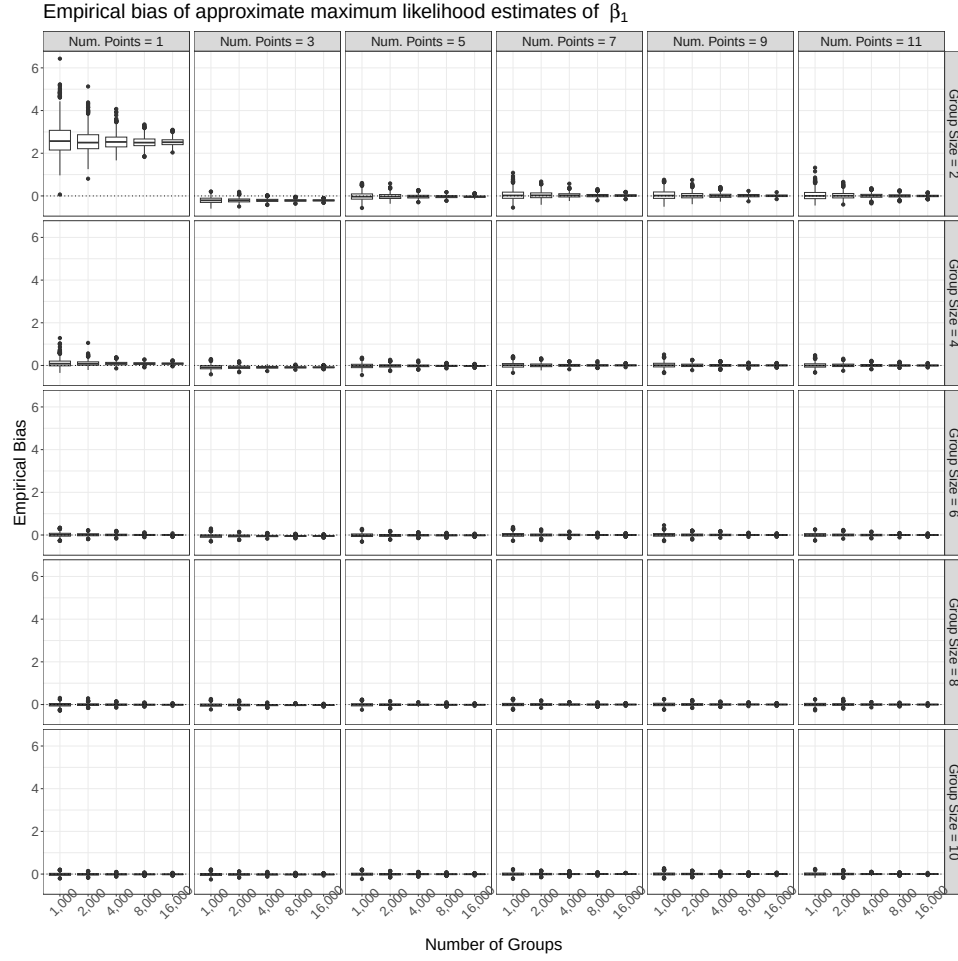


Figure 3. Empirical bias of β_1 for the simulation of Subsection 4.3. A Bernoulli random intercept model (Eq. (4.3)) was fit 1000 times to each combination of numbers of groups (m , x-axis), group size (n , rows) and numbers of quadrature points (k , columns). The y-axis shows the empirical bias of the approximate MLE, $\hat{\theta}$ as an estimate of θ . Shown are empirical bias across 1000 simulations ($\bullet$) along with Monte Carlo confidence intervals ($--$).

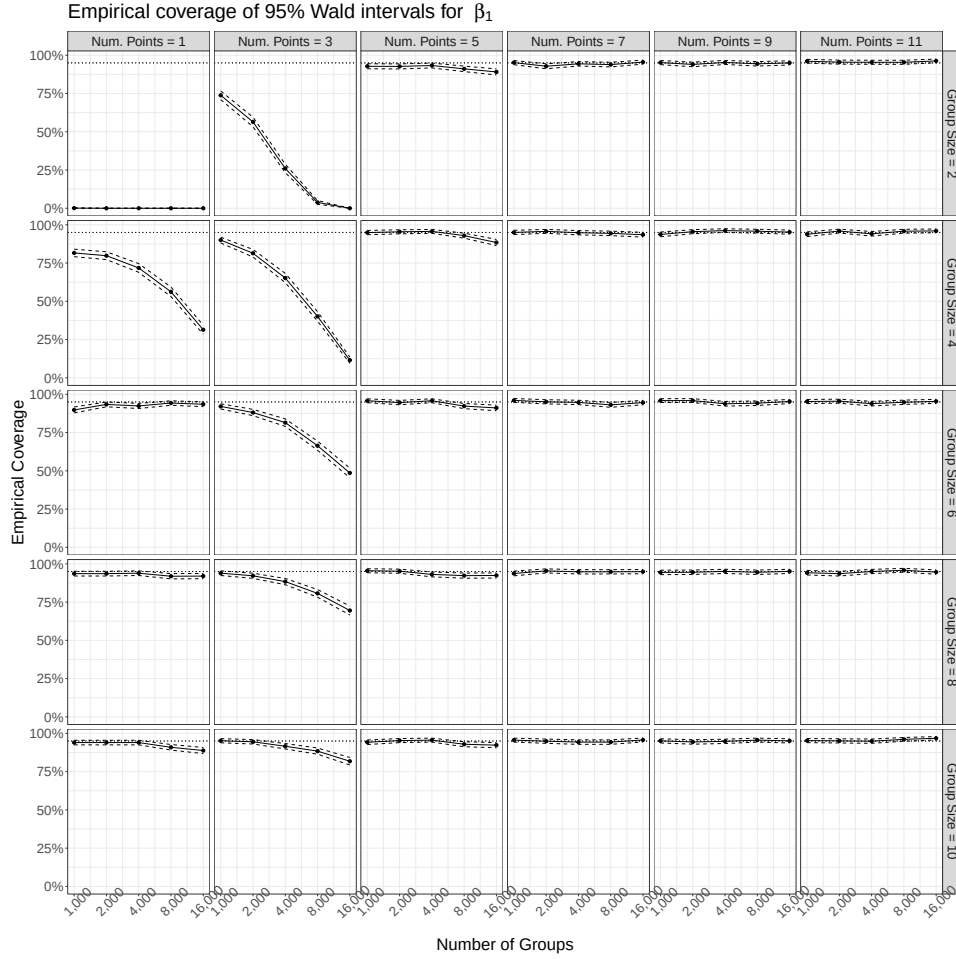


Figure 4. Empirical coverage of β_1 for the simulation of Subsection 4.3. A Bernoulli random intercept model (Eq. (4.3)) was fit 1000 times to each combination of numbers of groups (m , x-axis), group size (n , rows) and numbers of quadrature points (k , columns). The y-axis shows empirical coverages of 95% Wald confidence intervals centred at the approximate MLE, $\tilde{\theta}$, with standard errors computed using the diagonal of the inverse Hessian of the approximate marginal likelihood. Shown are empirical coverage proportions across 1000 simulations ($\bullet$) along with Monte Carlo confidence intervals ($--$).

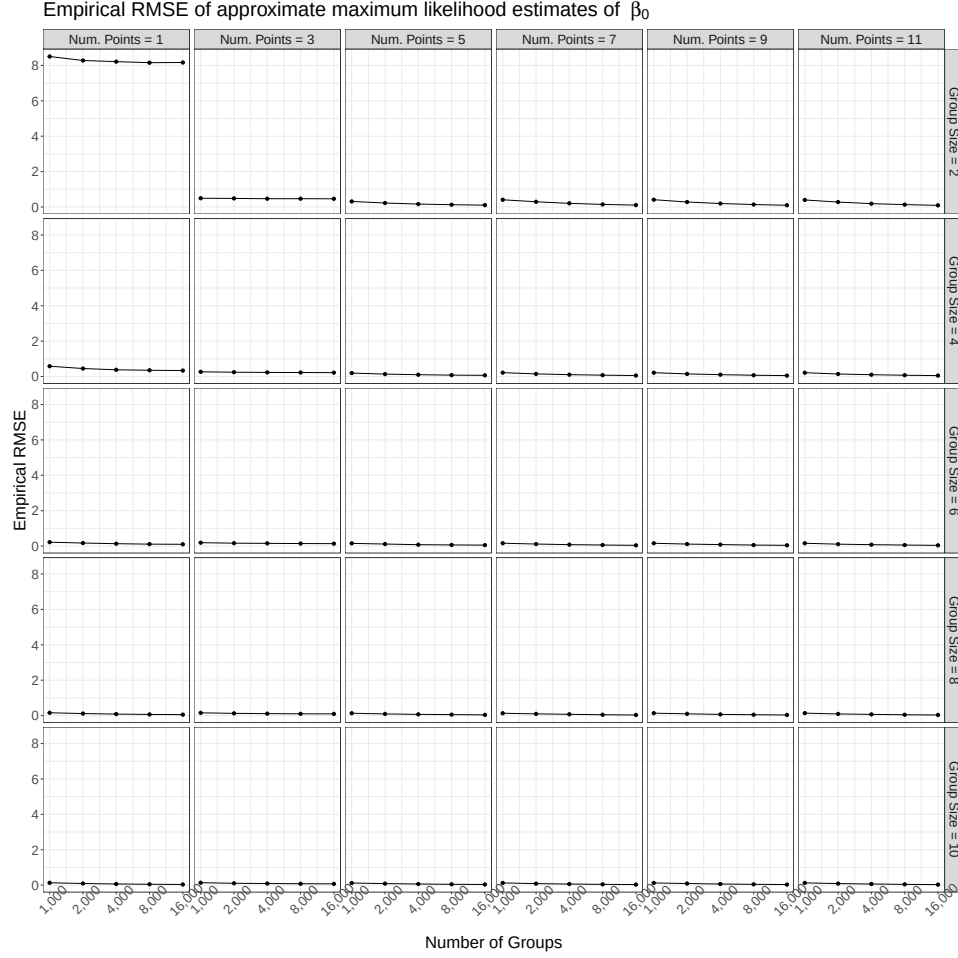


Figure 5. Empirical RMSE of β_0 for the simulation of Subsection 4.3. A Bernoulli random intercept model (Eq. (4.3)) was fit 1000 times to each combination of numbers of groups (m , x-axis), group size (n , rows) and numbers of quadrature points (k , columns). The y-axis shows the empirical RMSE of the approximate MLE, $\hat{\theta}$ as an estimate of θ . Shown are empirical RMSE across 1000 simulations ($\bullet$) along with Monte Carlo confidence intervals ($---$).

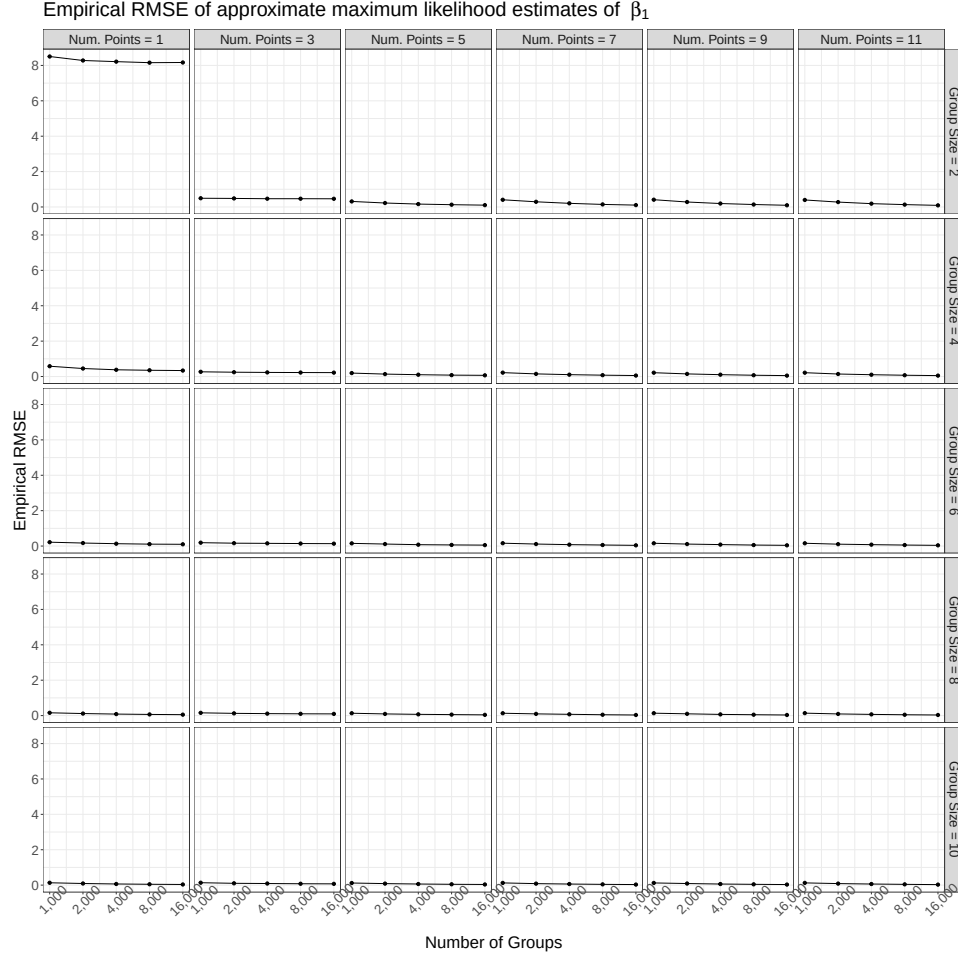


Figure 6. Empirical RMSE of β_1 for the simulation of Subsection 4.3. A Bernoulli random intercept model (Eq. (4.3)) was fit 1000 times to each combination of numbers of groups (m , x-axis), group size (n , rows) and numbers of quadrature points (k , columns). The y-axis shows the empirical RMSE of the approximate MLE, $\hat{\theta}$ as an estimate of θ . Shown are empirical RMSE across 1000 simulations ($\bullet$) along with Monte Carlo confidence intervals ($---$).