Fitting Generalized Linear Mixed Models using Adaptive Quadrature

Alex Stringer[1] and Blair Bilodeau[2]

Abstract

We describe how to approximate the intractable marginal likelihood that arises when fitting generalized linear mixed models. We prove that non-adaptive quadrature approximations yield high error asymptotically in every statistical model satisfying weak regularity conditions. We derive the rate of error incurred when using adaptive quadrature to approximate the marginal likelihood in a broad class of generalized linear mixed models, which includes non-exponential family response and non-Gaussian random effects distributions. We provide an explicit recommendation for how many quadrature points to use, and show that this recommendation recovers and explains many empirical results from published simulation studies and data analyses. Particular attention is paid to models for dependent binary and survival/time-to-event observations. Code to reproduce results in the manuscript is found at `https://github.com/awstringer1/glmm-aq-paper-code`.

1 Introduction

Generalized linear mixed models (GLMMs; Breslow and Clayton, 1993; Lee and Nelder, 1996; McCulloch, 1997) are used for modelling grouped data, including repeated measurements on subjects. A GLMM is a hierarchical model involving one or more latent group-level sources of stochastic heterogeneity, called random effects. Since the random effects are unobserved, inferences are based on a marginal likelihood, which is obtained from the marginal density of the observed response. Computing the marginal likelihood requires integrating the known joint likelihood with respect to the random effects. Except in the limited cases where the model's response and random effects distributions are deliberately chosen to make this integral tractable (Lee and Nelder, 1996), this integral must be approximated in order to evaluate the marginal likelihood, and hence make inferences about parameters of interest.

Gaussian quadrature (GQ) is a classical method for approximating the marginal likelihood in GLMMs (Lesaffre and Spiessens, 2001; Hedeker and Gibbons, 1994; Pan and Thompson, 2003), and has historically been the default method in many popular commercial and open-source software packages (Rabe-Hesketh et al., 2002; Wolfinger, 1999; Lillard and Panis, 2000; SERC, 1989). While it can no longer be described as popular, GQ continues to be used within the modern literature, both as a component of novel statistical methods (Jiang et al., 2021; Wu and Jones, 2021; Liu and Huang, 2008; Crowther et al., 2014) and in applications, for example in psychology (Bono et al., 2021). The GQ approximation to the marginal likelihood is straightforward to implement and fast

*Equal contribution authors
[1]Department of Statistics and Actuarial Science, University of Waterloo
[2]Department of Statistical Sciences, University of Toronto
Correspondence: alex.stringer[at]uwaterloo.ca

to evaluate. However, it has been observed that GQ can be inaccurate in practice (Pinheiro and Bates, 1995; Rabe-Hesketh et al., 2002), including in the analysis of binary (Lesaffre and Spiessens, 2001) and survival (Crowther et al., 2014) data. Whether to use GQ for fitting a GLMM remains an ambiguous decision for practitioners, and technical ambiguities of this nature have been noted (Bolker et al., 2008; Bono et al., 2021) to discourage the use of GLMMs in analyses where they are otherwise the most appropriate choice.

More recently, adaptive Gauss–Hermite quadrature (AQ; Naylor and Smith, 1982) has become a standard method for approximating the marginal likelihood in a limited class of GLMMs (Pinheiro and Bates, 1995; Kabaila and Ranathunga, 2019). In contrast to GQ, AQ has been shown to have compelling statistical properties in theory and practice. The Laplace approximation, which is AQ with only a single quadrature point, is well-understood theoretically (Tierney and Kadane, 1986; Kass et al., 1990), is the default estimation method for all models in the popular `lme4` software package (Bates et al., 2015), and is closely related to the frequently used penalized quasi-likelihood method (Breslow and Clayton, 1993). Beyond a single quadrature point, AQ has historically only been studied theoretically for a deterministic class of problems that does not include statistical inference (Liu and Pierce, 1994; Jin and Andersson, 2020), and only recently has been proven to perform well for stochastic problems (Bilodeau et al., 2021). While inferential properties for some specific GLMMs and inference methods have been considered under strong assumptions on the data (Vonesh, 1996; Bianconcini, 2014; Ogden, 2021), a fully stochastic analysis for GLMMs with general response and random effects distributions fit using AQ is not yet available.

Beyond theoretical considerations, the task of fitting a GLMM with AQ requires choosing the number of quadrature points to use, which includes deciding whether or not to use the Laplace approximation. This may be regarded as another potentially opaque, technical choice to be made by a practitioner, and hence a barrier to using GLMMs in practical applications. Popular software packages make conflicting, ad-hoc recommendations in their documentation with limited explanation (Bates et al., 2015; Rizopoulos, 2020; Rabe-Hesketh et al., 2002; Wolfinger, 1999; Lillard and Panis, 2000; SERC, 1989), adding to the potential confusion. A rigorously justified recommendation of how many quadrature points to use with AQ is of immediate practical benefit to those interested in fitting GLMMs to their data.

The primary contributions of this work are two-fold. First, we study the statistical error of approximating the integral of the likelihood in a GLMM using quadrature. To motivate our study of AQ over GQ, we first prove that any non-adaptive quadrature rule fails in the asymptotic data limit (Theorem 1), and therefore conclude that some adaptation is required. Then, we provide the asymptotic error rate of the AQ approximation to the marginal likelihood in a very broad class of GLMMs, including those with non-exponential family response and non-Gaussian random effects distributions (Theorem 2). This covers, for example, parametric survival models with Gamma frailties (Section 5; Crowther et al., 2014; Liu and Huang, 2008), which are not covered by existing theory. Our results help to explain earlier observations that AQ provides empirical improvements over GQ (Pinheiro and Bates, 1995; Lesaffre and Spiessens, 2001).

Our second contribution is an explicit recommendation of the number of quadrature points to use when fitting a GLMM using AQ. Our recommendation requires only the number of groups and number of observations per group, and is easy to calculate in practice. We discuss the theoretical justifications for our recommendation in Subsection 3.3, show that it recovers many existing empirically–motivated, ad-hoc recommendations in Section 4, and demonstrate its suitability for practical use through representative examples of fitting GLMMs to binary and survival data in Section 5. Our recommendation is:

When fitting a GLMM with M groups, where the smallest group has m observations, the number of adaptive quadrature points used should be at least

$$k(M, m) := \lceil (3/2) \log_m(M) - 2 \rceil.$$

In particular, one should avoid the Laplace approximation when $k(M, m) > 1$.

2 Preliminaries

2.1 Generalized Linear Mixed Models

A GLMM is parametrized by a total number of observations $n \in \mathbb{N}$, number of groups $M \in [n] = (1, \ldots, n)$, number of observations per group $(m_i)_{i \in [M]}$ satisfying $n = \sum_{i=1}^{M} m_i$, group-level random effect distributions $(G_i)_{i \in [M]}$ on $\mathbb{R}^p$, conditional response distributions $(F_i)_{i \in [M]}$ that are measurable maps from $\mathbb{R}$ to distributions on $\mathbb{R}$, fixed-effect linear coefficients $\boldsymbol{\beta} \in \mathbb{R}^d$, and a link function $h : \mathbb{R} \to \mathbb{R}$.

A GLMM for observed data $(\boldsymbol{x}_{ij}, \boldsymbol{v}_{ij}, y_{ij})_{i \in [M], j \in [m_i]}$, where $\boldsymbol{x}_{ij} \in \mathbb{R}^d$ and $\boldsymbol{v}_{ij} \in \mathbb{R}^p$, is of the form

$$\begin{aligned} y_{ij} \mid \mu_{ij} &\overset{\text{ind.}}{\sim} F_i(\mu_{ij}), \\ h(\mu_{ij}) &= \boldsymbol{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta} + \boldsymbol{v}_{ij}^{\mathrm{T}} \boldsymbol{U}_i, \\ \boldsymbol{U}_i &\overset{\text{ind.}}{\sim} G_i. \end{aligned} \tag{1}$$

The response and random effects distributions may depend on additional shared parameters $\boldsymbol{\sigma} \in \mathbb{R}^s$, usually representing the covariance of the random effects and any additional dispersion parameters in the response distribution. The parameters of inferential interest are $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\sigma}) \in \Theta = \mathbb{R}^d \times \mathbb{R}^s$.

We require that each G_i is zero-mean and has finite, positive definite covariance, but do not require them to be Gaussian. For theoretical guarantees, we require that F_i and G_i have densities f_i and g_i respectively that satisfy the regularity conditions of Appendix A, which relate to those given by Bilodeau et al. (2021), and are much weaker conditions than requiring the distributions belong to the exponential family or be Gaussian. We do not model the covariates $\boldsymbol{x}_{ij}$ and $\boldsymbol{v}_{ij}$, which may be fixed or random, and are implicitly conditioned upon for the remainder of the paper.

For $i \in [M]$, let $\boldsymbol{y}_i = (y_{ij})_{j \in [m_i]}$ denote the group's observations and denote the joint density of the group's responses and random effects by $\pi_i(\boldsymbol{y}_i, \boldsymbol{U}_i; \boldsymbol{\theta}) = \prod_{j=1}^{m_i} f_i(y_{ij}; \boldsymbol{\theta}, \boldsymbol{U}_i) g_i(\boldsymbol{U}_i; \boldsymbol{\sigma})$. Inferences for the unknown parameters $\boldsymbol{\theta}$ are based on the marginal likelihood

$$\pi(\boldsymbol{y}; \boldsymbol{\theta}) = \prod_{i=1}^{M} \pi_i(\boldsymbol{y}_i; \boldsymbol{\theta}) = \prod_{i=1}^{M} \int \pi_i(\boldsymbol{y}_i, \boldsymbol{U}; \boldsymbol{\theta}) \mathrm{d}\boldsymbol{U}, \tag{2}$$

The independence of $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_M$ ensures that the joint distribution of data and random effects factors into a product of p-dimensional integrals, where $p = \dim(\boldsymbol{U}_i)$ is typically small, and this is fundamental to the theory and computation in this setting. Inferences are based on an approximation to $\pi(\boldsymbol{y}; \boldsymbol{\theta})$ obtained by approximating each integral $\pi_i(\boldsymbol{y}_i; \boldsymbol{\theta})$ and then taking the product of the approximations.

2.2 Gauss–Hermite Quadrature

Let $\mathcal{Q}(1, k)$ be the set of points from k-point GQ in one dimension. Let $\mathcal{Q}(p, k)$ be an extension of $\mathcal{Q}(1, k)$ to p dimensions, for example one based on a product (Jin and Andersson, 2020),

sparse (Heiss and Winschel, 2008), or nested (Petras, 2003) rule, and let $\omega_k : \mathcal{Q}(p,k) \to \mathbb{R}$ be the corresponding weight function. The GQ approximation to $\pi_i(\boldsymbol{y}_i;\boldsymbol{\theta})$ is:

$$\widetilde{\pi}_i^{\texttt{GQ}}(\boldsymbol{y}_i;\boldsymbol{\theta}) = \sum_{\boldsymbol{z}\in\mathcal{Q}(p,k)} \pi_i\left(\boldsymbol{y}_i,\boldsymbol{z};\boldsymbol{\theta}\right)\omega_k(\boldsymbol{z}). \tag{3}$$

The $\boldsymbol{z}$ are chosen a-priori and widely tabulated. Computing $\widetilde{\pi}_i^{\texttt{GQ}}(\boldsymbol{y}_i;\boldsymbol{\theta})$ requires only $|\mathcal{Q}(p,k)|$ function evaluations.

However, $\widetilde{\pi}_i^{\texttt{GQ}}(\boldsymbol{y}_i;\boldsymbol{\theta})$ will be a poor approximation to $\pi_i(\boldsymbol{y}_i;\boldsymbol{\theta})$, and hence Eq. (2), whenever most of the mass of $\pi_i\left(\boldsymbol{y}_i,\cdot;\boldsymbol{\theta}\right)$ lies away from $\boldsymbol{z}\in\mathcal{Q}(p,k_i)$. We argue in Theorem 1 that this situation happens with high probability as $m_i\to\infty$, and hence $\widetilde{\pi}_i^{\texttt{GQ}}(\boldsymbol{y}_i;\boldsymbol{\theta})$ must somehow be adapted to the data. In finite samples, GQ can perform well when it happens, by coincidence, to be similar to the AQ approximation.

2.3 Adaptive Gauss–Hermite Quadrature

Let $\ell_i(\boldsymbol{y}_i,\boldsymbol{U}_i;\boldsymbol{\theta}) = \log\pi_i(\boldsymbol{y}_i,\boldsymbol{U}_i;\boldsymbol{\theta})$, $\widehat{\boldsymbol{U}}_i^{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{U}}\ell_i(\boldsymbol{y}_i,\boldsymbol{U};\boldsymbol{\theta})$, and $\boldsymbol{H}_i^{\boldsymbol{\theta}} = -\partial_{\boldsymbol{U}}^2\ell_i(\boldsymbol{y}_i,\boldsymbol{U}_i;\boldsymbol{\theta})$. Define $\boldsymbol{L}_i^{\boldsymbol{\theta}}$ to be the lower Cholesky triangle satisfying $(\boldsymbol{H}_i^{\boldsymbol{\theta}})^{-1} = \boldsymbol{L}_i^{\boldsymbol{\theta}}(\boldsymbol{L}_i^{\boldsymbol{\theta}})^{\mathrm{T}}$. The AQ approximation to $\pi_i(\boldsymbol{y}_i;\boldsymbol{\theta})$ is:

$$\widetilde{\pi}_i^{\texttt{AQ}}(\boldsymbol{y}_i;\boldsymbol{\theta}) = |\boldsymbol{L}_i^{\boldsymbol{\theta}}| \sum_{\boldsymbol{z}\in\mathcal{Q}(p,k)} \pi_i\left(\boldsymbol{y}_i, \boldsymbol{L}_i^{\boldsymbol{\theta}}\boldsymbol{z} + \widehat{\boldsymbol{U}}_i^{\boldsymbol{\theta}};\boldsymbol{\theta}\right)\omega_k(\boldsymbol{z}). \tag{4}$$

The corresponding likelihood approximation is $\widetilde{\pi}^{\texttt{AQ}}(\boldsymbol{y};\boldsymbol{\theta}) = \prod_{i=1}^M \widetilde{\pi}_i^{\texttt{AQ}}(\boldsymbol{y}_i;\boldsymbol{\theta})$. When $k=1$, $\widetilde{\pi}^{\texttt{AQ}}(\boldsymbol{y};\boldsymbol{\theta})$ is called a Laplace approximation. Because Eq. (4) centres and scales the points $\boldsymbol{z}\in\mathcal{Q}(p,k)$ according to the mode $\widehat{\boldsymbol{U}}_i^{\boldsymbol{\theta}}$ and curvature $\boldsymbol{H}_i^{\boldsymbol{\theta}}$ of $\ell_i(\boldsymbol{y}_i,\boldsymbol{U}_i;\boldsymbol{\theta})$, it will generally provide a very good approximation to $\pi_i(\boldsymbol{y}_i;\boldsymbol{\theta})$. We quantify this claim in Theorem 2.

3 Theoretical Guarantees

3.1 Error Incurred by Non-Adaptive Quadrature

We first show that non-adaptive, quadrature-based approximations (including GQ) to the marginal likelihood provably fail under weak regularity conditions that ensure concentration of the likelihood, and hence fail for models in which inference is facilitated by the usual first-order asymptotic theory.

Theorem 1. *Let $\mathcal{P} = \{\pi(\cdot \mid \boldsymbol{\xi}) : \boldsymbol{\xi}\in\mathfrak{M}\}$ be a statistical model on $\mathbb{R}^d$ with parameter space $\mathfrak{M}\subseteq\mathbb{R}^p$ and prior π on $\mathfrak{M}$ satisfying the regularity conditions of Kleijn and van der Vaart (2012, Theorem 2.1) with respect to the true distribution P^* of the observed data $\boldsymbol{y}$. For any fixed collection of points $\boldsymbol{z}_1,\ldots,\boldsymbol{z}_N\in\mathbb{R}^p$ and weights $\omega_1,\ldots,\omega_N\in\mathbb{R}_+$,*

$$\lim_{n\to\infty} P_n^*\left[\left|\frac{\sum_{i=1}^N \pi(\boldsymbol{y}\mid\boldsymbol{z}_i)\pi(\boldsymbol{z}_i)\,\omega_i}{\int\pi(\boldsymbol{y}\mid\boldsymbol{\xi})\pi(\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi}} - 1\right| < 1/2\right] = 0.$$

Remark 1. *For the random effects model, Theorem 1 can be applied for each fixed $\boldsymbol{\theta}$ to the model $\{p(\cdot\mid\boldsymbol{U},\boldsymbol{\theta}) : \boldsymbol{U}\in\mathbb{R}^p\}$ with prior distribution equal to the random effects distribution G. The necessary regularity assumptions are implied by joint regularity assumptions on $(\boldsymbol{\theta},\boldsymbol{U})$.* ◁

Remark 2. *The regularity conditions of Theorem 1 are implied by more classical ones, for example, those from Theorem 10.1 of van der Vaart (1998), but these require the model to be well-specified, which invalidates their use for the random effects model with arbitrary $\boldsymbol{\theta}$.* ◁

Remark 3. *A more precise analysis could be used to show that the error actually uniformly fails to converge to zero in a ball around the best model, but the simpler statement and proof of Theorem 1 suffices to capture the intuition behind the failure of non-adaptive methods.* ◁

Theorem 1 holds for any finite number of quadrature points, and hence the non-convergence of non-adapted quadrature cannot be mitigated by simply adding more points to an existing rule, a conclusion that agrees with the empirical evidence provided by Lesaffre and Spiessens (2001). This non-convergence can be mitigated by adapting the rule to the changing shape and location of the marginal likelihood.

3.2 Approximation Error for Adaptive Gauss–Hermite Quadrature in GLMMs

We now characterize the error incurred by approximating the likelihood in GLMMs using AQ.

Theorem 2. *Fix $M, k \in \mathbb{N}$, let $\widetilde{\pi}^{\texttt{AQ}}(\boldsymbol{y};\boldsymbol{\theta})$ be the AQ approximation of Eq. (4) with k quadrature points, and define δ_Θ as in Appendix A. Then:*

(a) *Under the assumptions of Appendix A of Bilodeau et al. (2021), for each $\boldsymbol{\theta} \in \mathcal{B}_{\delta_\Theta}(\boldsymbol{\theta}^*)$ there exists $C_{\boldsymbol{\theta}} > 0$ such that*

$$\lim_{m_1,\ldots,m_M\to\infty} P_n^*\left(|\log\widetilde{\pi}^{\texttt{AQ}}(\boldsymbol{y};\boldsymbol{\theta}) - \log\pi(\boldsymbol{y};\boldsymbol{\theta})| < C_{\boldsymbol{\theta}} \sum_{i=1}^{M} m_i^{-\lfloor (k+2)/3 \rfloor}\right) = 1.$$

(b) *Under the uniformity assumptions of Appendix A, there exists $C > 0$ such that*

$$\lim_{m_1,\ldots,m_M\to\infty} P_n^*\left(\sup_{\boldsymbol{\theta}\in\mathcal{B}_{\delta_\Theta}(\boldsymbol{\theta}^*)} |\log\widetilde{\pi}^{\texttt{AQ}}(\boldsymbol{y};\boldsymbol{\theta}) - \log\pi(\boldsymbol{y};\boldsymbol{\theta})| < C \sum_{i=1}^{M} m_i^{-\lfloor (k+2)/3 \rfloor}\right) = 1.$$

Theorem 2 generalizes Section 3.1 of Bianconcini (2014) to the much broader class of models in Eq. (1), and may be contrasted with Theorem 1 of Ogden (2021), who considered the error in higher-order Laplace approximations under alternative assumptions.

3.3 Selecting $k(M, m)$

Theorem 2 holds for every fixed $M > 0$, and the worst-case upper bound it provides on the error in the approximation to the log-marginal likelihood becomes looser for larger M. However, Nie (2007) showed that $M \to \infty$ is a necessary condition for consistency of the maximum likelihood estimator in GLMMs when the integration can be computed exactly. In practice, the analyst has a fixed M, but is able to choose k. We therefore recommend choosing k such that the error in the approximate marginal likelihood is asymptotically less than that in the maximum likelihood estimator. Specifically, we choose k to control the error in the worst of the M integral approximations that make up the approximate marginal likelihood.

Define $m = \min_{i\in M} m_i$ and $\varepsilon^*(k) = m^{-r(k)}$ for $r(k) = \lfloor (k+2)/3 \rfloor$. A corollary of Theorem 2(b) is

$$\sup_{\boldsymbol{\theta}\in\mathcal{B}_{\delta_\Theta}(\boldsymbol{\theta}^*)} \max_{i\in M} |\log\widetilde{\pi}_i^{\texttt{AQ}}(\boldsymbol{y}_i;\boldsymbol{\theta}) - \log\pi_i(\boldsymbol{y}_i;\boldsymbol{\theta})| = O_p\left(\varepsilon^*(k)\right).$$

For a fixed M and m, we recommend choosing $k = k(M, m)$ such that $M^{-1/2} \geq \varepsilon^*(k)$, using that $M^{-1/2}$ is the statistical estimation error in the absence of approximation error (Nie, 2007). This

yields our suggested minimum number of quadrature points,

$$k(M, m) = \lceil \frac{3}{2} \log_m(M) - 2 \rceil.$$

A Laplace approximation is obtained when $k = 1$. Accordingly, we recommend avoiding the Laplace approximation when $k(M, m) > 1$. This condition is a function of both M and m, and can be easily checked by a practitioner when faced with a new dataset to which they wish to fit a GLMM.

4 Comparison of $k(M, m)$ to Previous Recommendations

Grouped binary data is a standard benchmark scenario to assess the accuracy of approximation methods for GLMMs (Breslow and Clayton, 1993; Liu and Pierce, 1994; Lesaffre and Spiessens, 2001; Nie, 2007; Joe, 2008; Bianconcini, 2014; Jin and Andersson, 2020). In this section we compare a variety of empirically–motivated, ad-hoc recommendations made by previous authors to our own recommendation, $k(M, m)$, in models for grouped binary data. We revist such comparisons in Subsection 5.2 for survival data with non-Gaussian random effects.

For sources that considered AQ, we compare $k(M, m)$ to the number of quadrature points they recommended, and for sources that considered only the Laplace approximation (that is, $k = 1$), we compare whether or not $k(M, m) > 1$ with conclusions about whether or not the Laplace approximation was sufficiently accurate.

Following the earlier simulations of Zeger and Karim (1991), Breslow and Clayton (1993) simulate $M = 100$ clusters of size $m = 7$ from a binomial model, which they fit using penalized quasi-likelihood, a method based on the Laplace approximation. Their simulations offer empirical evidence that in the binary case, estimation of both the regression and variance parameters appears biased. We find $k(100, 7) = 2 > 1$, indicating that the Laplace approximation is not recommended, in agreement with Breslow and Clayton (1993). In a similar context, Nie (2007) uses $M = 100$ and $m \in \{7, 14, 28, 56\}$, but they consider only the error in the maximum likelihood estimator, essentially assuming no integration error. The number of points used is not reported by Nie (2007), however, $k(100, 7) = 2$ and $k(100, 14) = 1$, so at these cluster sizes, their practice of ignoring the integration error broadly matches our conclusion that the estimation error is dominated by that in the maximum likelihood estimator. Another example is Joe (2008), who simulates from a binary model fit using a Laplace approximation and AQ. They conclude that $k = 5$ is "essentially asymptotically unbiased" and that $k = 3$ gives "intermediate performance". They use $m \in \{2, 3, 5, 7, 9\}$, and, unfortunately, do not report M. Under our recommendation for k, $M \leq \{25, 82, 626, 2402, 6562\}$ should be respectively satisfied in order to ensure that the integral approximation error is smaller than the sampling error in the maximum likelihood estimator with these choices for m and $k = 5$.

Bianconcini (2014) simulates from a binary model with $M = 200$ and $m \in \{3, 5\}$. They emphasize the importance of selecting k, choosing $k = 5$ and 3 for $m = 3$ and 5 respectively, as was previously recommended in their specific context by Schilling and Bock (2005). In this case, $k(200, 3) = 6$ and $k(200, 5) = 3$, and hence $k(M, m)$ recovers these previous ad-hoc recommendations. Further, Bianconcini (2014) observes that when $m = 3$, fitting the model to data with $M = 1000$ results in less error than when $M = 200$ if $k = 5$, but when $M = 200$, there is essentially no change in error between using $k = 9$ and $k = 15$. We have $k(200, 3) = 6$ and $k(1000, 3) = 8$, explaining this pattern; when $M = 200$, we expect no further decrease in error from using $k > 6$, as the error is dominated by the sampling error in the maximum likelihood estimator rather than the integral approximation. In support of this point, Jin and Andersson (2020) simulate from a similar model to Bianconcini (2014), choosing $M = 10,000$ and $m = 10$. They conclude that AQ with $k = 3$

Table 1: Selected point estimates and standard errors obtained from fitting the binary logistic model to the `toenail` data of Lesaffre and Spiessens (2001) using AQ with various choices of k. The choice of $k(M, m) = 11$ appears to match the point at which the estimates stop changing with increasing k, at a computational time similar to lower choice of k and faster than higher choices. Numbers in brackets are estimated standard deviations. Timings are mean (SD) in seconds based on 100 runs.

k	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\sigma}^2$	Time
1	-2.510 (0.764)	-0.400 (0.047)	20.762	1.147 (0.157)
3	-2.026 (0.582)	-0.405 (0.047)	20.155	1.449 (0.088)
5	-1.458 (0.395)	-0.382 (0.043)	13.628	1.121 (0.085)
7	-1.500 (0.400)	-0.384 (0.043)	14.220	1.159 (0.091)
9	-1.575 (0.417)	-0.388 (0.044)	15.260	1.812 (0.094)
11	-1.629 (0.433)	-0.391 (0.044)	16.069	1.422 (0.088)
15	-1.647 (0.445)	-0.392 (0.045)	16.457	1.735 (0.334)
25	-1.615 (0.433)	-0.391 (0.044)	16.004	1.796 (0.188)

performs worse than $k = 4$ and $k = 6$, that there is a "lack of difference" between $k = 6$ and $k = 7$, and even consider $k = 15$ as being sufficiently accurate to be considered the ground truth. We find $k(10000, 10) = 4$, which is again consistent with, and serves to explain, these empirical findings.

Finally, in an extreme example, Liu and Pierce (1994) consider the case $M = 300$ and $m = 1$. Since our recommendations are based on a worst-case asymptotic analysis, we have $k(M, 1) = \infty$, emphasizing that when $m = 1$, the sample provides no help in obtaining a likelihood close enough to log-quadratic for the Laplace approximation to be accurate. In their Example 1, Liu and Pierce construct the model such that the integrand defining the joint likelihood is already nearly log-quadratic for any m, and report that the Laplace approximation appears sufficient in this case. However, in their Example 2, they consider a more complicated integrand, and find that (a) the Laplace approximation to the likelihood is poor, and (b) that the AQ approximation does not visually converge as k is taken larger. That the error in the likelihood approximation fails to converge is consistent with our worst-case observation that when $m = 1$, taking $k \to \infty$ does not guarantee that the approximation error goes to zero.

5 Applications

5.1 Toenail Data

Lesaffre and Spiessens (2001) report that the choice of k has a significant impact on inferences made in a grouped binary model for absence of toenail infection in $M = 294$ patients at $m_i \in \{2, \ldots, 7\}$ visits, after receiving one of two oral treatments for the condition. Specifically, they illustrate the confusion for practitioners surrounding the choice of k, listing conflicting recommendations present at the time in the documentation provided by various software packages (SERC, 1989; Wolfinger, 1999; Lillard and Panis, 2000; Rabe-Hesketh et al., 2002; Bates et al., 2015; Rizopoulos, 2020). Lesaffre and Spiessens conclude in this specific application that $k = 10$ appears reasonable based on extensive empirical observations.

These data have $m = 2$, and we find that $k(294, 2) = 11$, providing a theoretical justification for the empirically-motivated recommendation of Lesaffre and Spiessens (2001). Table 1 shows the results of fitting their model for various choices of k. We see that the estimates appear to stop

changing with k around $k = 11$. Further, we find that this yields comparable computational time to $k = 1$, while models with larger k start to exhibit an increase in computational time, an interesting phenomenon that we elaborate on in Appendix C.

5.2 Weibull Survival Model with Gamma Frailties

The modelling of grouped time-to-event data is a common scenario in which GLMMs with non-exponential family response and non-Gaussian random effects are used. Theorems 1 and 2 apply to this challenging case, and we wish to investigate the practical use of $k(M, m)$ for such models.

Liu and Huang (2008) consider a Weibull proportional hazards GLMM for grouped data. In simulations with $M = 100$ and $m = 6$, they (a) report that AQ with $k = 5$ and $k = 10$ give similar results, and (b) explicitly recommend GQ with $k = 30$ over AQ with any k. We find that $k(M, m) = 2$ in this case, explaining their findings that the error appears stable for $k \geq 5$.

Liu and Huang (2008) do not attempt to explain their recommendation of $k = 30$ for GQ, stating only that the situation warrants further attention, which we now provide. Their recommendation to use GQ appears to be based on simulations with relatively small $m = 6$ and where the true random effect is generated as $U_i \overset{i.i.d.}{\sim} \mathrm{N}(0, 1)$. A GQ rule with $k = 30$ has 99.65% of its total weight within the interval $(-3, 3)$, which also contains 99.73% of the mass of the true random effects distribution, and further, their true mode value $z^* = 0 \in \mathcal{Q}(1, k)$. Under their specific construction, the GQ rule can be expected to be very similar to the corresponding AQ rule. However, Theorem 1 says that no matter what k is, the GQ rule will relative error that fails to converge as $m \to \infty$. We therefore suggest that their claim that GQ attains good empirical performance can be explained by its overlap with AQ in their specific setting.

Crowther et al. (2014) and Sahu et al. (1997) consider the Weibull proportional hazards model fit to the classic dataset of McGilchrist and Aisbett (1991). These data contain information on the time to recurrance of infection of $M = 38$ patients' $m = 2$ kidneys. The log-hazard of infection recurrance is associated with age, sex, and a categorical disease status, and depends on three further parameters μ, α, σ^2 associated with the Weibull model. Crowther et al. (2014) report that AQ with $k = 9$ and $k = 10$ give nearly identical answers. Their empirical finding is again consistent with our recommendation of $k(38, 2) = 6$ for these data, providing an example of $k(M, m)$ being useful despite the low group size and the complexity of the model. Sahu et al. (1997) specifically consider the use of Gamma frailties, fitting their model in a Bayesian context using Gibbs sampling to circumvent the challenge of approximating the marginal likelihood.

We fit the Weibull proportional hazards model with multiplicative Gamma frailties (log-Gamma random effects) of Sahu et al. (1997) using AQ with various k in a range around $k(M, m) = 6$, and display the results in Table 2 for the parameters considered by Sahu et al. (1997). With $M = 38$ and $m = 2$ both small, and non-exponential family response and non-Gaussian random effects, this GLMM can be described as challenging. We see that the estimates appear to stop changing for k just above $k(M, m)$. The Laplace approximation has estimates that are far different from the others, in line with our recommendation to avoid it, since $k(M, m) > 1$.

6 Discussion

In this paper, we have (a) argued that fixed quadrature rules must be adapted to the data in order to have error converging to zero asymptotically, (b) provided a stochastic asymptotic rate for the error induced by approximating the intractable marginal likelihood in a very general class of GLMMs using adaptive quadrature, and (c) leveraged this rate to make a concrete recommendation on the

Table 2: Estimates and 95% Wald-type confidence intervals for the Weibull survival model fit using AQ for various k. These data have $k(M, m) = 6$, which appears to fall just below the point at which the estimates appear to stop changing.

k	$\beta_{\texttt{sex}}$	μ	α	σ^2
1	0.121 (-0.305,0.547)	1.863 (0.682,5.091)	0.410 (0.037,4.578)	0.203 (0.002,17.482)
3	-1.724 (-2.667,-0.782)	0.018 (0.004,0.087)	1.087 (0.795,1.486)	0.135 (0.002,10.312)
5	-1.872 (-2.920,-0.824)	0.016 (0.003,0.082)	1.145 (0.823,1.593)	0.243 (0.015,3.984)
6	-2.129 (-3.367,-0.890)	0.016 (0.003,0.087)	1.184 (0.828,1.693)	0.304 (0.026,3.601)
7	-1.944 (-2.995,-0.892)	0.016 (0.003,0.079)	1.170 (0.863,1.587)	0.310 (0.041,2.346)
9	-2.025 (-3.115,-0.934)	0.016 (0.003,0.080)	1.180 (0.866,1.607)	0.327 (0.046,2.316)
11	-1.898 (-2.956,-0.840)	0.016 (0.003,0.078)	1.175 (0.864,1.598)	0.325 (0.047,2.267)

number of quadrature points to use in practice. We have shown evidence that this recommendation recovers and explains the ad-hoc recommendations made by at least ten previous studies, and that it produces reasonable results in terms of accuracy and computation time for previously-considered analysis of grouped binary and survival data with Gaussian and non-Gaussian random effects.

Acknowledgements

We are grateful for helpful comments from Helen Ogden and Glen McGee. Blair Bilodeau is supported by an NSERC Canada Graduate Scholarship and the Vector Institute.

References

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

S. Bianconcini. Asymptotic properties of adaptive maximum likelihood estimators in latent variable models. *Bernoulli*, 20(3):1507–1531, 2014.

B. Bilodeau, A. Stringer, and Y. Tang. Stochastic convergence rates and applications of adaptive quadrature in Bayesian inference, 2021. arXiv:2102.06801.

B. M. Bolker, M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. H. Stevens, and J.-S. S. White. Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology and Evolution*, 42(3):127–135, 2008.

R. Bono, R. Alarcon, and M. J. Blanca. Report quality of generalized linear mixed models in psychology: A systematic review. *Frontiers in Psychology*, 12, 2021.

N. E. Breslow and D. G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25, 1993.

M. J. Crowther, M. P. Look, and R. D. Riley. Multilevel mixed effects parametric survival models using adaptive Gauss–Hermite quadrature with application to recurrent events and individual participant data meta-analysis. *Statistics in Medicine*, 33:3844–3858, 2014.

D. Hedeker and R. D. Gibbons. A random-effects ordinal regression model for multilevel analysis. *Biometrics*, 50(4):933–944, 1994.

F. Heiss and V. Winschel. Likelihood approximation by numerical integration on sparse grids. *Journal of Econometrics*, 144:62–80, 2008.

L. Jiang, Y. Ding, M. A. Sutherland, M. K. Hutchinson, C. Zhang, and B. Si. A novel sparse model-based algorithm to cluster categorical data for improved health screening and public health promotion. *IISE Transactions on Healthcare Systems Engineering*, 2021.

S. Jin and B. Andersson. A note on the accuracy of adaptive Gauss-Hermite quadrature. *Biometrika*, 107(3):737–744, 2020.

H. Joe. Accuracy of Laplace approximation for discrete response mixed models. *Computational Statistics and Data Analysis*, 52:5066–5074, 2008.

P. Kabaila and N. Ranathunga. On adaptive Gauss-Hermite quadrature for estimation in GLMM's. In H. Nguyen, editor, *Statistics and Data Science*, pages 130–139. Springer Singapore, 2019.

R. E. Kass, L. Tierney, and J. B. Kadane. The validity of posterior expansions based on Laplace's Method. *Bayesian and Likelihood Methods in Statistics and Econometrics*, pages 473–488, 1990.

B. Kleijn and A. van der Vaart. The Bernstein von-Mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354–381, 2012.

Y. Lee and J. A. Nelder. Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society, Series B (Methodology)*, 587(4):619–678, 1996.

E. Lesaffre and B. Spiessens. On the effect of the number of quadrature points in a logistic random-effects model: An example. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 50(3):325–335, 2001.

L. A. Lillard and C. W. A. Panis. *Multiprocess Multilevel Modelling, aML Release 1, User's Guide and Reference Manual*, 2000.

L. Liu and X. Huang. The use of Gaussian quadrature for estimation in frailty proportional hazards models. *Statistics in Medicine*, 27:2665–2683, 2008.

Q. Liu and D. A. Pierce. A note on Gauss-Hermite quadrature. *Biometrika*, 81(3):624–629, 1994.

C. E. McCulloch. Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, 92(437):162–170, 1997.

C. A. McGilchrist and C. W. Aisbett. Regression with frailty in survival analysis. *Biometrics*, 47 (2):461–466, 1991.

J. Naylor and A. F. M. Smith. Applications of a method for the efficient computation of posterior distributions. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 31(3):214–225, 1982.

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

H. Ogden. On the error in Laplace approximations of high-dimensional integrals. *Stat*, 10, 2021.

J. Pan and R. Thompson. Gauss-Hermite quadrature approximation for estimation in generalised linear mixed models. *Computational Statistics*, 18:57–78, 2003.

K. Petras. Smolyak cubature of given polynomial degree with few nodes for increasing dimension. *Numerische Mathematik*, 93:729–753, 2003.

J. C. Pinheiro and D. M. Bates. Approximations to the log-likelihood function in the nonlinear mixed effects models. *Journal of Computational and Graphical Statistics*, 4(1):12–35, 1995.

S. Rabe-Hesketh, A. Skrondal, and A. Pickles. Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal*, 2(1):1–21, 2002.

D. Rizopoulos. *GLMMadaptive: Generalized Linear Mixed Models using Adaptive Gaussian Quadrature*, 2020. URL `https://CRAN.R-project.org/package=GLMMadaptive`. R package version 0.7-15.

S. K. Sahu, D. K. Dey, H. Aslanidou, and D. Sinha. A Weibull regression model with gamma frailties for multivariate survival data. *Lifetime Data Analysis*, 3:123–137, 1997.

S. Schilling and R. Bock. High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika*, 70:533–555, 2005.

SERC. *EGRET Users' Manual*, 1989.

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

A. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

E. F. Vonesh. A note on the use of Laplace's approximation for nonlinear mixed effect models. *Biometrika*, 83(2):447–452, 1996.

R. D. Wolfinger. Fitting nonlinear mixed models with the new NLMIXED procedure. In *In Proceedings of the 24th SAS User's Group Meeting. Cary, NC: SAS Institute, Inc.*, 1999.

H. Wu and M. P. Jones. Proportional likelihood ratio mixed model for fiscrete longitudinal data. *Statistics in Medicine*, 40:2272–2285, 2021.

S. Zeger and M. R. Karim. Generalized linear models with random effects: A Gibbs sampling approach. *Journal of the American Statistical Association*, 86:79–86, 1991.

A Uniform Assumptions

We repeat here the assumptions from Bilodeau et al. (2021) used to prove their Theorem 1, but with the appropriate uniform modifications that imply part (b) of Theorem 2. We also use explicitly the notation of the present paper for GLMMs, while Bilodeau et al. (2021) describe conditions for a generic stochastic integral approximation.

For any data-generating distribution P_n^* and fixed M, we require the following to hold: there exist $\delta_\Theta, \delta > 0$, $\boldsymbol{\theta}^* \in \Theta$, and $\{\boldsymbol{U}_{\boldsymbol{\theta}}^* \mid \boldsymbol{\theta} \in \mathcal{B}_{\delta_\Theta}(\boldsymbol{\theta}^*)\} \subseteq \mathbb{R}^p$ for each such that all five statements are true.

Assumption 1. *There exists $t, D > 0$ such that for all $\boldsymbol{\alpha} \subseteq \mathbb{N}^p$ with $0 \leq |\boldsymbol{\alpha}| \leq t$, for all $i \in [M]$*

$$\lim_{n\to\infty} P_n^*\Big[\sup_{\boldsymbol{\theta}\in\mathcal{B}_{\delta_\Theta}(\boldsymbol{\theta}^*)} \sup_{\boldsymbol{U}\in\mathcal{B}_\delta(\boldsymbol{U}_{\boldsymbol{\theta}}^*)} |\partial_{\boldsymbol{U}}^{\boldsymbol{\alpha}} \ell_i(\boldsymbol{y}_i, \boldsymbol{U}; \boldsymbol{\theta})| < nD\Big] = 1.$$

Assumption 2. *There exist* $0 < \underline{\eta} \leq \overline{\eta} < \infty$ *such that for all* $i \in [M]$,

$$\lim_{n\to\infty} P_n^*\Big[n\underline{\eta} \leq \inf_{\boldsymbol{\theta}\in\mathcal{B}_{\delta_\Theta}(\boldsymbol{\theta}^*)} \inf_{\boldsymbol{U}\in\mathcal{B}_\delta(\boldsymbol{U}_{\boldsymbol{\theta}}^*)} \eta_p(-\partial_{\boldsymbol{U}}^2 \ell_i(\boldsymbol{y}_i, \boldsymbol{U}; \boldsymbol{\theta})) \leq \sup_{\boldsymbol{\theta}\in\mathcal{B}_{\delta_\Theta}(\boldsymbol{\theta}^*)} \sup_{\boldsymbol{U}\in\mathcal{B}_\delta(\boldsymbol{U}_{\boldsymbol{\theta}}^*)} \eta_1(-\partial_{\boldsymbol{U}}^2 \ell_i(\boldsymbol{y}_i, \boldsymbol{U}; \boldsymbol{\theta})) \leq n\overline{\eta}\Big] = 1,$$

where $\eta_j(\boldsymbol{H})$ *denotes the* j*th biggest eigenvalue of a matrix* $\boldsymbol{H}$.

Assumption 3. *There exists* $b > 0$ *such that for all* $i \in [M]$,

$$\lim_{n\to\infty} P_n^*\Big[\sup_{\boldsymbol{\theta}\in\mathcal{B}_{\delta_\Theta}(\boldsymbol{\theta}^*)} \sup_{\boldsymbol{U}\in[\mathcal{B}_\delta(\boldsymbol{U}_{\boldsymbol{\theta}}^*)]^c} \ell_i(\boldsymbol{y}_i, \boldsymbol{U}; \boldsymbol{\theta}) - \ell_i(\boldsymbol{y}_i, \boldsymbol{U}_{\boldsymbol{\theta}}^*; \boldsymbol{\theta}) \leq -nb\Big] = 1.$$

Assumption 4. *For any* $\beta > 0$ *and function* $G(n)$ *such that* $\lim_{n\to\infty} G(n) = \infty$, *for all* $i \in [M]$

$$\lim_{n\to\infty} P_n^*\left[\sup_{\boldsymbol{\theta}\in\mathcal{B}_{\delta_\Theta}(\boldsymbol{\theta}^*)} \frac{\sqrt{n}}{G(n)} \left\|\widehat{\boldsymbol{U}}_i^{\boldsymbol{\theta}} - \boldsymbol{U}^*\right\|_2 \leq \beta\right] = 1.$$

Assumption 5. *There exist* $0 < c_1 < c_2 < \infty$ *such that for all* $i \in [M]$,

$$c_1 \leq \inf_{\boldsymbol{\theta}\in\mathcal{B}_{\delta_\Theta}(\boldsymbol{\theta}^*)} \inf_{\boldsymbol{U}\in\mathcal{B}_\delta(\boldsymbol{U}_{\boldsymbol{\theta}}^*)} g_i(\boldsymbol{U}; \boldsymbol{\sigma}) \leq \sup_{\boldsymbol{\theta}\in\mathcal{B}_{\delta_\Theta}(\boldsymbol{\theta}^*)} \sup_{\boldsymbol{U}\in\mathcal{B}_\delta(\boldsymbol{U}_{\boldsymbol{\theta}}^*)} g_i(\boldsymbol{U}; \boldsymbol{\sigma}) \leq c_2.$$

B Proofs

of Theorem 1. We have

$$\frac{\sum_{i=1}^N \pi(\boldsymbol{y} \mid \boldsymbol{z}_i)\pi(\boldsymbol{z}_i)\,\omega_i}{\int \pi(\boldsymbol{y} \mid \boldsymbol{\xi})\pi(\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi}} = \frac{\sum_{i=1}^N \pi(\boldsymbol{z}_i \mid \boldsymbol{y})\pi(\boldsymbol{y})\,\omega_i}{\pi(\boldsymbol{y})} = \sum_{i=1}^N \pi(\boldsymbol{z}_i \mid \boldsymbol{y})\,\omega_i,$$

where $\pi(\boldsymbol{\xi} \mid \boldsymbol{y})$ denotes the posterior under π with normalizing constant $\pi(\boldsymbol{y})$. We consider two separate cases.

Let $\boldsymbol{\xi}^*$ be prescribed by Kleijn and van der Vaart (2012, Equation 2.1).

First, suppose $\boldsymbol{\xi}^* \notin \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N\}$, and observe that

$$\sum_{i=1}^N \pi(\boldsymbol{z}_i \mid \boldsymbol{y})\,\omega_i \leq N(\max_{i\in[N]} \omega_i)(\max_{i\in[N]} \pi(\boldsymbol{z}_i \mid \boldsymbol{y})).$$

Since $\boldsymbol{\xi}^* \notin \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N\}$, Kleijn and van der Vaart (2012, Theorem 2.1) implies $\max_{i\in[N]} \pi(\boldsymbol{z}_i \mid \boldsymbol{y}) \xrightarrow{p} 0$, giving the result.

Otherwise, suppose $\boldsymbol{\xi}^* \in \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N\}$, and observe that

$$\sum_{i=1}^N \pi(\boldsymbol{z}_i \mid \boldsymbol{y})\,\omega_i \geq (\min_{i\in[N]} \omega_i)(\max_{i\in[N]} \pi(\boldsymbol{z}_i \mid \boldsymbol{y})).$$

Since Kleijn and van der Vaart (2012, Theorem 2.1) implies $\pi(\boldsymbol{\xi}^* \mid \boldsymbol{y}) \xrightarrow{p} \infty$, $\boldsymbol{\xi}^* \in \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N\}$ implies the result. □

of Theorem 2. Fix M, and denote the data-generating probability measure by P_n^*. For each $i \in [M]$ and $\boldsymbol{\theta} \in \Theta$, we apply Theorem 1 of Bilodeau et al. (2021) to the likelihood $u \mapsto f_i(\boldsymbol{y}_i; \boldsymbol{\theta}, u)$ and prior $u \mapsto g_i(u; \boldsymbol{\theta})$ using k_i points for the approximation. Naively, under the appropriate regularity conditions this implies that for each $\boldsymbol{\theta} \in \Theta$ and $i \in [M]$, there exists a $C_{i,\boldsymbol{\theta}} > 0$ such that

$$\lim_{m_i \to \infty} P_n^* \left(\left| \frac{\widetilde{\pi}_i^{\texttt{AQ}}(\boldsymbol{y}_i; \boldsymbol{\theta})}{\pi_i(\boldsymbol{y}_i; \boldsymbol{\theta})} - 1 \right| < C_{i,\boldsymbol{\theta}}\, m_i^{-\lfloor (k_i+2)/3 \rfloor} \right) = 1.$$

However, this is a pointwise result in $\boldsymbol{\theta}$. In fact, under the appropriate further condition that the regularity conditions hold uniformly over the ball $\mathcal{B}_{\delta_\Theta}(\Theta)$ with fixed radius $\delta > 0$ (see Appendix A for the precise conditions), the following stronger statement holds. For each $i \in [M]$, there exists a $C_i > 0$

$$\lim_{m_i \to \infty} P_n^* \left(\sup_{\boldsymbol{\theta} \in \mathcal{B}_{\delta_\Theta}(\Theta)} \left| \frac{\widetilde{\pi}_i^{\texttt{AQ}}(\boldsymbol{y}_i; \boldsymbol{\theta})}{\pi_i(\boldsymbol{y}_i; \boldsymbol{\theta})} - 1 \right| < C_i\, m_i^{-\lfloor (k_i+2)/3 \rfloor} \right) = 1.$$

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 $i \in [M]$,

$$\lim_{m_i \to \infty} P_n^* \left(\sup_{\boldsymbol{\theta} \in \mathcal{B}_{\delta_\Theta}(\Theta)} |\log \widetilde{\pi}_i^{\texttt{AQ}}(\boldsymbol{y}_i; \boldsymbol{\theta}) - \log \pi_i(\boldsymbol{y}_i; \boldsymbol{\theta})| < 2C_i\, m_i^{-\lfloor (k_i+2)/3 \rfloor} \right) = 1.$$

Finally, since M is fixed, using a union bound gives that there exists some $C > 0$ such that

$$\lim_{m_1, \ldots, m_M \to \infty} P_n^* \left(\sup_{\boldsymbol{\theta} \in \mathcal{B}_{\delta_\Theta}(\Theta)} |\log \widetilde{\pi}^{\texttt{AQ}}(\boldsymbol{y}; \boldsymbol{\theta}) - \log \pi(\boldsymbol{y}; \boldsymbol{\theta})| < C \sum_{i=1}^{M} m_i^{-\lfloor (k_i+2)/3 \rfloor} \right) = 1.$$

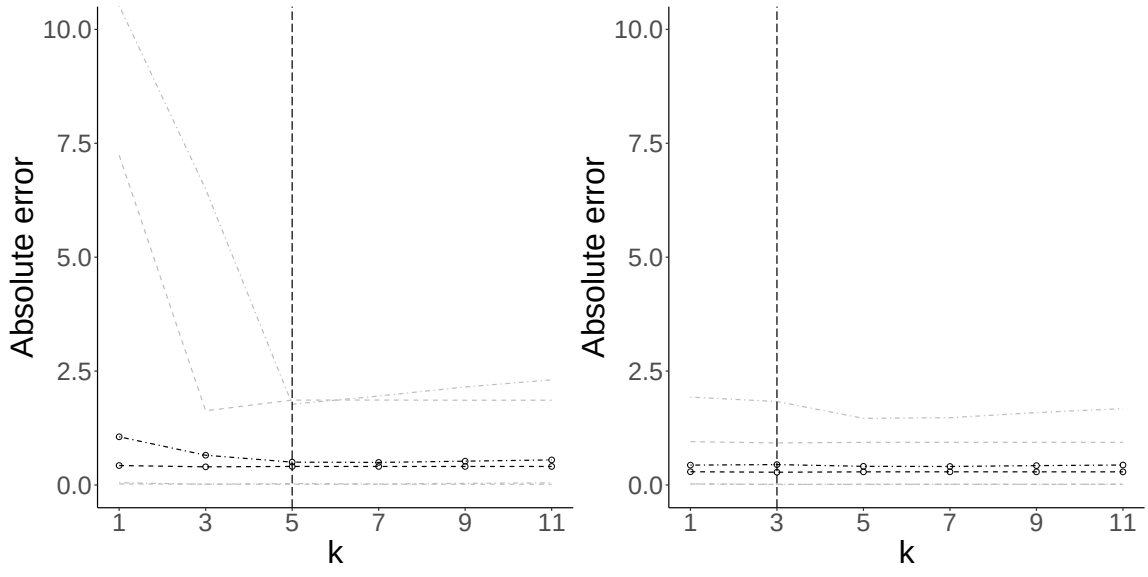
□

C Simulations

We conduct our own simulation study to evaluate the quality of inferences in a binary GLMM based on AQ with $k(M, m)$ points. Even in simulations, there is no correct or true value of k, and we suggest that a reasonable goal ought to be to fit the model as accurately as possible without employing wasteful computations in the form of using too many quadrature points. Accordingly, we study how both the approximation error and the computation time change with increasing k.

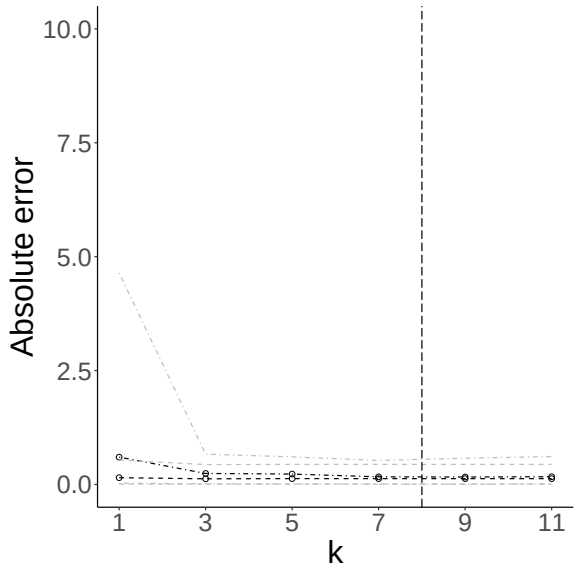
For each unique combination of $M \in \{100, 1000\}$, $m \in \{3, 5\}$, and $\sigma \in \{1, 3\}$ we sample 500 datasets from a binary regression model with logistic link, having $F_i(\mu_{ij}) = \text{Bern}(\mu_{ij})$, $\text{logit}(\mu_{ij}) = \boldsymbol{x}_{ij}^{\text{T}}\boldsymbol{\beta} + U_i$, $\boldsymbol{x}_{ij} = (1, x_i, t_j, x_i t_j)$ with $x_{ij} = \{0, 1\}$ with equal probability for each $i \in [M]$ and $t = 0, \ldots, (m-1)$, and $G_i = \text{N}(0, \sigma^2)$. We fit the model using `lme4::glmer` (Bates et al., 2015). We report the 2.5%, 50%, and 97.5% quantiles of the absolute error in estimating the intercept β_0 and the random effects standard deviation σ, as well as empirical coverages for 95% Wald-type confidence intervals for β_0. The parameters we report are the ones for which estimation appeared the most difficult; the other parameters were all estimated more accurately, and appeared less sensitive to the choice of k.

Figures 1–3 show the estimation errors and coverages from each configuration. The point in each plot where the error stops changing with k can be interpreted as the point where the approximation

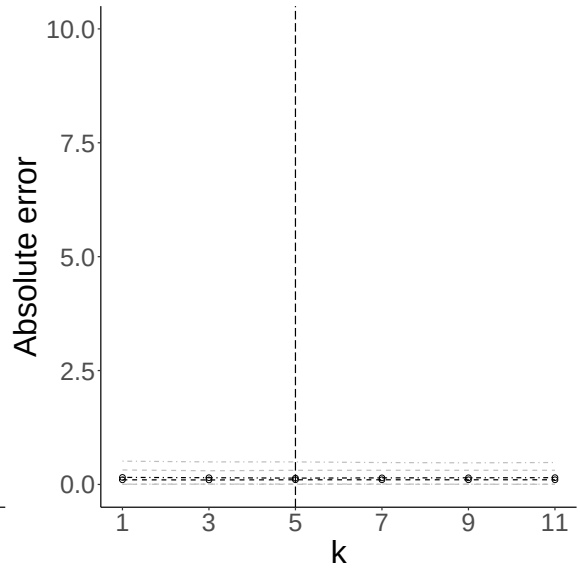


(a) $M = 100, m = 3$

(b) $M = 100, m = 5$

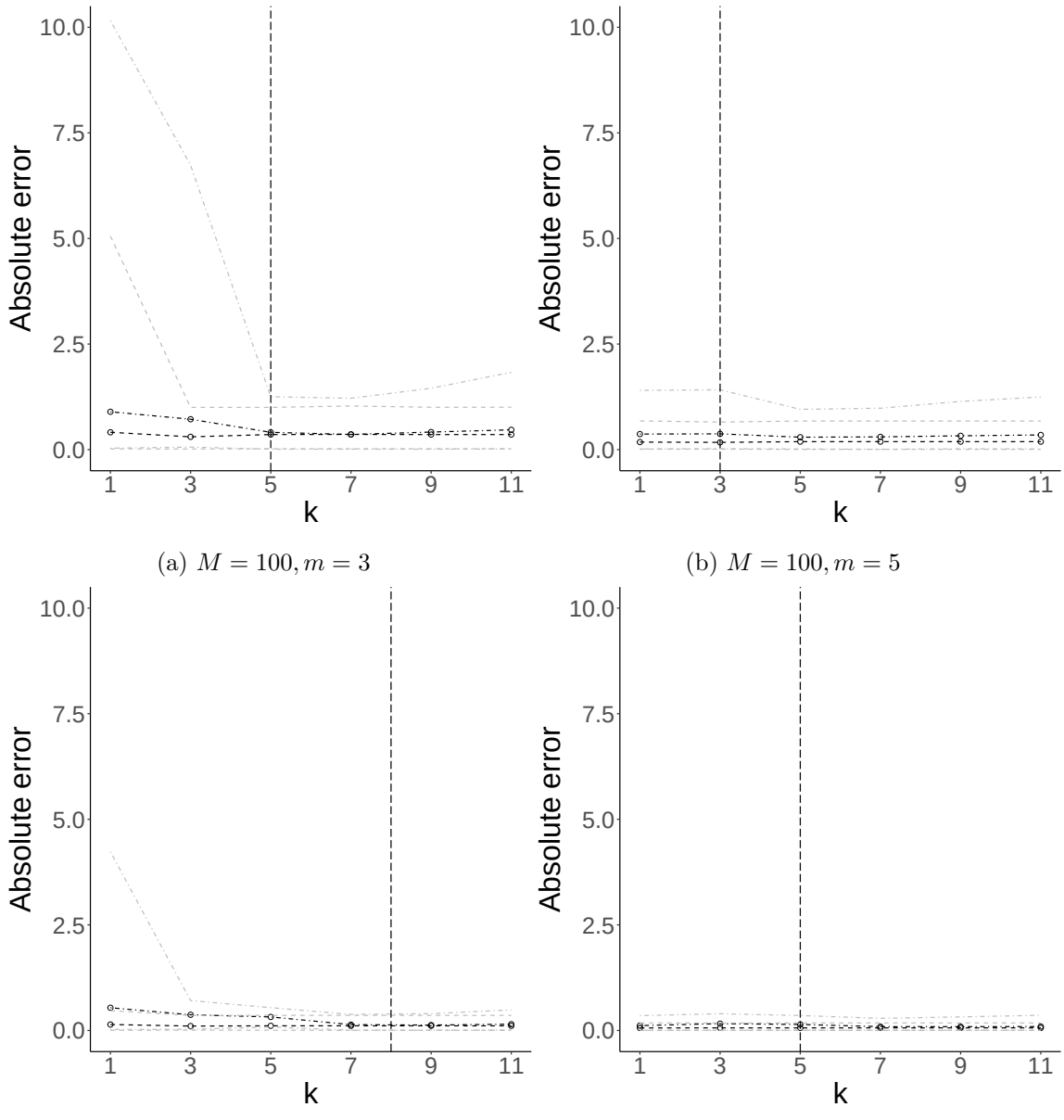


(c) $M = 1000, m = 3$



(d) $M = 1000, m = 5$

Figure 1: Approximation error for β_0 as a function of k for the simulation of Appendix C. Shown are square-roots of the $2.5\%, 50\%$, and 97.5% quantiles of the squared error over 500 simulations, for $\sigma = 1$ (- - -) and $\sigma = 3$ (- · -). Verticle line (– – –) shows recommended $k(M, m)$.



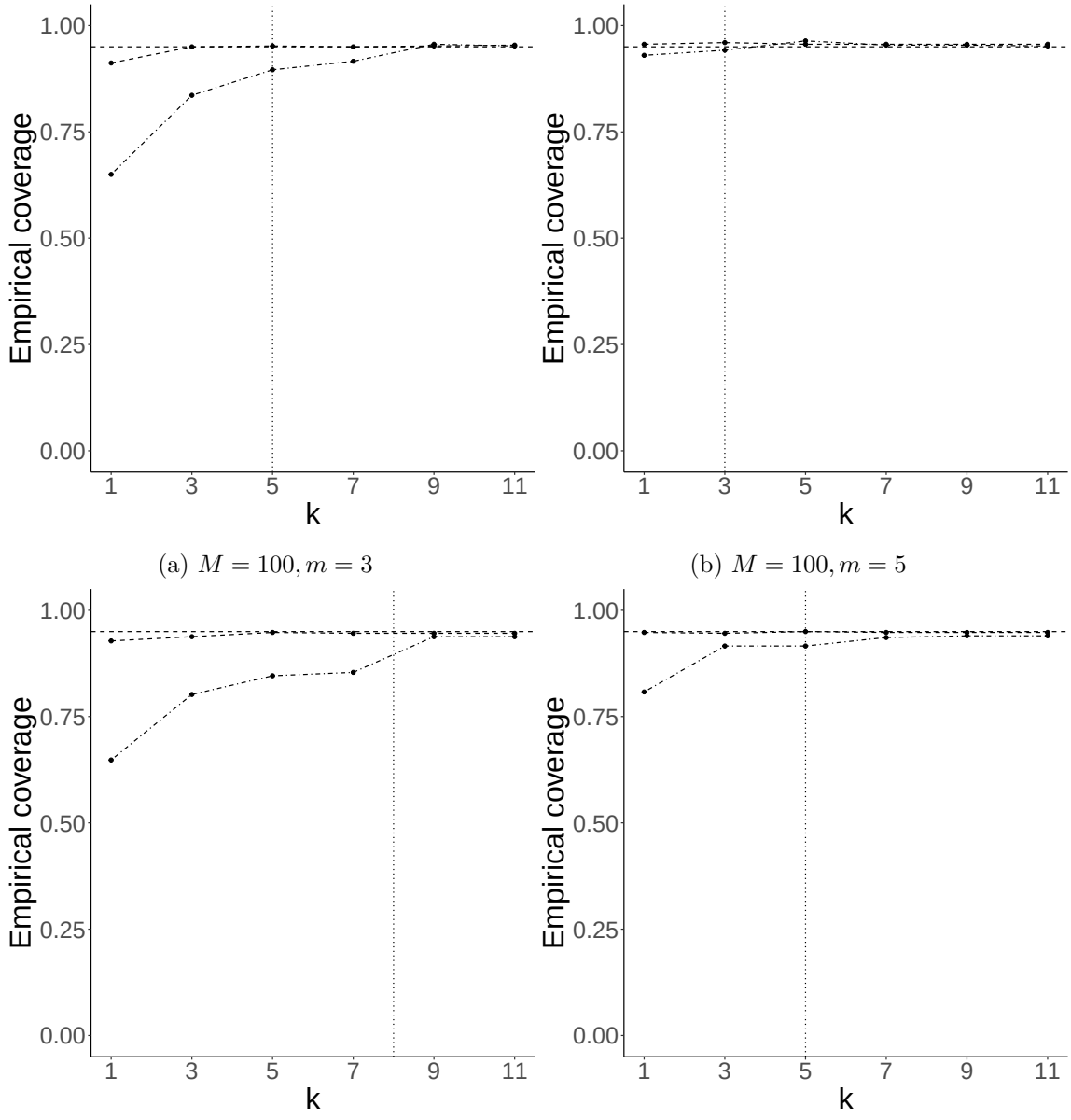
(a) $M = 100, m = 3$

(b) $M = 100, m = 5$

(c) $M = 1000, m = 3$

(d) $M = 1000, m = 5$

Figure 2: Approximation error for σ as a function of k for the simulation of Appendix C. Shown are square-roots of the 2.5%, 50%, and 97.5% quantiles of the squared error over 500 simulations, for $\sigma = 1$ (- - -) and $\sigma = 3$ (- · -). Verticle line (– – –) shows recommended $k(M, m)$.



(a) $M = 100, m = 3$

(b) $M = 100, m = 5$

(c) $M = 1000, m = 3$

(d) $M = 1000, m = 5$

Figure 3: Empirical coverages of 95% Wald-type confidence intervals for β_0 as a function of k for the simulation of Appendix C, for $\sigma = 1$ (- - -) and $\sigma = 3$ (- · -). Verticle line (– – –) shows recommended $k(M, m)$.

Table 3: Mean (SD) computation times in seconds, and relative change in computation time, for $k = 1$ and $k = k(M, m)$ for the simulations of Appendix C. The recommended $k(M, m)$ leads to at most a moderate increase, and in some cases a decrease, in computation time over the Laplace approximation.

		$M = 100$		$M = 1000$	
		$m = 3$	$m = 5$	$m = 3$	$m = 5$
$\sigma = 1$	Laplace	0.109(0.047)	0.120(0.022)	0.448(0.077)	0.905(0.150)
	$k(M, m)$	0.113(0.031)	0.131(0.024)	0.652(0.105)	1.096(0.182)
	Relative	3.54%	8.40%	31.38%	17.43%
$\sigma = 3$	Laplace	0.168(0.059)	0.158(0.042)	0.806(0.172)	1.185(0.396)
	$k(M, m)$	0.138(0.031)	0.158(0.035)	0.970(0.154)	1.320(0.207)
	Relative	−21.84%	−0.00%	16.91%	10.23%

error is dominated by the estimation error. We broadly see that $k(M, m)$ is located beyond the points where the error starts to level off, indicating slight conservatism. In the case of $\sigma = 3$ and $M = 100$ (Fig. 3c), the confidence interval coverages improve up to around $k \approx k(M, m) + 2$ or so. A standard deviation of $\sigma = 3$ indicates that most random intercepts can be a-priori expected to lie between -6 and 6 on the logit scale, a wide range which we suggest can be considered atypical, or at least very challenging.

Table 3 shows the absolute and relative computation times for using AQ with $k(M, m)$ compared to the Laplace approximation. The worst observed increase is around 32%, and occurs when σ and m are low but M is high. One case with high σ and low M actually has faster computation times when using $k(M, m)$ compared to Laplace. Bianconcini (2014) discusses how using a higher k, and hence potentially more accurate approximation, may lead to faster convergence of the optimization required to find the maximum likelihood estimate and less likelihood evaluations overall, leading to reduced computation times over less accurate likelihood approximations despite each individual evaluation being slower. Indeed, when $M = 100, m = 3$, and $\sigma = 3$, we find the likelihood optimization requires a mean (SD) of 519(136) and 415(87) evaluations for Laplace and $k(M, m)$ respectively. Though each evaluation is more expensive when $k > 1$, that there is less of them can lead to a decrease in total computation times, or at least mitigate any increase. This same pattern was observed in Table 1 in Subsection 5.1.