On a Hybrid Data Cloning Method and Its Application in Generalized Linear Mixed Models

Hossein Baghishani[a,*], Håvard Rue[b], Mohsen Mohammadzadeh[a]

[a] *Department of Statistics, Tarbiat Modares University, Tehran, Iran*
[b] *The Norwegian University of Science and Technology, Trondheim, Norway*

Abstract

Data cloning method is a new computational tool for computing maximum likelihood estimates in complex statistical models such as mixed models. This method is synthesized with integrated nested Laplace approximation to compute maximum likelihood estimates efficiently via a fast implementation in generalized linear mixed models. Asymptotic behavior of the hybrid data cloning method is discussed. The performance of the proposed method is illustrated through a simulation study and real examples. It is shown that the proposed method performs well and rightly justifies the theory. Supplemental materials for this article are available online.

Keywords: Approximate Bayesian Inference, Asymptotic Normality, Data Cloning, Generalized Linear Mixed Models, Integrated Nested Laplace Approximation, Stein's Identity.

1. Introduction

Non-Gaussian repeated measurements such as longitudinal and clustered data are common in many sciences such as biology, ecology, epidemiology and medicine. The Generalized Linear Mixed Models (GLMMs) are a flexible and extensively used class of models for modeling these types of data. As an extension of generalized linear models (GLMs) (McCullagh and Nelder, 1989), a GLMM assumes that the response variable follows a distribution from the exponential family and is conditionally independent given latent variables, while the latent variables are modeled by random effects that are typically Gaussian (Breslow and Clayton, 1993).

[*] To whom correspondence should be addressed.

Email addresses: hbaghishani@modares.ac.ir (Hossein Baghishani), hrue@math.ntnu.no (Håvard Rue), mohsen_m@modares.ac.ir (Mohsen Mohammadzadeh)

January 14, 2011

Statistical inferences in such models have been the subject of a great deal of research over the two past decades. Both frequentist and Bayesian methods have been developed for inference in GLMMs (McCulloch, 1997). Computational difficulties rendered likelihood based inferences for GLMMs prohibitive, i.e. computing the likelihood function needed for such inferences requires computing an intractable, high dimensional integral. Due to the advances in Markov chain Monte Carlo (MCMC) sampling methods, nowadays, a commonly used approach for inference in these models is based on the Bayesian paradigm. However, Bayesian inferences depend on the choice of the prior distributions and the specification of prior distributions is not straightforward in particular for variance components (Fong *et al.*, 2010). Moreover, MCMC algorithms applied to these models come with a wide range of problems in terms of convergence and computational time.

A recent suitable alternative method to carry out statistical inferences in a GLMM could be the data cloning (DC) method, which was first introduced by Lele *et al.* (2007) in ecological studies. This method, as a computational trick, uses an MCMC algorithm to sample from an artificially constructed distribution, named DC-based distribution, for computing maximum likelihood estimates (MLE) and their variance estimates. The trick is to generate samples from a DC-based distribution constructed by duplicating the original data set enough times, k say, such that the sample mean as well as the scaled sample variance converge to MLE and its variance estimate.

Computation, however, is an important issue since the method applies intensive MCMC simulations to k clones of the data. This issue becomes more drastic when one requires using an increasing sequence of k. As Christian P. Robert have discussed in his personal weblog (`http://xianblog.wordpress.com`) if k is large enough, the MCMC algorithm will face difficulties in the exploration of the parameter space, and hence in the subsequent discovery of the global modes, while, if k is too small, there is no certainty that the algorithm will identify the right mode. The practical implementation thus requires using an increasing sequence of k's, which is very demanding on computing time, especially when k is large, and thus cancels somehow the appeal of the method.

The computational challenges of DC method motivates us to synthesize DC method with integrated nested Laplace approximation (INLA), which is the main purpose of this paper, introduced by Rue and Martino (2007) and Rue *et al.* (2009) to compute MLE efficiently. It would be expected that synthesizing of the DC method by INLA can reduce the computational efforts severely.

Asymptotic behavior of the proposed method for a GLMM is explored as well.

The paper, generally, is organized into two parts. The first (main) part illustrates computational aspects of the proposed hybrid method accessible to most of the readers who are less interested in theoretical aspects (Sections 2-5). In this part, we present our main results, however in an imprecise form. Formal statements of our results are given and proved in the second part (Appendices).

In the next section we describe the model and INLA methodology. Section 3 describes the new hybrid DC (HDC) method. The performance of the method is explored through simulation studies and real data examples in Sections 4 and 5. Finally, Section 6 concludes with a brief discussion. All technical details and proofs of our results are relegated to the Appendices A-C and more technical details are available in the online supplemental file.

2. Model and INLA

In this section we introduce our basic model and INLA methodology.

2.1. The Model

Generalized linear mixed models are flexible models for modeling non-Gaussian repeated measurements. On the basis of the GLM, the GLMM assumes that the responses are independent conditional on the random effects and are distributed according to a member of the exponential family.

We consider clustered data in which repeated measures of a response variable are taken on a random sample of m clusters. Consider the response vectors $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})^T$, $i = 1, \ldots, m$. Let $n = \sum_{i=1}^m n_i$ be the total sample size. Conditional on $r \times 1$ vector of unobservable cluster-specific random effects $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{ir})^T$, these data are distributed according to a member of the exponential family:

$$f(y_{ij}|\boldsymbol{u}_i, \boldsymbol{\beta}) = \exp\{y_{ij}(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{v}_{ij}^T\boldsymbol{u}_i) - a(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{v}_{ij}^T\boldsymbol{u}_i) + c(y_{ij})\},$$

for $i = 1, \ldots, m; j = 1, \ldots, n_i$, in which $\boldsymbol{x}_{ij}$ and $\boldsymbol{v}_{ij}$, are the corresponding p- and r-dimensional covariate vectors associated with the fixed effects and the random effects respectively, $\boldsymbol{\beta}$ is a p-dimensional vector of unknown regression parameters, and $a(\cdot)$ and $c(\cdot)$ are specific functions. Here $\tau_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{v}_{ij}^T\boldsymbol{u}_i$ is the canonical parameter. Let $\mu_{ij} = E[Y_{ij}|\boldsymbol{\beta}, \boldsymbol{u}_i] = a^{'}(\tau_{ij})$ with $g(\mu_{ij}) = \eta_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{v}_{ij}^T\boldsymbol{u}_i$, where $g(\cdot)$ is a monotonic link function. Furthermore, assume $\boldsymbol{u}_i$ comes from a

Gaussian distribution, $\boldsymbol{u}_i|Q \sim N(\mathbf{0}, Q^{-1})$, in which the precision matrix $Q = Q(\boldsymbol{\theta})$ depends on parameters $\boldsymbol{\theta}$. Let $\boldsymbol{\theta}$ denote the $d \times 1$ vector of the variance components for which prior $\pi(\boldsymbol{\theta})$ is assigned. Let also $\boldsymbol{\psi} = (\boldsymbol{\beta}, \boldsymbol{u})$ denote the $q \times 1$ vector of parameters assigned Gaussian priors. Moreover, let $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m)$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, an open subset of $\Re^d$, and $\boldsymbol{\psi} \in \boldsymbol{\Psi}$, an open subset of $\Re^q$. The posterior density is defined by

$$\pi(\boldsymbol{\psi}, \boldsymbol{\theta}|\boldsymbol{y}) \propto \pi(\boldsymbol{\psi}|\boldsymbol{y}, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{y}) \propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{\beta})|Q(\boldsymbol{\theta})|^{1/2} \exp\left\{-\frac{1}{2}\boldsymbol{u}^T Q(\boldsymbol{\theta})\boldsymbol{u} + \sum_{i=1}^{m} \log f(\boldsymbol{y}_i|\boldsymbol{\psi})\right\}. \quad (1)$$

2.2. Integrated Nested Laplace Approximation

Because of usefulness and easy implementation of MCMC methods, the most commonly used approach for inference in the GLMMs is based on Bayesian methods and MCMC sampling. Considering (1) the main aim is to compute the posterior marginals $\pi(\psi_l|\boldsymbol{y})$, $l = 1, \ldots, q$ and $\pi(\theta_v|\boldsymbol{y})$, $v = 1, \ldots, d$. It is well known, however, that MCMC methods tend to exhibit poor performance when applied to such models (Rue *et al.*, 2009).

INLA is a new tool for Bayesian inference on latent Gaussian models introduced by Rue *et al.* (2009). The method combines Laplace approximations and numerical integration in a very efficient manner. INLA substitutes MCMC simulations with accurate, deterministic approximations to posterior marginal distributions. The quality of such approximations is high in most cases such that even very long MCMC runs could not detect any error in them.

We can write

$$\begin{aligned}
\pi(\psi_l|\boldsymbol{y}) &= \int \pi(\psi_l|\boldsymbol{\theta}, \boldsymbol{y})\pi(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}, \quad l = 1, \ldots, q, \\
\pi(\theta_v|\boldsymbol{y}) &= \int \pi(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}_{-v}, v = 1, \quad \ldots, d,
\end{aligned}$$

where $\boldsymbol{\theta}_{-v}$ is equal to $\boldsymbol{\theta}$ with eliminated vth element. The key feature of INLA is to use this form to construct nested approximations

$$\begin{aligned}
\tilde{\pi}(\psi_l|\boldsymbol{y}) &= \int \tilde{\pi}(\psi_l|\boldsymbol{\theta}, \boldsymbol{y})\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}, \\
\tilde{\pi}(\theta_v|\boldsymbol{y}) &= \int \tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}_{-v},
\end{aligned}$$

where Laplace approximation is applied to carry out the integrations required for evaluation of $\tilde{\pi}(\psi_l|\boldsymbol{\theta}, \boldsymbol{y})$. We have to note that the Laplace approximation does depend on reparameterization, but for these models, it is natural to keep the parameterization unchanged. For more details we refer the readers to Rue *et al.* (2009), page 387. The approximate posterior marginals obtained from INLA can then be used to compute summary statistics of interest, such as posterior means, variances or quantiles.

3. A Hybrid Data Cloning Method

In this section we describe how DC method could be synthesized with INLA. According to the theoretical results given in the Appendices A-C, we need to establish the asymptotic normality of the HDC-based distribution to be able to use the INLA within DC method. For this purpose, we first establish the asymptotic normality of the approximate posterior as well as the DC-based distributions. These two results will enable us to establish the asymptotic normality of the HDC-based distribution. The formal statements and their proofs are provided in the Appendices A-C.

A DC-based distribution is constructed by duplicating the original data set k times. In other words, we create a k-repeated cloned data set $\boldsymbol{y}^{(k)} = (\boldsymbol{y}, \ldots, \boldsymbol{y})$ where the observed data vector is repeated k times. In this way, the covariates are cloned as well. Finally, k independent copies of the random effects, $\boldsymbol{u}$, are generated from its Gaussian density, thus contributing to the cloned likelihood. Then, using the new cloned data, DC-based distribution will be constructed. Figure 1 shows schematically the steps of cloning data to construct DC-based distribution.

Let $\pi^{(k)}(\boldsymbol{\psi}, \boldsymbol{\theta}|\boldsymbol{y}) \propto \pi^{(k)}(\boldsymbol{\psi}|\boldsymbol{y}, \boldsymbol{\theta})\pi^{(k)}(\boldsymbol{\theta}|\boldsymbol{y})$ be the artificially constructed density, DC-based density, from k identical and independent clones of the data and prior distributions. Although this distribution looks like a Bayesian posterior distribution, but it is constructed from two functions which are not in fact a prior distribution and a likelihood. However, considering them as prior and likelihood can be mimicked virtually (Baghishani and Mohammadzadeh, 2011). A simple explanation for why the data cloning method works, is that by cloning the data the effect of the prior is diminished and the DC-based estimators converge to the MLEs. Furthermore, by the Central Limit Theorem, the DC-based estimators are approximately normally distributed and their variances are scaled versions of the asymptotic variances estimates for the MLEs, i.e.

$$\begin{aligned}
\mathrm{E}^{(k)}(\boldsymbol{\psi}, \boldsymbol{\theta}|\boldsymbol{y}) &\longrightarrow (\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\theta}}), \\
\mathrm{Var}^{(k)}(\boldsymbol{\psi}, \boldsymbol{\theta}|\boldsymbol{y}) &\longrightarrow k^{-1} \times \mathrm{Var}((\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\theta}})),
\end{aligned}$$

as $k \longrightarrow \infty$.

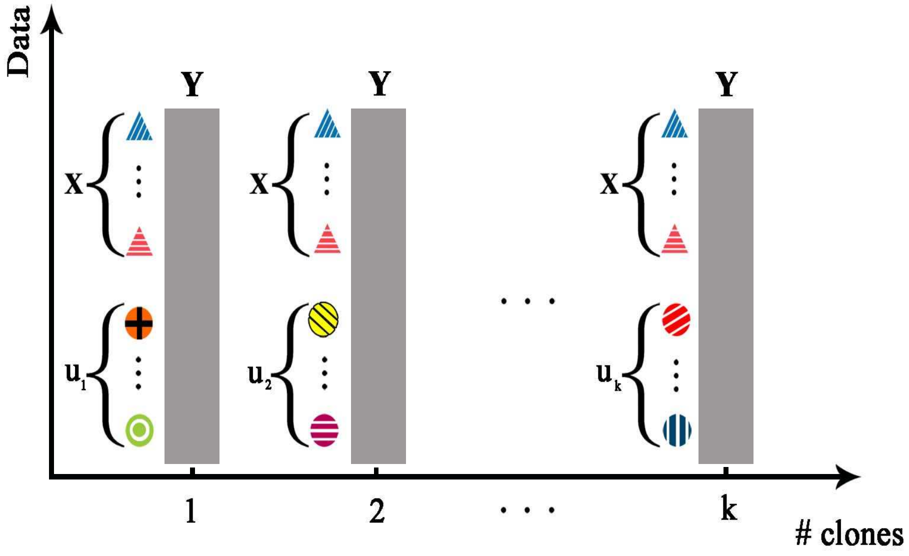


Figure 1: Cloning original data for constructing DC-based distribution

Following, we will combine the DC method with INLA by three steps. We can write

$$\log \pi(\boldsymbol{\psi}, \boldsymbol{\theta}|\boldsymbol{y}) \propto \log \pi(\boldsymbol{\psi}|\boldsymbol{y}, \boldsymbol{\theta}) + \log \pi(\boldsymbol{\theta}|\boldsymbol{y}) = \ell_n(\boldsymbol{\psi}) + \ell_n(\boldsymbol{\theta}).$$

For our first step, we show the approximations of $\ell_n(\boldsymbol{\psi})$ and $\ell_n(\boldsymbol{\theta})$ by $\tilde{\ell}_n(\boldsymbol{\psi})$ and $\tilde{\ell}_n(\boldsymbol{\theta})$ which obtained by using INLA. Therefore (1) is approximated by

$$\tilde{\pi}(\boldsymbol{\psi}, \boldsymbol{\theta}|\boldsymbol{y}) = \exp\{\tilde{\ell}_n(\boldsymbol{\psi}) + \tilde{\ell}_n(\boldsymbol{\theta})\}.$$

Result 1. *The approximate posterior distribution obtained by using INLA, $\tilde{\pi}(\boldsymbol{\psi}, \boldsymbol{\theta}|\boldsymbol{y})$, converges to a multivariate normal distribution as $n \longrightarrow \infty$.*

Precise statement of Result 1 and its proof are provided in the Appendix A. Result 1 states that the INLA-based estimators in a GLMM are asymptotically normally distributed. However, for our purpose we need two additional following results. Let

$$\log \pi^{(k)}(\boldsymbol{\psi}, \boldsymbol{\theta}|\boldsymbol{y}) \propto \log \pi^{(k)}(\boldsymbol{\psi}|\boldsymbol{y}, \boldsymbol{\theta}) + \log \pi^{(k)}(\boldsymbol{\theta}|\boldsymbol{y}) = \ell_n^{(k)}(\boldsymbol{\psi}) + \ell_n^{(k)}(\boldsymbol{\theta}).$$

Then, the DC-based density is given by

$$\pi^{(k)}(\boldsymbol{\psi}, \boldsymbol{\theta}|\boldsymbol{y}) = \exp\{\ell_n^{(k)}(\boldsymbol{\psi}) + \ell_n^{(k)}(\boldsymbol{\theta})\}.$$

Result 2. *The DC-based distribution, $\pi^{(k)}(\boldsymbol{\psi},\boldsymbol{\theta}|\boldsymbol{y})$, converges to a multivariate normal distribution as $k \longrightarrow \infty$.*

Precise statement of Result 2 and its proof are provided in the Appendix B. This result, recently, have also presented and proved by Lele *et al.* (2007) and Baghishani and Mohammadzadeh (2011). Combining Results 1 and 2, we can establish the asymptotic normality of the new HDC-based distribution. Let $\tilde{\ell}_n^{(k)}(\boldsymbol{\theta})$ and $\tilde{\ell}_n^{(k)}(\boldsymbol{\psi})$ be the corresponding approximates of $\ell_n^{(k)}(\boldsymbol{\theta})$ and $\ell_n^{(k)}(\boldsymbol{\psi})$ respectively obtained by INLA. Then,

$$\log \tilde{\pi}^{(k)}(\boldsymbol{\psi},\boldsymbol{\theta}|\boldsymbol{y}) \propto \tilde{\ell}_n^{(k)}(\boldsymbol{\theta}) + \tilde{\ell}_n^{(k)}(\boldsymbol{\psi}).$$

Now we can state the following result, which its precise statement and proof are provided in the Appendix C.

Result 3. *The HDC-based distribution, $\tilde{\pi}^{(k)}(\boldsymbol{\psi},\boldsymbol{\theta}|\boldsymbol{y})$, converges to a multivariate normal distribution as $k \longrightarrow \infty$.*

By having Result 3, we can develop a new HDC algorithm to carry out likelihood based inferences in GLMMs with high accuracy and low computational costs. This claim is illustrated in the following two sections through simulation studies and real examples.

4. Simulation Study

In this section, we present a simulation study designed to assess the performance of the HDC algorithm across a range of conditions that are realistic for clustered/longitudinal data. We also compare its performance to that of the adaptive Gauss-Hermit quadrature (AGHQ) method.

All analyses were conducted in R version 2.10.1 and on a Windows workstation using two Intel 2.53 GHz processors. The R packages *INLA* (`www.r-inla.org`) and *lme4* (`www.r-project.org`) were used for HDC and AGHQ based analyses, respectively. The R script for reproducing all results in the paper, can be downloaded from the `www.r-inla.org` web site in the Case Studies section.

The setting mimics a study with clustered count data where both continuous and dichotomous covariates influence the distribution of the response variable. In addition, there might be an overall heterogeneity between clusters. Namely, the data were generated according to the random intercept model:

$$\ln(\mu_{ij}) = \eta_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + u_i, \quad i = 1,\dots,m; j = 1,\dots,n_i,$$

where $u_i|\sigma^2 \overset{iid}{\sim} N(0, \sigma^2)$ with $\sigma = 0.75$ and $\boldsymbol{\beta} = (1, 2, -3)$. The covariate x_{1ij} was binary taking a value of 1 with probability 0.6. The covariate x_{2ij} was generated according to the uniform distribution, $U(0, 1)$.

We examined all combinations of three number of clones, k, (20, 40 and 80), three number of clusters, m, (20, 50 and 100) and two cluster sizes, n_i, (2 and 4), resulting in 18 conditions. For each condition, we simulated 100 data sets and estimated the model using HDC and AGHQ methods. For each condition, the same 100 data sets were analyzed by the two methods to enable accurate comparisons. For AGHQ method, 15 quadrature points were used.

To show that the proposed HDC based inferences are invariant to the choice of the priors, we used three different sets of prior distributions: a log-gamma prior for $\log(\sigma^{-2})$, an informative prior, with mean equal to 0.5 and variance equal to 0.25, a log-uniform prior for $\log(\sigma^{-2})$, a non-informative prior, and a log-gamma prior, a vague prior, with mean equal to 60 and variance equal to 1200. We also used the default INLA Gaussian priors for fixed effects, $\boldsymbol{\beta}$.

Tables 1–3 show the mean parameter estimates (Est.), their standard deviations (SD) and mean squared errors (MSE) over 100 simulated data sets. Due to space limitations, the results for $m = 20, 100$ and informative and non-informative priors are only reported in Tables 1–3. The full results are given in the online supplemental file. In addition, because the results for the AGHQ method are identical for different values of k, they are not shown in Tables 2 and 3. Here, HDC_1 and HDC_2 refer to the HDC method that uses informative and non-informative priors, respectively. According to the results, the estimates for regression parameters are captured very accurately under various conditions. Although, generally, the variance component, σ, is well estimated, it is slightly underestimated. Further, in most cases, when the number of clusters increases the estimate of variance component gets closer to the true value. The results obtained using AGHQ method are also quite close to the results obtained using HDC method with different priors. It is seen that, HDC method performs equivalently for small, medium and large k to estimate the parameters. A large difference in favor of selecting large k is seen in summarizing the precision of a variance component estimate by giving an approximate standard deviation. Those who use such standard deviations require that the distribution of the parameter estimate be symmetric or at least approximately symmetric. Douglas M. Bates has described that, in chapter drafts 1 and 2 of the book he is writing (http://lme4.R-forge.R-project.org/book/), the distribution

Table 1: Results from AGHQ and HDC estimations on 100 simulated data sets from the random intercept Poisson model with $k = 20$

				AGHQ			HDC_1			HDC_2		
m	n_i	Par.	True	Est.	SD	MSE	Est.	SD	MSE	Est.	SD	MSE
20	2	β_0	1	0.966	0.411	0.168	0.962	0.412	0.169	0.967	0.411	0.168
		β_1	2	1.987	0.364	0.132	1.988	0.365	0.132	1.987	0.364	0.132
		β_2	-3	-2.936	0.532	0.284	-2.939	0.531	0.283	-2.936	0.531	0.284
		σ	0.75	0.691	0.177	0.035	0.708	0.168	0.030	0.688	0.179	0.036
	4	β_0	1	1.026	0.249	0.062	1.024	0.250	0.062	1.026	0.249	0.062
		β_1	2	1.999	0.195	0.038	2.000	0.195	0.038	2.000	0.195	0.038
		β_2	-3	-3.007	0.254	0.064	-3.007	0.254	0.064	-3.007	0.254	0.064
		σ	0.75	0.702	0.129	0.019	0.710	0.126	0.017	0.699	0.129	0.019
100	2	β_0	1	0.985	0.168	0.028	0.986	0.167	0.028	0.986	0.168	0.028
		β_1	2	2.004	0.151	0.023	2.003	0.151	0.022	2.003	0.151	0.023
		β_2	-3	-2.987	0.205	0.042	-2.986	0.205	0.042	-2.986	0.205	0.042
		σ	0.75	0.743	0.073	0.005	0.739	0.072	0.005	0.736	0.073	0.005
	4	β_0	1	1.007	0.129	0.017	1.008	0.129	0.017	1.008	0.129	0.017
		β_1	2	2.000	0.088	0.008	2.000	0.088	0.008	2.000	0.088	0.008
		β_2	-3	-2.997	0.128	0.016	-2.997	0.128	0.016	-2.997	0.128	0.016
		σ	0.75	0.737	0.071	0.005	0.732	0.070	0.005	0.729	0.070	0.005

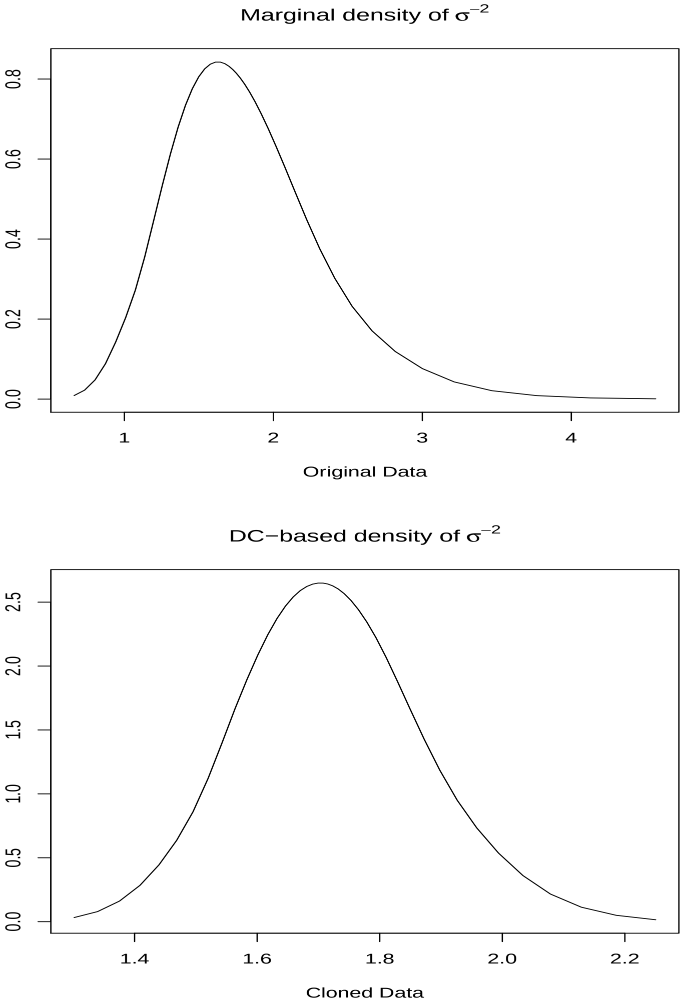


Figure 2: Marginal density (top panel) and HDC-based density (bottom panel) of precision parameter of random effect in the simulated model with $m = 50$, $n_i = 2$ and $k = 10$.

of the estimator of a variance component is more like a scaled chi-squared distribution which is asymmetric. It is nonsensical to believe that the distribution of variance estimators in much more complex situations should be well approximated by a normal distribution. Then, in most cases, using such approximate standard deviations is woefully inadequate.

However, following asymptotic results of Section 3, the distribution of the HDC-based estimator of a variance component is asymptotically normal and hence symmetric. Figure 2, as an example, reveals this fact. Therefore, the approximate standard deviations obtained using HDC method with a large enough k could be useful to form reliable statistical inferences for variance components.

For more explanation, Table 4 gives selected results of average standard deviations of the

Table 2: Results from AGHQ and HDC estimations on 100 simulated data sets from the random intercept Poisson model with $k = 40$

				HDC_1			HDC_2		
m	n_i	Par.	True	Est.	SD	MSE	Est.	SD	MSE
20	2	β_0	1	0.965	0.411	0.169	0.967	0.411	0.168
		β_1	2	1.987	0.365	0.132	1.987	0.364	0.132
		β_2	-3	-2.937	0.531	0.284	-2.936	0.532	0.284
		σ	0.75	0.696	0.171	0.032	0.686	0.178	0.035
	4	β_0	1	1.026	0.250	0.062	1.027	0.249	0.062
		β_1	2	2.000	0.195	0.038	2.000	0.195	0.038
		β_2	-3	-3.007	0.254	0.064	-3.007	0.254	0.064
		σ	0.75	0.703	0.126	0.018	0.697	0.128	0.019
100	2	β_0	1	0.986	0.168	0.028	0.986	0.168	0.028
		β_1	2	2.003	0.151	0.023	2.003	0.151	0.023
		β_2	-3	-2.986	0.205	0.042	-2.986	0.205	0.042
		σ	0.75	0.737	0.073	0.005	0.736	0.073	0.005
	4	β_0	1	1.008	0.129	0.017	1.008	0.129	0.017
		β_1	2	2.000	0.088	0.008	2.000	0.088	0.008
		β_2	-3	-2.997	0.128	0.016	-2.997	0.128	0.016
		σ	0.75	0.726	0.069	0.005	0.725	0.070	0.005

Table 3: Results from AGHQ and HDC estimations on 100 simulated data sets from the random intercept Poisson model with $k = 80$

				HDC_1			HDC_2		
m	n_i	Par.	True	Est.	SD	MSE	Est.	SD	MSE
20	2	β_0	1	0.967	0.412	0.169	0.968	0.411	0.168
		β_1	2	1.987	0.365	0.132	1.987	0.364	0.132
		β_2	-3	-2.936	0.531	0.284	-2.935	0.531	0.284
		σ	0.75	0.689	0.172	0.033	0.684	0.176	0.035
	4	β_0	1	1.027	0.249	0.062	1.027	0.249	0.062
		β_1	2	2.000	0.195	0.038	2.000	0.195	0.038
		β_2	-3	-3.007	0.254	0.064	-3.007	0.254	0.064
		σ	0.75	0.699	0.127	0.019	0.696	0.128	0.019
100	2	β_0	1	0.986	0.168	0.028	0.987	0.168	0.028
		β_1	2	2.003	0.151	0.023	2.003	0.151	0.023
		β_2	-3	-2.986	0.205	0.042	-2.986	0.205	0.042
		σ	0.75	0.735	0.074	0.006	0.734	0.074	0.006
	4	β_0	1	1.007	0.128	0.017	1.008	0.129	0.017
		β_1	2	2.000	0.088	0.007	2.000	0.088	0.008
		β_2	-3	-2.997	0.128	0.015	-2.997	0.128	0.016
		σ	0.75	0.726	0.069	0.005	0.725	0.070	0.005

Table 4: Average standard deviations of the variance component on 100 simulated data sets from the random intercept Poisson model with $n_i = 2$

k	m	AGHQ	HDC_1	HDC_2	HDC_3
20	20	0.184	0.159	0.160	0.157
	50	0.120	0.098	0.098	0.098
40	20	0.184	0.157	0.158	0.157
	50	0.120	0.098	0.097	0.099
80	20	0.184	0.156	0.156	0.156
	50	0.120	0.089	0.091	0.089

variance component across 100 simulated data sets. The results for HDC method were calculated using averaging $\sqrt{k}$ times approximated standard deviations of the variance component. According to our simulations, it is seen that there is a significant discrepancy between the results obtained using two different methods in favor of the HDC method. In addition, the standard deviations decrease slightly when k increases. Hence, by selection of a large enough k we can obtain a reliable estimate of precision of variance component. For regression parameters the results obtained using AGHQ and HDC methods are in agreement.

Another advantage of the HDC method is its implementation. In our simulation study, we considered a random intercept model for which the AGHQ method could be implemented. However, the AGHQ method is impractical, e.g. using R package *lme4*, for crossed designs. For the AGHQ method with 15 quadrature points, evaluation of the deviance at a single parameter value would require 15^r individual GLM deviance evaluations, where r is the total number of random effects. Even to evaluate the deviance for the GLMM at the starting values of the parameters could take several days and that is before the optimization with respect to the parameter values begins. But, for a crossed random effects model, the HDC method could be implemented easily. All computations required by the HDC method are efficiently performed by the R *INLA* package.

The computing time for HDC method on a typical data set was 31 s for $k = 100$, while usual DC estimates for the same data set and the same k was computed in 2026 s. For DC method we used the one-block MCMC sampler described in chapter 4 in the book by Rue and Held (2005).

5. Real Data Examples

This section provides examples of applications of the proposed hybrid method. We consider three examples by which both nested (Subsection 5.1) and crossed (Subsections 5.2 and 5.3) random effects are introduced. These examples have been considered previously by Breslow and Clayton (1993) and Fong *et al.* (2010).

Fong *et al.* (2010) analyzed these data sets by using INLA method and gave a number of prescriptions for prior specification especially for variance components. They also noticed that sometimes specification of a prior for variance components is not straightforward. But DC-based results are invariant to the choice of the priors.

5.1. Overdispersion

This example concerns data on the proportion of seeds that germinated on each of $m = 21$ plates arranged according to a 2×2 factorial design with respect to seed variety and type of root extract (Crowder, 1978). The sampling model is $Y_i|\boldsymbol{\beta}, u_i \sim Bin(n_i, p_i)$ where, for plate i, y_i is the number of germinating seeds and n_i is the total number of seeds for $i = 1, \ldots, m$. To account for the extraneous between plate variability, Breslow and Clayton (1993) introduced plate-level random effects and then fitted two main effects and interaction models:

$$\begin{aligned} logit(p_i) &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i, \qquad (2) \\ logit(p_i) &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + u_i, \end{aligned}$$

where $u_i|\sigma^2 \overset{iid}{\sim} N(0, \sigma^2)$, and x_{1i}, x_{2i} represent the seed variety and type of root extract for plate i, respectively.

To show that the DC-based estimators are invariant to the choice of the priors, we used three different sets of prior distributions. Following Fong *et al.* (2010) for the main effects model, the first set includes $N(0, 10)$ for fixed effects and $Ga(0.5, 0.0164)$ for σ^{-2}. The second set includes $N(1, 100)$ for fixed effects and a flat prior for the log-precision, $\log(\sigma^{-2})$ and the third set includes $N(-2, 1)$, $N(-1, 1)$ for fixed effects and $Ga(0.01, 0.01)$, as a vague prior, for σ^{-2}. For the interaction model, the first set includes $N(1, 100)$ for fixed effects and $Ga(0.5, 0.0164)$ for σ^{-2}. The second set includes $N(-1, 1)$, $N(-2, 1)$, $N(-1, 1)$ for fixed effects and a flat prior for $\log(\sigma^{-2})$ and the third set includes $N(-2, 100)$, $N(2, 100)$, $N(1, 100)$ for fixed effects and $Ga(0.01, 0.01)$ for σ^{-2}.

Table 5: The proportion of seeds that germinated on each of 21 plates arranged according to a 2 by 2 factorial layout by seed (x_1) and type of root extract (x_2)

Original Data					Cloned Data				
r	n	x_1	x_2	$plate$	$r.k$	$n.k$	$x.k_1$	$x.k_2$	$plate.k$
10	39	0	0	1	10	39	0	0	1
23	62	0	0	2	23	62	0	0	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3	7	1	1	21	3	7	1	1	21
					10	39	0	0	22
					23	62	0	0	23
					⋮	⋮	⋮	⋮	⋮
					3	7	1	1	$21k$

To implement the method using R *INLA* package, we first prepared cloned data by duplicating the original data. Notice that the number of plates for cloned data, *plate.k*, is $m \times k = 21k$. Let $x.k_{1i}$ and $x.k_{2i}$ denote the new cloned covariates. Let also $r.k_i$ and $n.k_i$ denote the proportion and the total number of seeds for $i = 1, \ldots, 21k$ respectively. Table 5 shows how the cloned data set is provided. Fitting the model (2), say, is done by calling the *inla()* function:

```
> formula = r.k ~ x.k1+x.k2+f(plate.k,model="iid",param=c(.5,.0164))
> result = inla(formula,data=clone.data,family="binomial",Ntrials=n.k)
```

The option *param=c(.5,.0164)* specifies the parameters for the gamma prior for σ^{-2}. The *summary()* function is available to obtain summary results:

```
> summary(result)

Fixed effects:
                  mean         sd 0.025quant   0.5quant 0.975quant        kld
(Intercept) -0.3890820 0.01163724 -0.4119139 -0.3890906 -0.3662326 0.07416333
x.k1        -0.3452493 0.01491875 -0.3746143 -0.3452261 -0.3160662 0.03071876
x.k2         1.0290746 0.01449569  1.0008734  1.0290775  1.0572161 0.30432151


Random effects:
Name      Model         Max KLD
```

```
plate.k   IID model   0.00026


Model hyperparameters:
                      mean    sd      0.025quant 0.5quant 0.975quant
Precision for plate.k 11.7230  0.6342 10.4854    11.7246  12.9717


Expected number of effective parameters(std dev): 1653.33(36.31)
Number of equivalent replicates : 2.540
```

The variances estimates for the MLEs can be obtained as the variances estimates of DC-based estimators multiplied by k. For example, the square root of variance estimate for β_2 in model (2) is equal to $\sqrt{200} \times 0.0145 = 0.205$ The estimate and standard deviation of the variance parameter σ can be easily extracted by calling the *inla.hyper()* and *inla.expectation()* functions:

```
> result.hyperpar = inla.hyperpar(result)
> precision = result.hyperpar$marginals[[1]]
> m1 = inla.expectation(function(x) (1/x)^0.5, precision)
> m2 = inla.expectation(function(x) 1/x, precision)
> sd = sqrt(k*(m2 - m1^2))
> print(c(mean=m1, sd=sd))
     mean        sd
0.2922371 0.10823
```

Table 6 presents the results obtained by HDC method with $k = 200$ for three different prior sets in the main and interaction effects models respectively. These results are compared with MLEs obtained by Breslow and Clayton (1993) using Gaussian quadrature. There is surprisingly very close correspondence between the MLE and obtained results from HDC method for different priors. Furthermore, DC-based densities for different priors are indistinguishable. These findings are illustrated in Figure 3 for interaction effect and precision parameter of random effect in the interaction model. For other fixed effects the densities are very close for different priors.

To compare the accuracy and computational costs for HDC method and DC method, we generated 10000 samples from DC-based distribution by using a one block MCMC sampler. The great advantage of the HDC method is the high accuracy and low computational cost. We obtained the results for HDC method in the main effects model in less than 6 s, whereas the MCMC samples

Table 6: MLEs and HDC-based estimates with $k = 200$ for Seed data in the main and interaction effects model. [a] from Breslow and Clayton (1993). Standard deviations are in brackets.

Model	Par.	MLE[a]	HDC_1	HDC_2	HDC_3
Main	Intercept	-0.389 (0.166)	-0.389 (0.165)	-0.389 (0.164)	-0.389 (0.165)
	Seed	-0.347 (0.215)	-0.346 (0.211)	-0.345 (0.211)	-0.345 (0.211)
	Extract	1.029 (0.205)	1.029 (0.203)	1.029 (0.203)	1.029 (0.203)
	σ	0.295 (0.112)	0.292 (0.108)	0.292 (0.110)	0.292 (0.110)
Interaction	Intercept	-0.548 (0.167)	-0.548 (0.165)	-0.548 (0.165)	-0.548 (0.166)
	Seed	0.097 (0.278)	0.097 (0.276)	0.098 (0.276)	0.098 (0.276)
	Extract	1.337 (0.237)	1.337 (0.235)	1.337 (0.235)	1.337 (0.235)
	Interaction	-0.811 (0.385)	-0.810 (0.383)	-0.810 (0.382)	-0.810 (0.382)
	σ	0.236 (0.110)	0.234 (0.108)	0.233 (0.108)	0.234 (0.108)

for DC-based distribution required about 557 s. High accuracy of HDC method is shown in Figure 4 as well.

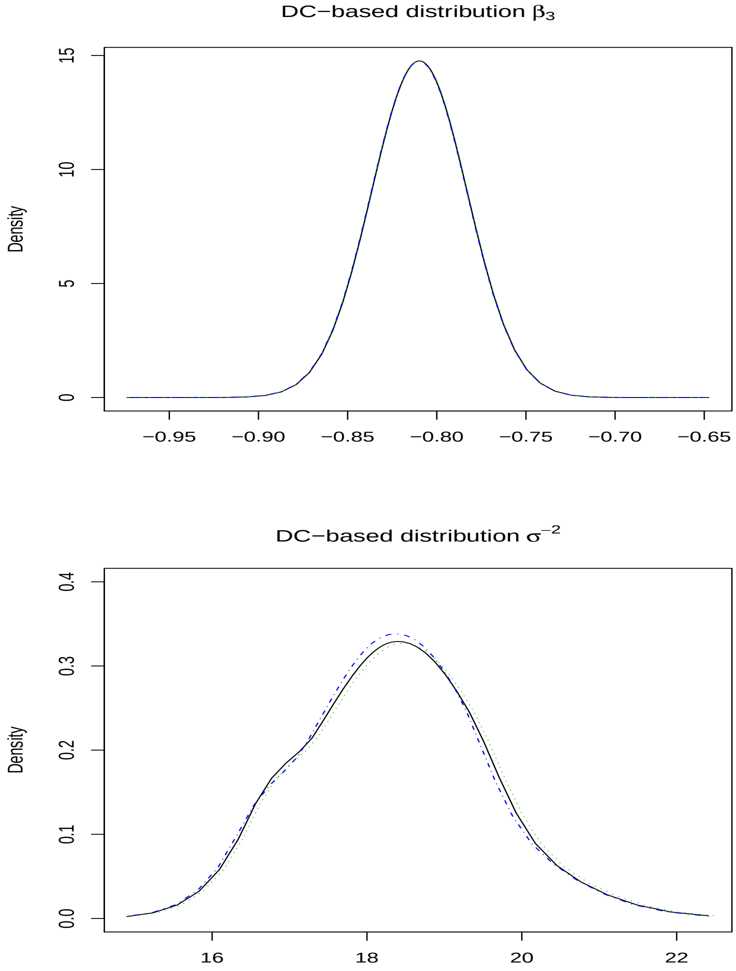
5.2. Longitudinal Data

Epilepsy data of Thall and Vail (1990) are a well known dataset that was analyzed several times by various authors. They presented data from a clinical trial of 59 epileptics who were randomized to a new drug (Trt=1) or a placebo (Trt=0). Baseline data available at entry into the trial included the number of epileptic seizures recorded in the preceding 8-weeks period and age in years. The logarithm of $\frac{1}{4}$ the number of baseline seizures (Base) and the logarithm of age (Age) were treated as covariables in the analysis. A multivariate response variable consisted of the counts of epileptic seizures, y_{ij}, for patient i during the 2-weeks before each of four clinic visits j (Visit, coded $-3, -1, +1, +3$), with $Y_{ij}|\boldsymbol{\beta}, u_i \overset{iid}{\sim} Po(\mu_{ij}), i = 1, \dots, 59; j = 1, \dots, 4$. An indicator of the fourth visit was also constructed to model to account its effect. Following Fong *et al.* (2010), we concentrate on the three random effects models fitted by Breslow and Clayton (1993):

$$\log(\mu_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + u_{1i}, \quad (3)$$

$$\log(\mu_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + u_{1i} + u_{0ij}, \quad (4)$$

$$\log(\mu_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + u_{1i} + u_{2i}V_j/10, \quad (5)$$


DC−based distribution β_3

DC−based distribution σ^{-2}

Figure 3: HDC-based densities of interaction effect (top panel) and precision parameter of random effect (bottom panel) in the interaction model with $k = 200$; The graphs showing the densities for the first (solid), second (dots), and third (dot-small dash) prior sets.

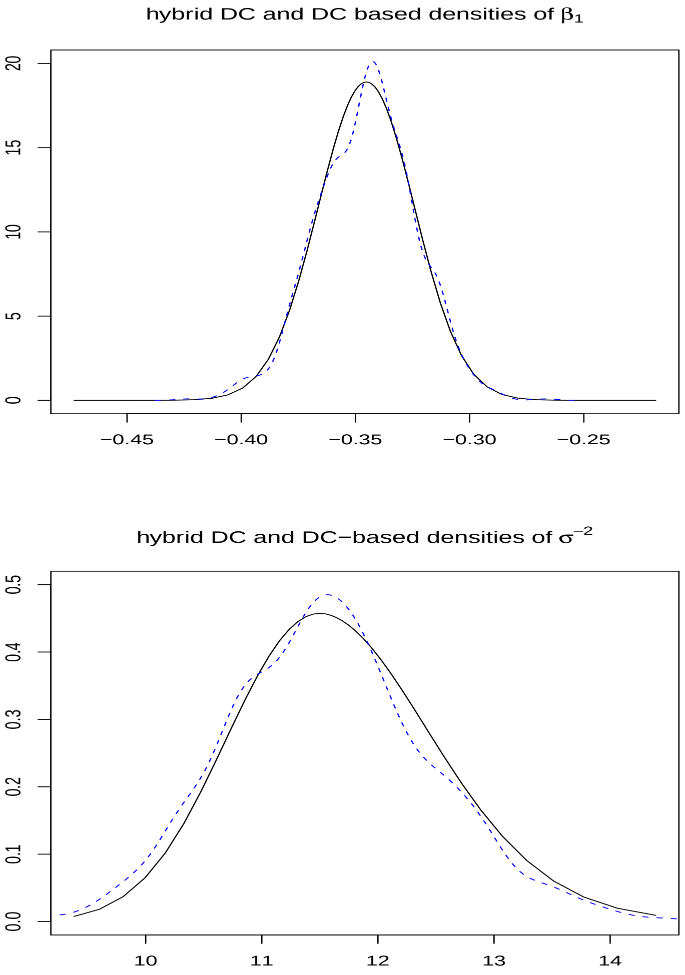


Figure 4: Comparison of HDC (solid) and DC-based (dashes) densities of seed variety effect (top panel) and precision parameter of random effect (bottom panel) in the main effects model with $k = 100$.

where $\boldsymbol{x}_{ij}$ is a 6×1 vector containing a 1 for intercept, the baseline by treatment interaction and above mentioned covariates and $\boldsymbol{\beta}$ is the associated fixed effect. All three models include patient-specific random effects $u_{1i} \sim N(0, \sigma_1^2)$, while in model (4) we introduce independent measurement errors $u_{0ij} \sim N(0, \sigma_0^2)$ and model (5) includes random effects on the slope associated with visit, u_{2i}, with

$$\begin{pmatrix} u_{1i} \\ u_{2i} \end{pmatrix} \sim N(\mathbf{0}, Q^{-1}).$$

According to Fong *et al.* (2010) we assume $Q \sim Wishart(r, T)$ with

$$T = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix}.$$

Here, similar to the previous subsection, we used three different sets of prior distributions. The first set for three models are priors considered by Fong *et al.* (2010). The second and third sets for the first model include $N(1, 100)$ for $\boldsymbol{\beta}$, a flat prior for $\log(\sigma_1^{-2})$ and $N(0, 10)$, $N(-1, 10)$, $N(-2, 100)$, $N(2, 100)$, $N(0, 10)$, $N(1, 100)$ for $\boldsymbol{\beta}$ and $Ga(0.05, 0.02)$ for σ_1^{-2} respectively. For the second model the second prior set includes $N(1, 100)$ for $\boldsymbol{\beta}$ and flat priors for $\log(\sigma_1^{-2})$ and $\log(\sigma_0^{-2})$. The third set includes $N(0, 10)$, $N(-1, 10)$, $N(-2, 100)$, $N(2, 100)$, $N(0, 10)$, $N(1, 100)$ for $\boldsymbol{\beta}$ and $Ga(0.05, 0.02)$ and $Ga(0.01, 0.01)$ for σ_1^{-2} and σ_0^{-2} respectively. Finally, for the third model the second and third prior sets include the same priors for fixed effects as second model but they include $r = 4$ and $T = diag(3, 4)$ and $r = 6$ and $T = diag(0.5, 0.5)$ respectively.

Table 7 presents the results obtained by HDC method with $k = 100$. The results are compared with AGHQ and INLA results. It is clear that the results are indistinguishable for different priors. Note that $\sigma_2^2 = T_{22}^{-1}$.

Figures 5 and 6 show the DC-based densities of fixed effects and precision parameters of random effects for third model (5) respectively. According to the figures, all three HDC-based densities for different prior sets are coincided. The results remain the same for two other models as well.

5.3. Crossed Random Effects: The Salamander Data

McCullagh and Nelder (1989) described an interesting dataset on the success of matings between male and female salamander of two population types, roughbutts (RB) and whitesides (WS). The experimental design involves three experiments having multiple pairings, with each salamander being involved in multiple matings, so that crossed random effects are required. The first

Table 7: HDC-based estimates obtained using $k = 100$ for Epilepsy data in the models (3)-(5) compared with AGHQ and INLA estimates. Standard deviations are in brackets.

Model	Par.	AGHQ	INLA	HDC_1	HDC_2	HDC_3
(3)	Intercept	-1.34 (1.18)	-1.30 (1.30)	-1.34 (1.18)	-1.34 (1.17)	-1.34 (1.17)
	Base	0.88 (0.13)	0.88 (0.15)	0.88 (0.13)	0.88 (0.13)	0.88 (0.13)
	Trt	-0.93 (0.40)	-0.94 (0.44)	-0.93 (0.40)	-0.93 (0.40)	-0.93 (0.40)
	Base×Trt	0.34 (0.20)	0.34 (0.22)	0.34 (0.20)	0.34 (0.20)	0.34 (0.20)
	Age	0.48 (0.35)	0.47 (0.38)	0.48 (0.34)	0.49 (0.34)	0.48 (0.34)
	V_4 or $V/10$	-0.16 (0.05)	-0.16 (0.05)	-0.16 (0.05)	-0.16 (0.05)	-0.16 (0.05)
	σ_1	0.50 (0.09)	0.50 (0.06)	0.50 (0.04)	0.50 (0.04)	0.50 (0.04)
(4)	Intercept	-1.41 (1.16)	-1.40 (1.32)	-1.41 (1.17)	-1.41 (1.18)	-1.41 (1.17)
	Base	0.88 (0.13)	0.88 (0.15)	0.88 (0.13)	0.88 (0.14)	0.88 (0.13)
	Trt	-0.95 (0.40)	-0.96 (0.45)	-0.95 (0.40)	-0.95 (0.40)	-0.95 (0.40)
	Base×Trt	0.35 (0.20)	0.35 (0.23)	0.35 (0.20)	0.35 (0.20)	0.35 (0.20)
	Age	0.49 (0.34)	0.48 (0.39)	0.48 (0.34)	0.49 (0.34)	0.48 (0.34)
	V_4 or $V/10$	-0.10 (0.09)	-0.10 (0.09)	-0.10 (0.09)	-0.10 (0.09)	-0.10 (0.09)
	σ_0	0.36 (0.08)	0.41 (0.04)	0.36 (0.04)	0.36 (0.04)	0.36 (0.04)
	σ_1	0.46 (0.09)	0.54 (0.06)	0.46 (0.06)	0.46 (0.06)	0.46 (0.06)
(5)	Intercept	-1.37 (1.17)	-1.36 (1.30)	-1.36 (1.19)	-1.38 (1.20)	-1.37 (1.19)
	Base	0.89 (0.13)	0.88 (0.14)	0.89 (0.13)	0.88 (0.13)	0.89 (0.13)
	Trt	-0.93 (0.40)	-0.94 (0.44)	-0.93 (0.40)	-0.93 (0.40)	-0.93 (0.40)
	Base×Trt	0.34 (0.20)	0.34 (0.22)	0.34 (0.20)	0.34 (0.20)	0.34 (0.20)
	Age	0.48 (0.35)	0.47 (0.38)	0.47 (0.35)	0.48 (0.35)	0.48 (0.35)
	V_4 or $V/10$	-0.27 (0.16)	-0.27 (0.16)	-0.26 (0.17)	-0.27 (0.17)	-0.27 (0.16)
	σ_1	0.50 (0.09)	0.56 (0.08)	0.50 (0.06)	0.50 (0.06)	0.50 (0.06)
	σ_2	0.73 (0.11)	0.70 (0.14)	0.73 (0.15)	0.73 (0.15)	0.72 (0.15)

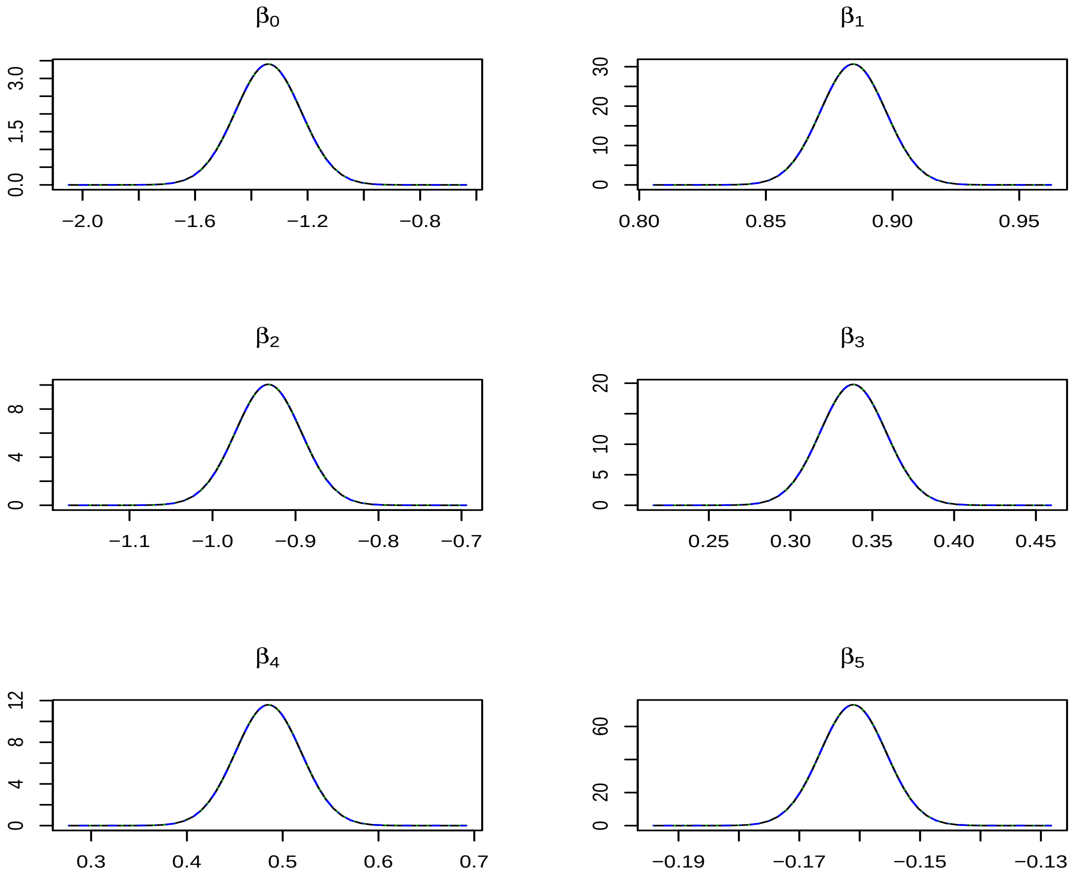


Figure 5: HDC-based densities of fixed effects in the model (5) with $k = 100$; The graphs showing the densities for the first (solid), second (dots), and third (dot-dash) prior sets.

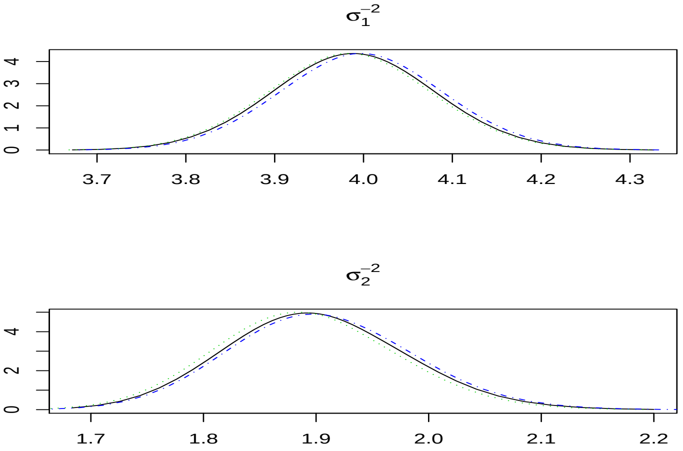


Figure 6: HDC-based densities of precision parameters of random effects in the model (5) with $k = 100$; The graphs showing the densities for the first (solid), second (dots), and third (dot-dash) prior sets.

experiment conducted during the summer of 1986 and the second and third conducted in the fall. Each experiment involved 30 matings of each of the four gender-population combinations. There are 360 binary responses in total. This complex data is reanalyzed by several authors such as Karim and Zeger (1992), Breslow and Clayton (1993), Bellio and Varin (2005) and Fong *et al.* (2010).

Suppose y_{ijk} be the binary response for female i and male j in experiment k. Here, we focus on model that was considered by Fong *et al.* (2010):

$$logitPr(Y_{ijk} = 1|\boldsymbol{\beta}, u^f_{ik}, u^m_{jk}) = \boldsymbol{x}^T_{ijk}\boldsymbol{\beta}_k + u^f_{ik} + u^m_{jk},$$

where $\boldsymbol{x}_{ijk}$ is a 4×1 vector representing the intercept, an indicator WS_f of whiteside females, an indicator WS_m of whiteside males and their interaction and $\boldsymbol{\beta}_k$ is the corresponding fixed effect. As Fong *et al.* (2010) have mentioned this model allows the fixed effects to vary by experiment and the model contains six random effects

$$u^f_{ik} \overset{iid}{\sim} N(o, \sigma^2_{fk}), u^m_{ik} \overset{iid}{\sim} N(o, \sigma^2_{mk}), \quad k = 1, 2, 3$$

one for each of males and females, and in each experiment.

Again similar to the previous subsections, we used three different sets of prior distributions and the first set presents priors considered by Fong *et al.* (2010). The second set include $N(0, 10)$ for

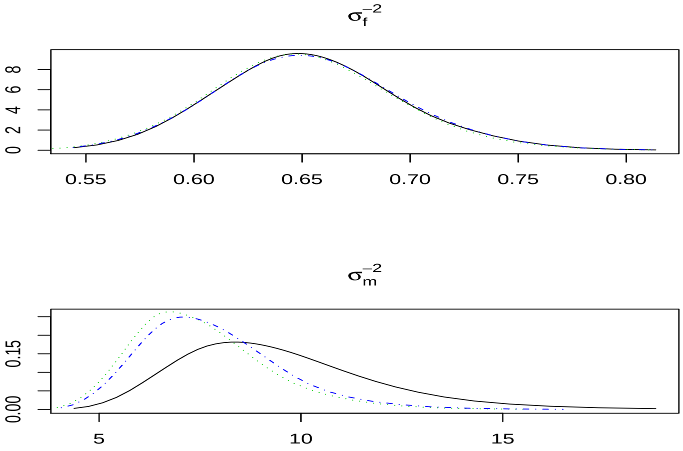


Figure 7: HDC-based densities of precision parameters of random effects obtained for summer experiment with $k = 100$; The graphs showing the densities for the first (solid), second (dots), and third (dot-dash) prior sets.

$\boldsymbol{\beta}$ and flat priors for both $\log(\sigma_{fk}^{-2})$ and $\log(\sigma_{mk}^{-2})$. The third set also include $N(0, 10)$ for $\boldsymbol{\beta}$ and $Ga(0.1, 0.1)$ for both σ_{fk}^{-2} and σ_{mk}^{-2}.

Table 8 shows the results obtained by HDC method with $k = 100$. The results are compared with Laplace approximation (LA) and INLA results. As one can see the results are indistinguishable for different priors. The results are also close to LA but, in some cases, there are some differences between their standard deviations. Also there is strong discrepancy between HDC and INLA estimates, usually with slightly larger standard deviations under the latter.

Figures 7–9 show the DC-based densities of precision parameters of random effects obtained for three experiments. According to the figures, in most cases, there are a good matching between curves.

We also compared the computational costs for HDC and DC methods by using 10000 generated samples from DC-based distribution. The computing time on summer experiment data for HDC method with $k = 100$ was about 65 s, whereas for DC method was about 5331 s.

6. Discussion

Although Fong *et al.* (2010) gave a number of prescriptions for prior specification especially for variance components in a GLMM, but sometimes specification of a prior in these models is not straightforward. On the other side, the DC-based inferences are invariant to the choice of the priors.

Table 8: HDC-based estimates obtained using $k = 100$ for Salamander data for summer and fall experiments compared with LA and INLA estimates. Standard deviations are in brackets.

Model	Par.	LA	INLA	HDC_1	HDC_2	HDC_3
Summer	Intercept	1.34 (0.62)	1.48 (0.72)	1.34 (0.62)	1.32 (0.62)	1.34 (0.62)
	WS_f	-2.94 (0.88)	-3.26 (1.01)	-2.94 (0.89)	-2.91 (0.90)	-2.94 (0.90)
	WS_m	-0.42 (0.63)	-0.50 (0.73)	-0.43 (0.64)	-0.42 (0.63)	-0.43 (0.64)
	$WS_f \times WS_m$	3.18 (0.94)	3.52 (1.03)	3.17 (0.94)	3.14 (0.96)	3.18 (0.96)
	σ_{f1}	1.25 (0.10)	1.29 (0.46)	1.24 (0.38)	1.24 (0.38)	1.23 (0.38)
	σ_{m1}	0.27 (0.05)	0.78 (0.29)	0.34 (0.32)	0.38 (0.31)	0.37 (0.26)
First Fall	Intercept	0.57 (0.67)	0.56 (0.71)	0.54 (0.64)	0.55 (0.65)	0.55 (0.65)
	WS_f	-2.46 (0.93)	-2.51 (1.01)	-2.37 (0.94)	-2.39 (0.94)	-2.40 (0.95)
	WS_m	-0.77 (0.72)	-0.75 (0.75)	-0.72 (0.69)	-0.73 (0.70)	-0.73 (0.70)
	$WS_f \times WS_m$	3.71 (0.96)	3.74 (1.03)	3.55 (1.01)	3.58 (1.01)	3.60 (1.03)
	σ_{f2}	1.35 (0.11)	1.38 (0.50)	1.30 (0.42)	1.30 (0.42)	1.31 (0.42)
	σ_{m2}	0.96 (0.09)	1.00 (0.36)	0.90 (0.38)	0.91 (0.39)	0.91 (0.40)
Second Fall	Intercept	1.02 (0.65)	1.07 (0.73)	0.99 (0.64)	0.93 (0.64)	1.00 (0.64)
	WS_f	-3.23 (0.83)	-3.39 (0.92)	-3.16 (0.85)	-3.15 (0.86)	-3.18 (0.80)
	WS_m	-0.82 (0.86)	-0.85 (0.94)	-0.79 (0.85)	-0.75 (0.85)	-0.79 (0.85)
	$WS_f \times WS_m$	3.82 (0.99)	4.02 (1.05)	3.74 (1.03)	3.75 (1.04)	3.76 (0.97)
	σ_{f3}	0.59 (0.07)	0.81 (0.28)	0.55 (0.40)	0.54 (0.42)	0.54 (0.43)
	σ_{m3}	1.36 (0.11)	1.47 (0.48)	1.33 (0.39)	1.33 (0.43)	1.33 (0.42)

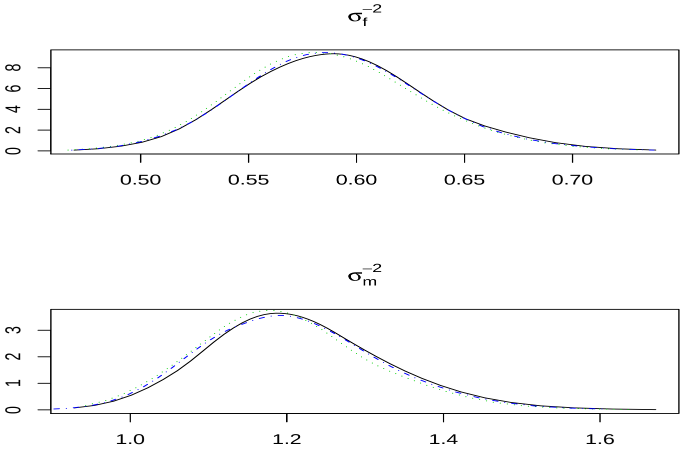


Figure 8: HDC-based densities of precision parameters of random effects obtained for first fall experiment with $k = 100$; The graphs showing the densities for the first (solid), second (dots), and third (dot-dash) prior sets.

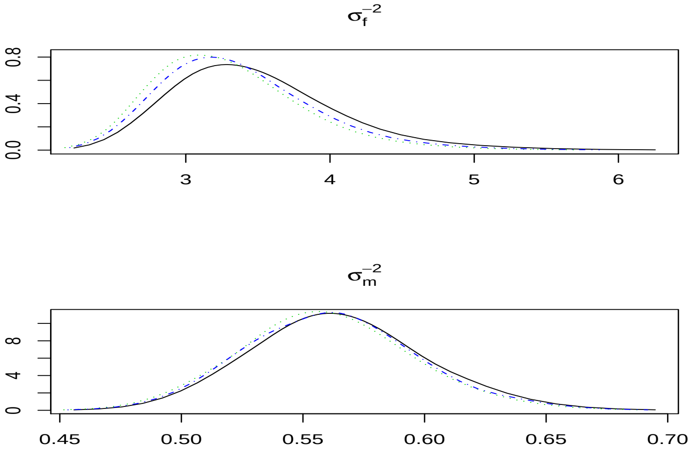


Figure 9: HDC-based densities of precision parameters of random effects obtained for second fall experiment with $k = 100$; The graphs showing the densities for the first (solid), second (dots), and third (dot-dash) prior sets.

Computation and convergence, however, are issues since the usual implementation of DC method is via MCMC. On the other hand, INLA provides precise estimates in seconds and minutes, even for models involving thousands of variables, in situations where any MCMC computation typically takes hours or even days. In this paper, we synthesized these two approaches and introduced a new HDC method so that its performance, according to the obtained results, is very good and inherits invariance property of DC method as well.

The benefits of our proposed method are the simplicity of implementation using R INLA package and to obtain MLE efficiently. The most available alternative methods to compute MLE in GLMMs, especially in models with crossed random effects, have disadvantages in the sense of consistent estimation, loss of efficiency, computational time required and convergence assessment, e.g. penalized quasi likelihood (Breslow and Clayton, 1993), composite likelihood (Bellio and Varin, 2005) and Monte Carlo expectation maximization (Booth *et al.*, 2001).

A disadvantage of our work is that, according to INLA methodology, the prior distributions for fixed effects $\boldsymbol{\beta}$ must be Gaussian. However, we can use Gaussian priors with high variances to consider approximately flat priors and the results, theoretically, are invariant to the choice of the priors as well. Alongside good performance of DC method, Baghishani and Mohammadzadeh (2011) have mentioned some of its limitations and their possible solutions. Although we did not discussed about selecting the number of clones, k, but, in general, selecting k under different circumstances such as sample size, random effects dimension and number of parameters needs further research.

In this paper, we assumed that the dimension of the random effects is fixed. But in some frameworks such as spatial models and spline smoothing models, the number of random effects (spline basis) increases with the sample size. Exploring to extend the HDC method to such frameworks would be interesting and is a topic of further research.

Based on INLA methodology, we need to assume a Gaussian distribution for random effects. However, it is usual to assume other distributions for the random effects such as (penalized) mixture of normals (Komarek and Lesaffre, 2008) or skew normal distribution (Hosseini *et al.*, 2011). It might be possible to synthesize DC method with other approximate Bayesian methods like the method proposed by Eidsvik *et al.* (2009) to be able to use other distributions for random effects, which needs further research in itself.

Finally, as Ponciano *et al.* (2009) have noticed, nowadays, the choice between Bayesian and frequentist approaches in GLMMs is no longer a matter of feasibility but rather can be based on the philosophical views of researchers.

Acknowledgment

We thank the associate editor and three anonymous referees for their valuable comments and suggestions which led to a considerably improvement of the paper. The first author wishes to thank the NTNU Department of Mathematical Sciences, Norway, for their support during his visit there. The third author would like to thank partial support of the Ordered and Spatial Data Center of Excellence of Ferdowsi University of Mashhad.

A. Asymptotic Normality of Approximate Posterior Distribution

A formal statement of Result 1, stated in Section 3, is given and proved in this section. The necessary lemmas for the following proofs are given in the supplemental materials. We first introduce some notation and calculations.

Assume that the functions $\ell_n(\boldsymbol{\theta})$ and $\tilde{\ell}_n(\boldsymbol{\theta})$ as well as $\ell_n(\boldsymbol{\psi})$ and $\tilde{\ell}_n(\boldsymbol{\psi})$ are twice continuously differentiable with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ respectively. Let $\nabla\ell_n(\boldsymbol{\theta})$, $\nabla\tilde{\ell}_n(\boldsymbol{\theta})$, $\nabla\ell_n(\boldsymbol{\psi})$ and $\nabla\tilde{\ell}_n(\boldsymbol{\psi})$ be the vectors of first-order partial derivatives with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ respectively. Furthermore, let $\nabla^2\ell_n(\boldsymbol{\theta})$, $\nabla^2\tilde{\ell}_n(\boldsymbol{\theta})$, $\nabla^2\ell_n(\boldsymbol{\psi})$ and $\nabla^2\tilde{\ell}_n(\boldsymbol{\psi})$ be the matrices of second-order partial derivatives with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ respectively. Here and subsequently, let $\hat{\boldsymbol{\psi}}_n$ be the mode of $\ell_n(\boldsymbol{\psi})$, satisfying $\nabla\ell_n(\boldsymbol{\psi}) = 0$ and $\hat{\boldsymbol{\theta}}_n$ be the mode of $\ell_n(\boldsymbol{\theta})$, satisfying $\nabla\ell_n(\boldsymbol{\theta}) = 0$.

To facilitate asymptotic theory arguments, whenever the $\hat{\boldsymbol{\psi}}_n$ and $\hat{\boldsymbol{\theta}}_n$ exist and $-\nabla^2\ell_n(\boldsymbol{\psi})$, $-\nabla^2\tilde{\ell}_n(\boldsymbol{\psi})$, $-\nabla^2\ell_n(\boldsymbol{\theta})$ and $-\nabla^2\tilde{\ell}_n(\boldsymbol{\theta})$ are positive definite, we define F_n, G_n, Q_n, V_n, $\boldsymbol{z}_n$, $\boldsymbol{w}_n$, $\boldsymbol{x}_n$ and $\boldsymbol{s}_n$ as follow:

$$\begin{aligned}
F_n^T F_n &= -\nabla^2\ell_n(\hat{\boldsymbol{\psi}}_n), \;\; \boldsymbol{z}_n = F_n(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}_n), \\
Q_n^T Q_n &= -\nabla^2\tilde{\ell}_n(\hat{\boldsymbol{\psi}}_n), \;\; \boldsymbol{x}_n = Q_n(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}_n), \\
G_n^T G_n &= -\nabla^2\ell_n(\hat{\boldsymbol{\theta}}_n), \;\; \boldsymbol{w}_n = G_n(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n), \\
V_n^T V_n &= -\nabla^2\tilde{\ell}_n(\hat{\boldsymbol{\theta}}_n), \;\; \boldsymbol{s}_n = V_n(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n).
\end{aligned}$$

Then the joint approximate posterior density of $(\boldsymbol{x}_n, \boldsymbol{s}_n)$ is given by

$$\tilde{\pi}_n(\boldsymbol{x}_n, \boldsymbol{s}_n|\boldsymbol{y}) \propto \tilde{\pi}_n(\boldsymbol{\psi}(\boldsymbol{x}_n), \boldsymbol{\theta}(\boldsymbol{s}_n)|\boldsymbol{y}) \propto e^{\tilde{\ell}_n(\boldsymbol{\psi})-\tilde{\ell}_n(\hat{\boldsymbol{\psi}}_n)} e^{\tilde{\ell}_n(\boldsymbol{\theta})-\tilde{\ell}_n(\hat{\boldsymbol{\theta}}_n)}. \qquad (6)$$

Let $\boldsymbol{\theta}_0$ and $\boldsymbol{\psi}_0$ denote the true underlying parameters and the true realization of random effects respectively. Let also $\tilde{P}_n^c$ and $\tilde{E}_n^c$ denote the approximate conditional probability and expectation given data $\boldsymbol{y}$. In what follows, all probability statements are with respect to the true underlying probability distribution. Then we must show

$$\tilde{P}_n^c((\boldsymbol{x}_n^T, \boldsymbol{s}_n^T)^T \in B) \longrightarrow \Phi_{q+d}(B), \ \ as \ \ n \to \infty,$$

where B is any Borel set in $\Re^{q+d}$ and Φ_{q+d} is the standard $q+d$-variate Gaussian distribution.

To conduct the posterior distribution in a form suitable for Stein's Identity, we need the following calculations. For converting $\tilde{\ell}_n(\boldsymbol{\psi})$ into a form close to normal, we first take the Taylor's expansion

$$\ell_n(\boldsymbol{\psi}) = \ell_n(\hat{\boldsymbol{\psi}}_n) + \frac{1}{2}(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}_n)^T \nabla^2 \ell_n(\boldsymbol{\psi}^*)(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}_n) + R_n$$

where $\boldsymbol{\psi}^*$ lies between $\boldsymbol{\psi}$ and $\hat{\boldsymbol{\psi}}_n$.

Now by Remark 1 in the supplement we have,

$$\tilde{\ell}_n(\boldsymbol{\psi}) = \tilde{\ell}_n(\hat{\boldsymbol{\psi}}_n) + \frac{1}{2}(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}_n)^T \nabla^2 \tilde{\ell}_n(\boldsymbol{\psi}^*)(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}_n) + R_n^{'},$$

where

$$\begin{aligned} R_n^{'} = R_n \ \ &+ \ \ [\ell_n(\hat{\boldsymbol{\psi}}_n) - \tilde{\ell}_n(\hat{\boldsymbol{\psi}}_n)] + [\tilde{\ell}_n(\boldsymbol{\psi}) - \ell_n(\boldsymbol{\psi})] \\ &+ \ \ \frac{1}{2}(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}_n)^T [\nabla^2 \ell_n(\boldsymbol{\psi}^*) - \nabla^2 \tilde{\ell}_n(\boldsymbol{\psi}^*)](\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}_n). \end{aligned}$$

Let

$$k_n(\boldsymbol{\psi}) = -\frac{1}{2}(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}_n)^T [\nabla^2 \tilde{\ell}_n(\hat{\boldsymbol{\psi}}_n) - \nabla^2 \tilde{\ell}_n(\boldsymbol{\psi}^*)](\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}_n).$$

Thus,

$$\tilde{\ell}_n(\boldsymbol{\psi}) \approx \tilde{\ell}_n(\hat{\boldsymbol{\psi}}_n) - \frac{1}{2}\|\boldsymbol{x}_n\|^2 + k_n(\boldsymbol{\psi}). \qquad (7)$$

With parallel arguments, we have

$$\begin{aligned} \tilde{\ell}_n(\boldsymbol{\theta}) \ &\approx \ \tilde{\ell}_n(\hat{\boldsymbol{\theta}}_n) + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)^T \nabla^2 \tilde{\ell}_n(\boldsymbol{\theta}^*)(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n), \\ l_n(\boldsymbol{\theta}) \ &= \ -\frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)^T [\nabla^2 \tilde{\ell}_n(\hat{\boldsymbol{\theta}}_n) - \nabla^2 \tilde{\ell}_n(\boldsymbol{\theta}^*)](\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n), \\ \tilde{\ell}_n(\boldsymbol{\theta}) \ &\approx \ \tilde{\ell}_n(\hat{\boldsymbol{\theta}}_n) - \frac{1}{2}\|\boldsymbol{s}_n\|^2 + l_n(\boldsymbol{\theta}), \qquad (8) \end{aligned}$$

where $\boldsymbol{\theta}^*$ lies between $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}_n$. Therefore we can rewrite (6) as

$$\tilde{\pi}_n(\boldsymbol{x}_n, \boldsymbol{s}_n|\boldsymbol{y}) \propto \phi_q(\boldsymbol{x}_n)\phi_d(\boldsymbol{s}_n)f_n(\boldsymbol{x}_n, \boldsymbol{s}_n), \tag{9}$$

where $f_n(\boldsymbol{x}_n, \boldsymbol{s}_n) = \exp\{k_n(\boldsymbol{\psi}) + l_n(\boldsymbol{\theta})\}$ and $\phi_t(\cdot)$ display the standard t-variate Gaussian density.

Suppose $\nabla_{\boldsymbol{x}_n} f(\boldsymbol{x}_n, \boldsymbol{s}_n)$ and $\nabla_{\boldsymbol{s}_n} f_n(\boldsymbol{x}_n, \boldsymbol{s}_n)$ denote the partial derivatives of $f_n(\boldsymbol{x}_n, \boldsymbol{s}_n)$ with respect to $\boldsymbol{x}_n$ and $\boldsymbol{s}_n$ respectively. Hence,

$$\begin{aligned} \frac{\nabla_{\boldsymbol{x}_n} f_n(\boldsymbol{x}_n, \boldsymbol{s}_n)}{f_n(\boldsymbol{x}_n, \boldsymbol{s}_n)} &= (Q_n^T)^{-1}\nabla k_n(\boldsymbol{\psi}), & (10)\\ \frac{\nabla_{\boldsymbol{s}_n} f_n(\boldsymbol{x}_n, \boldsymbol{s}_n)}{f_n(\boldsymbol{x}_n, \boldsymbol{s}_n)} &= (V_n^T)^{-1}\nabla l_n(\boldsymbol{\theta}). & (11) \end{aligned}$$

Let also $D = D_1 \cup D_2$, where $D_1 = \{\nabla\tilde{\ell}_n(\hat{\boldsymbol{\psi}}_n) = 0, -\nabla^2\tilde{\ell}_n(\hat{\boldsymbol{\psi}}_n) > 0\}$ and $D_2 = \{\nabla\tilde{\ell}_n(\hat{\boldsymbol{\theta}}_n) = 0, -\nabla^2\tilde{\ell}_n(\hat{\boldsymbol{\theta}}_n) > 0\}$. Here $A > 0$ means that the matrix A is positive definite.

A.1. Derivation and Proof of Result 1

Let

$$S = \left\{(\boldsymbol{x}_n, \boldsymbol{s}_n) : \boldsymbol{x}_n = Q_n(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}_n), \boldsymbol{s}_n = V_n(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n); \boldsymbol{\psi} \in N_1, \boldsymbol{\theta} \in N_2\right\}, \tag{12}$$

where N_1 and N_2 is given in the supplement. The following theorem deals with the asymptotic normality of approximate posterior distribution. The regularity conditions $(B1)$-$(B4)$, $(C1)$-$(C4)$ and $(F1)$-$(F3)$ are given in the supplement.

Theorem 1. *Let $h \in H_s$, the prior $\pi(\boldsymbol{\theta})$ satisfies $(F1)$-$(F3)$, $\tilde{\ell}_n(\boldsymbol{\theta})$ satisfies $(B1)$-$(B4)$ and $\tilde{\ell}_n(\boldsymbol{\psi})$ satisfies $(C1)$-$(C4)$. Then, $\tilde{E}_n^c[h(\boldsymbol{x}_n, \boldsymbol{s}_n)] \xrightarrow{p} \Phi h$.*

Proof. Note that Uh and $\pi(\boldsymbol{\theta})$ are bounded by Lemma 3.1 of Weng and Tsai (2008) and $(F1)$-$(F2)$. From (10), (11) and Lemma 3 in the supplement, for a.e. on D, we have

$$\tilde{E}_n^c[h(\boldsymbol{x}_n, \boldsymbol{s}_n)] - \Phi h = \tilde{E}_{n,\boldsymbol{x}_n}^c + \tilde{E}_{n,\boldsymbol{s}_n}^c,$$

where

$$\begin{aligned} \tilde{E}_{n,\boldsymbol{x}_n}^c &= \tilde{E}_n^c\left\{(Uh_1(\boldsymbol{x}_n))^T(Q_n^T)^{-1}\nabla k_n(\boldsymbol{\psi})\right\} = I_{\boldsymbol{x}_n}, & (13)\\ \tilde{E}_{n,\boldsymbol{s}_n}^c &= \tilde{E}_n^c\left\{(Uh_2(\boldsymbol{s}_n))^T(V_n^T)^{-1}\nabla l_n(\boldsymbol{\theta})\right\} = I_{\boldsymbol{s}_n}. & (14) \end{aligned}$$

Since $P(D_2^c) \longrightarrow 0$ by $(B1)$ and $P(D_1^c) \longrightarrow 0$ by $(C1)$, it suffices to show $I_{\boldsymbol{s}_n} \xrightarrow{p} 0$ and $I_{\boldsymbol{x}_n} \xrightarrow{p} 0$.

From (14) we have,

$$I_{\boldsymbol{s}_n} = \frac{\int_S (Uh_2(\boldsymbol{s}_n))^T (V_n^T)^{-1} \nabla l_n(\boldsymbol{\theta}) \phi_d(\boldsymbol{s}_n)\phi_q(\boldsymbol{x}_n) f_n(\boldsymbol{s}_n, \boldsymbol{x}_n) d\boldsymbol{s}_n d\boldsymbol{x}_n}{\int_S \phi_d(\boldsymbol{s}_n)\phi_q(\boldsymbol{x}_n) f_n(\boldsymbol{s}_n, \boldsymbol{x}_n) d\boldsymbol{s}_n d\boldsymbol{x}_n}.$$

The denominator is bounded below by some $K_1 > 0$ by Lemma 5(D1) in the supplement. Then we just need to show that the numerator converges to 0 in probability. First we decompose the numerator into two integrals over $\|\boldsymbol{s}_n\| \leq b_{1n}$ and $\|\boldsymbol{s}_n\| > b_{1n}$ and call the corresponding integrals as $I_{\boldsymbol{s}_n,1}$ and $I_{\boldsymbol{s}_n,2}$ respectively. With respect to $(F1)$-$(F2)$, Lemma 3.1 of Weng and Tsai (2008), and

$$(V_n^T)^{-1} \nabla l_n(\boldsymbol{\theta}) = \left\{ I_d - (V_n^T)^{-1} [-(\frac{\partial^2 \tilde{\ell}_n}{\partial \theta_i \partial \theta_j}(\theta^{*ij}))] V_n^{-1} \right\} \boldsymbol{s}_n,$$

there exists a constant $C_1 > 0$ such that

$$\begin{aligned} |I_{\boldsymbol{s}_n,1}| &\leq \int_{\|\boldsymbol{s}_n\| \leq b_{1n}} |(Uh_2(\boldsymbol{s}_n))^T (V_n^T)^{-1} \nabla l_n(\boldsymbol{\theta})| e^{\tilde{\ell}_n(\boldsymbol{\theta}) - \tilde{\ell}_n(\hat{\boldsymbol{\theta}})} e^{\tilde{\ell}_n(\boldsymbol{\psi}) - \tilde{\ell}_n(\hat{\boldsymbol{\psi}})} d\boldsymbol{s}_n d\boldsymbol{x}_n \\ &\leq C_1 \sup_{\boldsymbol{\theta}: \|\boldsymbol{s}_n\| \leq b_{1n}} \|I_d - (V_n^T)^{-1} [-(\frac{\partial^2 \tilde{\ell}_n}{\partial \theta_i \partial \theta_j})(\theta^{*ij})] V_n^{-1}\| \\ &\times \int_{\|\boldsymbol{s}_n\| \leq b_{1n}} \|\boldsymbol{s}_n\| e^{\tilde{\ell}_n(\boldsymbol{\theta}) - \tilde{\ell}_n(\hat{\boldsymbol{\theta}})} e^{\tilde{\ell}_n(\boldsymbol{\psi}) - \tilde{\ell}_n(\hat{\boldsymbol{\psi}})} d\boldsymbol{s}_n d\boldsymbol{x}_n. \end{aligned}$$

Using $(B2)$ and Lemma 4 part 2 in the supplement, we conclude that $I_{\boldsymbol{s}_n,1} \xrightarrow{p} 0$. Next, by $(B3)$, $(F1)$-$(F2)$ and Lemma 3.1 of Weng and Tsai (2008), there exists a constant $C_2 > 0$ such that

$$|I_{\boldsymbol{s}_n,2}| \leq C_2 \int_{S \cap \{\|\boldsymbol{s}_n\| > b_{1n}\}} \|\boldsymbol{s}_n\|^{r_1} e^{\tilde{\ell}_n(\boldsymbol{\theta}) - \tilde{\ell}_n(\hat{\boldsymbol{\theta}})} e^{\tilde{\ell}_n(\boldsymbol{\psi}) - \tilde{\ell}_n(\hat{\boldsymbol{\psi}})} d\boldsymbol{s}_n d\boldsymbol{x}_n,$$

which using Lemma 4 part 2 in the supplement, converges to 0 in probability. Hence, $I_{\boldsymbol{s}_n} \xrightarrow{p} 0$.

Similarly, $I_{\boldsymbol{x}_n} \xrightarrow{p} 0$ follows from $(F1)$-$(F2)$, $(C2)$-$(C3)$, Lemma 4 part 2 in the supplement, Lemma 3.1 of Weng and Tsai (2008) and

$$(Q_n^T)^{-1} \nabla k_n(\boldsymbol{\psi}) = \left\{ I_q - (Q_n^T)^{-1} [-(\frac{\partial^2 \tilde{\ell}_n}{\partial \psi_i \partial \psi_j}(\psi^{*ij}))] Q_n^{-1} \right\} \boldsymbol{x}_n.$$

This completes the proof. □

B. Asymptotic Normality of DC-Based Distribution

We define $F_{n,k}$, $G_{n,k}$, $\boldsymbol{z}_{n,k}$ and $\boldsymbol{w}_{n,k}$ as follow:

$$\begin{aligned} F_{n,k}^T F_{n,k} &= -\nabla^2 \ell_n^{(k)}(\hat{\boldsymbol{\psi}}_n), \quad \boldsymbol{z}_{n,k} = F_{n,k}(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}_n), \\ G_{n,k}^T G_{n,k} &= -\nabla^2 \ell_n^{(k)}(\hat{\boldsymbol{\theta}}_n), \quad \boldsymbol{w}_{n,k} = G_{n,k}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n), \end{aligned}$$

Therefore,

$$\pi_n^{(k)}(\boldsymbol{z}_{n,k}, \boldsymbol{w}_{n,k}|\boldsymbol{y}) \propto \pi_n^{(k)}(\boldsymbol{\psi}(\boldsymbol{z}_{n,k}), \boldsymbol{\theta}(\boldsymbol{w}_{n,k})|\boldsymbol{y}) \propto e^{\ell_n^{(k)}(\boldsymbol{\psi})-\ell_n^{(k)}(\hat{\boldsymbol{\psi}}_n)}e^{\ell_n^{(k)}(\boldsymbol{\theta})-\ell_n^{(k)}(\hat{\boldsymbol{\theta}}_n)}.$$

Theorem 2 below shows the asymptotic distribution of the DC-based distribution is normal.

Theorem 2. *Let $h \in H_s$, the prior $\pi(\boldsymbol{\theta})$ satisfies (F1)-(F3), $\ell_n^{(k)}(\boldsymbol{\theta})$ and $\ell_n^{(k)}(\boldsymbol{\psi})$ satisfies appropriate conditions, similar to conditions (B1)-(B4) and (C1)-(C4) of Theorem 1 replacing n with nk. Then, $E_{nk}^c[h(\boldsymbol{z}_{n,k}, \boldsymbol{w}_{n,k})] \xrightarrow{p} \Phi h$ as $k \to \infty$.*

Proof. The proof is based on similar techniques of the proof of Theorem 1. □

C. Asymptotic Normality of Hybrid DC-Based Distribution

Lemma 2 in the supplement and the expressions (7) and (8) in Appendix A, illustrate why we need asymptotic normality of the HDC-based distribution to be able to use the INLA within DC. Combining obtained results of Theorems 1 and 2, we can establish the asymptotic normality of the HDC-based distribution. Let

$$\begin{aligned} Q_{n,k}^T Q_{n,k} &= -\nabla^2 \tilde{\ell}_n^{(k)}(\hat{\boldsymbol{\psi}}_n), \ \ \boldsymbol{x}_{n,k} = Q_{n,k}(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}_n), \qquad (15) \\ V_{n,k}^T V_{n,k} &= -\nabla^2 \tilde{\ell}_n^{(k)}(\hat{\boldsymbol{\theta}}_n), \ \ \boldsymbol{s}_{n,k} = V_{n,k}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n), \qquad (16) \end{aligned}$$

Now we can state the following theorem.

Theorem 3. *Let $h \in H_s$, the prior $\pi(\boldsymbol{\theta})$ satisfies (F1)-(F3), $\tilde{\ell}_n^{(k)}(\boldsymbol{\theta})$ and $\tilde{\ell}_n^{(k)}(\boldsymbol{\psi})$ satisfies appropriate conditions replacing n with nk. Then, $\tilde{E}_{nk}^c[h(\boldsymbol{x}_{n,k}, \boldsymbol{s}_{n,k})] \xrightarrow{p} \Phi h$ as $k \to \infty$.*

Proof. The proof follows by combining (15) and (16) with Theorem 1. □

References

Baghishani, H. & Mohammadzadeh, M. (2011). A data cloning algorithm for computing maximum likelihood in spatial generalized linear mixed models, *Computational Statistics and Data Analysis,* **55**, 1748-1759.

Bellio, R. & Varin, C. (2005). A pairwise likelihood approach to generalized linear models with crossed random effects, *Statistical Modelling,* **5**, 217-227.

Booth, J. G., Hobert, J. P. & Jank W. (2001). A survey of Monte Carlo algorithms for maximizing the likelihood of a two-stage hierarchical model, *Statistical Modelling* **1**, 333-349.

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

Crowder, M. J. (1978). Beta-binomial ANOVA for proportions, *Applied Statistics,* **27**, 24-37.

Eidsvik, J., Martino, S. & Rue, H. (2009). Approximate Bayesian inference in spatial generalized linear mixed models, *Scandinavian Journal of Statistics,* **36**, 1-22.

Fong, Y., Rue, H. & Wakefield, J. (2010). Bayesian inference for generalized linear mixed models, *Biostatistic,* **11**, 397-412.

Hosseini, F., Eidsvik, J. & Mohammadzadeh, M. (2011). Approximate Bayesian inference in spatial generalized linear mixed models with skew normal latent variables, *Computational Statistics and Data Analysis,* **55**, 1791-1806.

Karim, M. R., & Zeger, S. L. (1992). Generalized linear models with random effects: Salamander mating revisited, *Biometrics,* **48**, 631-644.

Komarek, A. & Lesaffre, E. (2008). Generalized linear mixed model with a penalized Gaussian mixture as a random effects distribution, *Computational Statistics and Data Analysis,* **52**, 3441-3458.

Lele, S. R., Dennis, B. & Lutscher, F. (2007). Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods, *Ecology Letters,* **10**, 551-563.

McCullagh, P. & Nelder, J. A. (1989). *Generalized linear models,* Chapman and Hall, London.

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

Ponciano, J. M., Taper, M. L., Dennis, B., & Lele, S. R. (2009). Hierarchical models in ecology: Confidence intervals, hypothesis testing, and model selection using data cloning, *Ecology,* **90**, 356-362.

Rue, H. & Held, L. (2005). *Gaussian Markov random fields: theory and applications,* Chapman & Hall/CRC, Boca Raton, London.

Rue, H. & Martino, S. (2007). Approximate Bayesian inference for hierarchical Gaussian Markov random fields models, *Journal of Statistical Planning and Inference,* **137**, 3177-3192.

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

Thall, P. F., & Vail, S. C. (1990). Some covariance models for longitudinal count data with overdispersion, *Biometrics,* **46**, 657-671.

Weng, R. C. & Tsai, W. C. (2008). Asymptotic posterior normality for multiparameter problems, *Journal of Statistical Planning and Inference,* **138**, 4068-4080.