Biometrika (2012), **99**, 1, *pp*. 1–13
© 2012 Biometrika Trust
Printed in Great Britain

Advance Access publication on 31 July 2012

On asymptotic validity of naive inference with an approximate likelihood

BY H. E. OGDEN

Mathematical Sciences, University of Southampton, Southampton SO17 1BJ, U.K.
h.e.ogden@soton.ac.uk

SUMMARY

Many statistical models have likelihoods which are intractable: it is impossible or too expensive to compute the likelihood exactly. In such settings, a common approach is to replace the likelihood with an approximation, and proceed with inference as if the approximate likelihood were the true likelihood. In this paper, we describe conditions which guarantee that this naive inference with an approximate likelihood has the same first-order asymptotic properties as that with the true likelihood. We investigate the implications of these results using a Laplace approximation to the likelihood in a simple two-level latent variable model, and using reduced-dependence approximations to the likelihood in an Ising model.

Some key words: Intractable likelihood, Ising model, Laplace approximation, Latent variable model

1. INTRODUCTION

For many models, it is impossible or infeasibly expensive to evaluate the likelihood function, typically because it involves a high-dimensional sum or integral. In such cases, a common approach is to find an approximation $\tilde{L}(\cdot)$ to the likelihood $L(\cdot)$, and to use $\tilde{L}(\cdot)$ in place of $L(\cdot)$ to conduct inference about the model parameters.

For instance, one could construct a point estimate by maximizing the approximate likelihood, and form confidence intervals based on the curvature of the approximated log-likelihood about its maximum. From a Bayesian perspective, an approximate posterior $\tilde{\pi}(\theta; y) \propto \tilde{L}(\theta; y)\pi(\theta)$ could be formed by replacing the true likelihood by the approximate likelihood.

Such an approach is commonly used in practice. In latent variable models, where the likelihood is an integral over the latent variables, naive inference using a Laplace approximation to the likelihood is used in both maximum likelihood (Pinheiro & Bates, 1995; Bates et al., 2015) and Bayesian (Rue et al., 2009) settings. In Markov random field models, where the likelihood involves an intractable normalizing constant, inference is often conducted by replacing the exact normalizing constant with an approximation (Friel et al., 2009; Tjelmeland & Austad, 2012).

In this paper, we provide conditions under which the approximate maximum likelihood estimator is consistent and has the same asymptotic normal distribution as the true maximum likelihood estimator, hypothesis tests based on the approximate likelihood remain valid, and in Bayesian analysis the distance between the approximate posterior and the exact posterior shrinks to zero.

Douc et al. (2004) show that the approximate maximum likelihood estimator will have the correct asymptotic normal distribution if the error in the log-likelihood, $\epsilon_n(\theta) = \log \tilde{L}_n(\theta) - \log L_n(\theta)$, tends in probability to zero as $n \to \infty$, uniformly in θ. We argue that this measure is often too strict, and give examples of situations in which $\epsilon_n(\theta)$ grows rapidly with n and yet the

inference remains asymptotically valid. Our conditions are based instead on $\nabla_\theta \epsilon_n(\theta)$, the error in the approximation to the score function.

We provide two examples to demonstrate how the conditions may be used in practice. The first is a simple two-level latent variable model, with m_n repeated observations for each of n items. We obtain the rate at which m_n must grow with n in order for the Laplace approximation to give asymptotically valid inference. If m_n grows with n at a slower rate, the estimator remains consistent, but loses efficiency relative to the true maximum likelihood estimator, and naively constructed confidence intervals have coverage lower than nominal.

The second example is an Ising model on an $m \times m$ lattice, with the class of reduced-dependence approximations (Friel et al., 2009) used to approximate the likelihood. The reduced-dependence approximation is controlled by a tuning parameter k, where larger values of k provide a more accurate approximation at a higher cost. We obtain the rate at which $k = k_m$ must grow with m to provide asymptotically valid inference, as $m \to \infty$. For parameter values associated with weak dependence, we show that the reduced-dependence approximation with a suitable choice of k_m may be used to obtain asymptotically valid inference at cost polynomial in m, in contrast with the exponential cost of computing the likelihood exactly.

2. ASYMPTOTIC VALIDITY OF APPROXIMATE LIKELIHOOD INFERENCE

2·1. *Setup and notation*

Consider a sequence of models indexed by n, with common parameter $\theta \in \Theta \subseteq \mathbb{R}^p$. Write $\ell_n(\theta; y)$ for the log-likelihood given observed data y under model n, $u_n(\theta; y) = \nabla_\theta \ell_n(\theta; y)$ for the corresponding score function and $J_n(\theta; y) = -\nabla_\theta^T \nabla_\theta \ell_n(\theta; y)$ for the observed information. For convenience we sometimes drop the data y from the notation. Suppose that the data were generated from the model for some $\theta_0 \in \Theta$, and that as $n \to \infty$, the amount of information provided by the data about the parameter grows at some rate r_n, so that there exists a positive-definite matrix $I(\theta)$ such that $\bar{J}_n(\theta) = r_n^{-1} J_n(\theta) \to I(\theta)$ in probability. This includes the case with n independent replications as a special case, with $r_n = n$, but we also wish to allow for more complex settings.

Let $\tilde{\ell}_n(\cdot; y)$ be an approximate log-likelihood, which in general may be any function of the parameter θ to be used in place of $\ell_n(\cdot; y)$. Our focus is on approximate likelihoods formed by replacing an intractable quantity, such as a high-dimensional integral or sum, with a numerical approximation to it. Write $\hat{\theta}_n$ and $\tilde{\theta}_n$ for the estimators maximizing $\ell_n(\theta)$ and $\tilde{\ell}_n(\theta)$ respectively. Suppose that $\ell_n(\theta)$ and $\tilde{\ell}_n(\theta)$ are both three times differentiable, and write $\tilde{u}_n(\theta) = \nabla_\theta \tilde{\ell}_n(\theta)$ and $\tilde{J}_n(\theta) = -\nabla_\theta^T \nabla_\theta \tilde{\ell}_n(\theta)$ for the approximate score and information.

Write $\epsilon_n(\theta) = \tilde{\ell}_n(\theta) - \ell_n(\theta)$ for the pointwise error in the log-likelihood, $\delta_n(\theta) = \|\nabla_\theta \epsilon_n(\theta)\| = \|\tilde{u}_n(\theta) - u_n(\theta)\|$ for the absolute error in the score, and $\gamma_n(\theta) = \|\nabla_\theta^T \nabla_\theta \epsilon_n(\theta)\| = \|J_n(\theta) - \tilde{J}_n(\theta)\|$ for the absolute error in the observed information matrix. For concreteness, we use the L_1 norms $\|a\| = \sum_i |a_i|$ for a vector a, and $\|A\| = \max_j \{\sum_i |A_{ij}|\}$ for a matrix A, although the same results hold for any choice of norms.

Write $\delta_n^\infty(S) = \sup_{\theta \in S} \delta_n(\theta)$ for the uniform error in the score over any set $S \subseteq \Theta$, and let $\delta_n^\infty = \delta_n^\infty(\Theta)$. Similarly, define $\gamma_n^\infty(S) = \sup_{\theta \in S} \gamma_n(\theta)$ and $\gamma_n^\infty = \gamma_n^\infty(\Theta)$. For any $\theta_0 \in \Theta$, write $B_t(\theta_0) = \{\theta \in \Theta : \|\theta - \theta_0\| \leq t\}$ for the ball of radius t about θ_0.

2·2. *Approximate maximum likelihood inference*

First, we describe sufficient conditions to ensure that $\tilde{\theta}_n$ is consistent. The proofs of all results are given in the Appendix.

We will assume some standard regularity conditions on the model. Writing $\bar{u}_n(\theta) = r_n^{-1}u_n(\theta)$, and $\bar{u}(\theta)$ for the limit as $n \to \infty$, we assume that $\sup_{\theta\in\Theta}\|\bar{u}_n(\theta) - \bar{u}(\theta)\| \to 0$ in probability. We assume $\bar{u}(\cdot)$ is such that, for any $\epsilon > 0$, $\int_{\theta\in\Theta:\|\theta-\theta_0\|\geq\epsilon}\|\bar{u}(\theta)\|d\theta > \bar{u}(\theta_0) = 0$. These conditions are stronger than necessary, and we expect the same result to hold in many other situations where the true maximum likelihood estimator is consistent.

THEOREM 1. *Suppose $\delta_n^\infty = o_p(r_n)$ as $n \to \infty$. Then $\tilde{\theta}_n \to \theta_0$ in probability, as $n \to \infty$.*

We now give conditions to ensure that $\tilde{\theta}_n$ retains the same limiting distribution as $\hat{\theta}_n$. Since $\hat{\theta}_n - \theta_0$ is $O_p(r_n^{-1/2})$, this is equivalent to finding conditions under which $\hat{\theta}_n - \tilde{\theta}_n$ is $o_p(r_n^{-1/2})$. The following lemma bounds the distance between $\tilde{\theta}_n$ and $\hat{\theta}_n$ in terms of the error in the score function near θ_0.

LEMMA 1. *Suppose that $\delta_n^\infty = o_p(r_n)$, and that there exists $t > 0$ such that $\delta_n^\infty\{B_t(\theta_0)\} = o_p(a_n)$. Then $\tilde{\theta}_n - \hat{\theta}_n = o_p(a_n r_n^{-1})$.*

Applying Lemma 1 with $a_n = r_n^{1/2}$ leads directly to the asymptotic normality result.

THEOREM 2. *Suppose that $\delta_n^\infty = o_p(r_n)$, and that there exists $t > 0$ such that $\delta_n^\infty\{B_t(\theta_0)\} = o_p(r_n^{1/2})$. Then $r_n^{1/2}(\tilde{\theta}_n - \theta_0) \to N\{0, I(\theta_0)^{-1}\}$ in distribution, as $n \to \infty$.*

It is also desirable for test statistics constructed by using the approximate likelihood in place of the true likelihood to have the correct asymptotic distribution.

Consider testing the hypothesis $H_0 : \theta \in \Theta_R$, where $\Theta_R \subset \Theta$ and $\dim(\Theta_R) = q$. The likelihood ratio test statistic is $\Lambda_n = 2\{\ell_n(\hat{\theta}_n) - \ell_n(\hat{\theta}_n^R)\}$ where $\hat{\theta}_n^R = \arg\max_{\theta\in\Theta_R}\{\ell_n(\theta)\}$, and the approximate likelihood ratio test statistic is $\tilde{\Lambda}_n = 2\{\tilde{\ell}_n(\tilde{\theta}_n) - \tilde{\ell}_n(\tilde{\theta}_n^R)\}$, where $\tilde{\theta}_n^R = \arg\max_{\theta\in\Theta_R}\{\tilde{\ell}_n(\theta)\}$.

Under the conditions used to show that $\tilde{\theta}_n$ has the correct limiting distribution, plus a bound on the error in the information around θ_0, $\tilde{\Lambda}_n$ is asymptotically equivalent to Λ_n under H_0.

THEOREM 3. *Suppose that $\delta_n^\infty = o_p(r_n)$, and that there exists $t > 0$ such that $\delta_n^\infty\{B_t(\theta_0)\} = o_p(r_n^{1/2})$ and $\gamma_n^\infty\{B_t(\theta_0)\} = o_p(r_n)$. Then, under H_0, $\tilde{\Lambda}_n - \Lambda_n = o_p(1)$.*

The Wald and score test statistics, W_n and S_n, are asymptotically equivalent to the likelihood ratio test, so under H_0, all three statistics have limiting distribution χ^2_{p-q}. Under the conditions of Theorem 3, $I(\theta_0)$ is consistently estimated by $r_n^{-1}\tilde{J}_n(\tilde{\theta}_n)$, so the approximate Wald and score test statistics $\tilde{W}_n$ and $\tilde{S}_n$ are also asymptotically equivalent to Λ_n.

2·3. *Approximate Bayesian inference*

We now consider the approximate posterior $\tilde{\pi}(\theta \mid y) \propto \tilde{L}_n(\theta; y)\pi(\theta)$, where we suppose that the prior is such that $\log\pi(\cdot)$ is three times differentiable. Under the same conditions that were used to show asymptotic correctness of maximum likelihood inference, the total variation distance between the approximate and exact posteriors,

$$d_{TV}\{\tilde{\pi}(\theta \mid y), \pi(\theta \mid y)\} = \frac{1}{2}\int_\Theta \big|\tilde{\pi}(\theta \mid y) - \pi(\theta \mid y)\big|d\theta,$$

tends to zero as $n \to \infty$.

THEOREM 4. *Suppose that $\delta_n^\infty = o_p(r_n)$, and that there exists $t > 0$ such that $\delta_n^\infty\{B_t(\theta_0)\} = o_p(r_n^{1/2})$ and $\gamma_n^\infty\{B_t(\theta_0)\} = o_p(r_n)$. Then $d_{TV}\{\tilde{\pi}(\theta \mid y), \pi(\theta \mid y)\} = o_p(1)$ as $n \to \infty$.*

If the Bernstein–von Mises Theorem (van der Vaart, 1998, Chapter 10) holds for the exact posterior distribution, under the conditions of Theorem 4, it will also hold for the approximate posterior distribution $\tilde{\pi}(\theta \mid y)$. In that case, credible regions formed from the approximate posterior distribution will also be valid confidence sets.

2·4. *Adjusted approximate likelihood inference*

If Lemma 1 holds for some $a_n > r_n^{1/2}$, then $\tilde{\theta}_n$ will still be a consistent estimator, but might not have the same limiting distribution as $\hat{\theta}_n$. The approximate likelihood may still be useful in practice, provided that the inference is adjusted accordingly.

The sandwich information matrix (Godambe, 1960) is $G_n(\theta) = I_n(\theta)H_n(\theta)^{-1}I_n(\theta)$ where $H_n(\theta) = \mathrm{var}\{\nabla_\theta \tilde{u}_n(\theta)\}$ and $I_n(\theta) = E\{\tilde{J}_n(\theta)\}$. Under suitable regularity conditions, $s_n^{1/2}(\tilde{\theta}_n - \theta_0) \to N\{0, \bar{G}(\theta_0)^{-1}\}$ in distribution, where s_n is the rate of convergence of $\tilde{\theta}_n$, chosen such that $G_n(\theta_0) = O(s_n)$, and $\bar{G}(\theta) = \lim_{n\to\infty} s_n^{-1} G_n(\theta)$.

Composite likelihood estimators (Lindsay, 1988) also have this type of asymptotic behaviour, and many methods which have been proposed to adjust inference using a composite likelihood could also be used to adjust the inference with an approximate likelihood, provided that $\tilde{\theta}_n$ is a consistent estimator. For example, Varin et al. (2011) describe various methods to approximate the variance of a composite likelihood estimator, and list some modifications to the composite likelihood ratio test statistic designed to ensure that the resulting test statistic has an approximate χ^2_{p-q} distribution. From a Bayesian perspective, the adjustments proposed by Pauli et al. (2011) and Ribatet et al. (2012) to posterior distributions based on composite likelihoods could also be used in the context of approximate likelihood inference.

3. EXAMPLES

3·1. *Latent variable model*

Suppose $Y_i \sim \mathrm{Binomial}(m, p_i)$, where $\mathrm{logit}\, p_i = b_i$ and $b_i \sim N(0, \theta^2)$, for $i = 1, \ldots, n$. The likelihood is $L_n(\theta) = \prod_{i=1}^n L_i(\theta)$, where

$$L_i(\theta) = \int_{-\infty}^{\infty} \{\mathrm{logit}^{-1}(b_i)\}^{y_i} \{1 - \mathrm{logit}^{-1}(b_i)\}^{m-y_i} \phi(b_i; 0, \theta) db_i$$

and $\phi(\cdot; \mu, \sigma)$ is the $N(\mu, \sigma^2)$ density function.

If we take a Laplace approximation $\tilde{L}_n(\theta)$ to the likelihood, it is intuitively clear that $m = m_n$ will have to grow with n to give valid inference as $n \to \infty$. It is less obvious whether any choice of m_n that grows with n will give valid inference, or whether m_n needs to grow with n at some minimum rate. Rue et al. (2009) suggest that any m_n which grows with n will suffice, conjecturing that the error rate is the number of latent variables over the total number of observations, although they note that this rate is not established rigorously. In this case, the error rate refers to the error in approximating $\pi(\theta \mid y)$ with $\tilde{\pi}(\theta \mid y)$, found by using a Laplace approximation to the likelihood. The integrated nested Laplace approximations proposed by Rue et al. (2009) are based on this $\tilde{\pi}(\theta \mid y)$, with further approximations used to approximate the marginal posterior distribution of each component of θ, if $p > 1$. In this example θ is a scalar, so the integrated nested Laplace approximation to the posterior distribution is precisely $\tilde{\pi}(\theta \mid y)$.

The factorization of the likelihood allows us to study the error for each item $\epsilon_i(\theta) = \ell_i(\theta) - \ell_i(\theta)$ separately, and then combine the errors. In the Supplementary Material, we show that for each fixed $\theta \in \Theta$, $\delta_i(\theta) = \|\nabla_\theta \epsilon_i(\theta)\| = O_p(m_n^{-2})$, so $\delta_n(\theta) \leq \sum_{i=1}^n \delta_i(\theta) = O_p(n m_n^{-2})$.

However, the conditions are in terms of uniform rather than pointwise errors. Since $\delta_i(\theta)$ is maximized at a point $\theta^* = O_p(m_n^{-1/2})$ and is decreasing for all $\theta > \theta^*$, $\delta_i^\infty = \delta_i(\theta^*) = O_p(m_n^{-1/2})$, so $\delta_n^\infty = O_p(nm_n^{-1/2})$.

The amount of information that the data provides about θ is bounded for fixed n as $m_n \to \infty$ by the available information on θ given the value of each b_i. So if $m_n \to \infty$ as $n \to \infty$, then $r_n = O(n)$, and since $\delta_n^\infty = o_p(r_n)$, $\tilde{\theta}_n$ will be consistent.

Provided that $\theta_0 \neq 0$, choose any fixed $t \in (0, \theta_0)$. Then, for sufficiently large n, $\delta_n^\infty\{B_t(\theta_0)\} = \delta_n(\theta_0 - t) = O_p(nm_n^{-2})$ and $\gamma_n^\infty\{B_t(\theta_0)\} = \gamma_n(\theta_0 - t) = O_p(nm_n^{-2})$. If m_n grows at a rate faster than $n^{1/4}$, then $\delta_n^\infty\{B_t(\theta_0)\} = o_p(n^{1/2})$ and $\gamma_n^\infty\{B_t(\theta_0)\} = o_p(n)$, so the Laplace approximation to the likelihood will give first-order correct inference.

To illustrate these results, we simulate 10000 realizations from the model with $\theta_0 = 0.5$, values of n between 1000 and 10000, and $m_n = \min\{1, 5 + 4(n^a - 1000^a)\}$, for $a = 0.2$, 0.25 or 0.3. The three choices of m_n are shown in Figure 1a.

A very accurate approximation to the likelihood may be obtained by using adaptive Gaussian quadrature with 20 quadrature points to approximate each of the univariate integrals $L_i(\theta)$, and we use this as a proxy for the true likelihood $L(\theta)$.

As $n \to \infty$, we have $r_n = O(n)$, but for smaller sample sizes $E\{\|J_n(\hat{\theta}_n)\|\}$ still grows with m_n. This quantity may be approximated by $\hat{r}_n = \hat{E}\{\|J_n(\hat{\theta}_n)\|\}$, where $\hat{E}(\cdot)$ is the sample mean over the 10000 realizations. The functional form of m_n was chosen to make $\hat{r}_n^{-1/2}\delta_n(\theta_0)$ approximately constant when $a = 0.25$, as shown in Figure 1b. The same quantity grows with n for $a = 0.2$ and shrinks with n for $a = 0.3$.

Figure 1c shows the root mean squared error for the Laplace estimator, and as expected, the estimator is consistent for all three choices of a. The root mean squared error of the Laplace estimator divided by that of the maximum likelihood estimator, shown in Figure 1d, grows in the $a = 0.2$ case, stays approximately constant if $a = 0.25$, and shrinks towards 1 if $a = 0.3$.

Figure 1e shows the empirical coverage of nominal 90% likelihood ratio type confidence intervals for θ. The upper three lines show that the intervals constructed using the true likelihood have very close to nominal coverage for each a. The lower three lines show the coverage of the approximate likelihood intervals, which decreases with n for $a = 0.2$, and increases towards the nominal 90% level for $a = 0.3$.

Figure 1f shows the total variation distance between $\tilde{\pi}(\theta \mid y)$ and $\pi(\theta \mid y)$, with prior $\pi(\theta) \propto 1/\theta$. The distance between the approximate posterior and the exact posterior grows when $a = 0.2$, stays approximately constant when $a = 0.25$, and shrinks towards zero when $a = 0.3$. The behaviour for $a = 0.2$ refutes the conjecture of Rue et al. (2009) that the error in the approximate posterior distribution should shrink to zero provided that m_n grows with n.

3·2. *Ising model*

We consider a simple Ising model for $n = rc$ variables $y_i \in \{-1, 1\}$, arranged on an $r \times c$ lattice, with parameters $\theta = (\alpha, \beta)$, so that

$$\text{pr}(Y = y; \theta) = Z_{r,c}(\theta)^{-1} \exp\{\alpha V_0(y) + \beta V_1(y)\},$$

where $V_0(y) = \sum_i y_i$, and $V_1(y) = \sum_{i \sim j} y_i y_j$. Here $i \sim j$ indicates that i and j have an edge between them in the lattice, and $Z_{r,c}(\theta) = \sum_{y \in \{-1,1\}^n} \exp\{\alpha V_0(y) + \beta V_1(y)\}$ is a normalizing constant. The likelihood function $L(\theta; y) = \text{pr}(Y = y; \theta)$ depends on $Z_{r,c}(\theta)$, and it is the computation of this normalizing constant that makes evaluation of the likelihood function difficult. By using variable elimination (e.g. Jordan, 2004), $Z_{r,c}(\theta)$ may be computed at cost $O\{rc2^{\min(r,c)}\}$, which remains infeasibly expensive if both r and c are large.

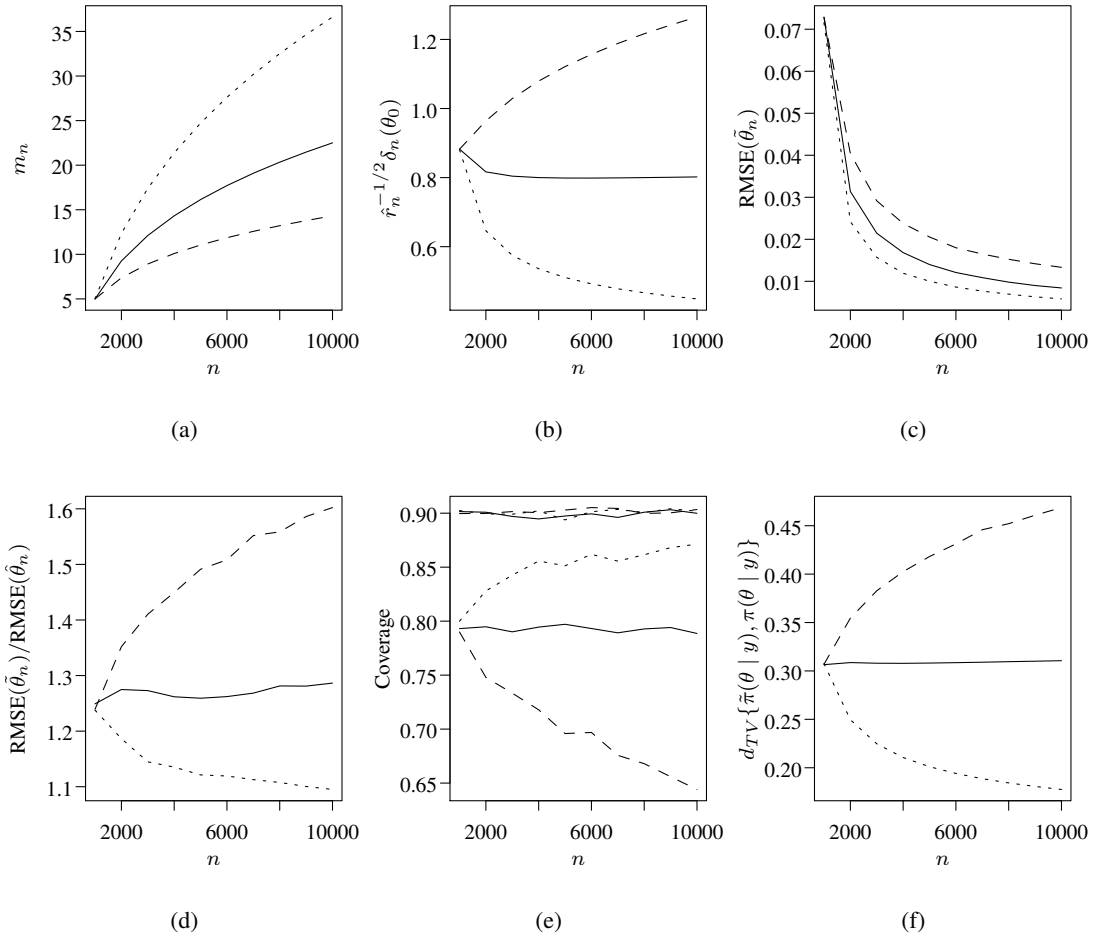
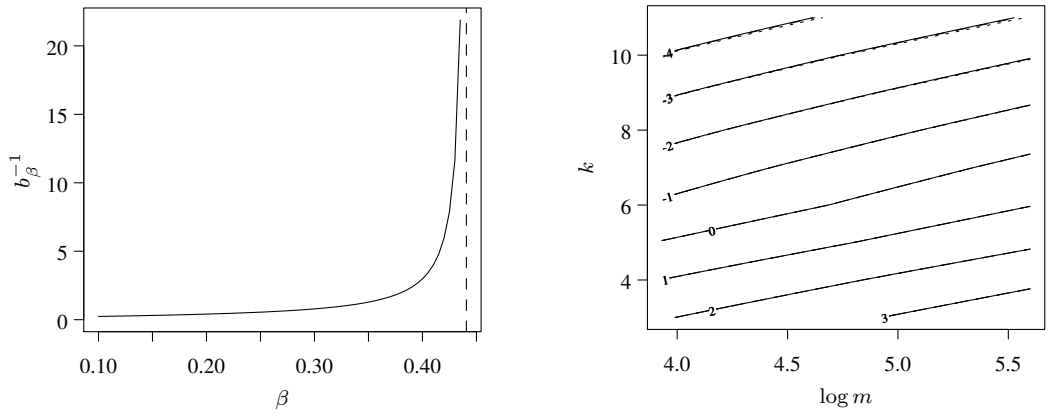


Fig. 1: Comparison of exact and approximate likelihood inference for a two-level model with m_n observations on each of n items, where $m_n = \min\{1, 5 + 4(n^a - 1000^a)\}$, for $a = 0.2$ (dashed lines), 0.25 (solid lines) and 0.3 (dotted lines). The quantities shown are: (a) the three choices of m_n; (b) the rescaled error in the score function $\hat{r}_n^{-1/2}\delta_n(\theta_0)$; (c) the root mean squared error of $\tilde{\theta}_n$; (d) the root mean squared error of $\tilde{\theta}_n$ relative to that of $\hat{\theta}_n$; (e) the coverage of 90% confidence intervals, constructed by using the true likelihood (upper three lines) or the Laplace approximation to the likelihood (lower three lines); (f) the total variation distance between the approximate and exact posterior distributions. RMSE denotes root mean squared error.

Many methods for approximating $Z_{r,c}(\theta)$ have been proposed. We study properties of inference using the reduced-dependence approximations introduced by Friel et al. (2009), a family of approximations controlled by a positive integer tuning parameter, which we call k. The approximation for fixed k is $\tilde{Z}^{(k)}_{r,c}(\theta) = Z_{k,c}(\theta)^{r-k+1}/Z_{k-1,c}(\theta)^{r-k}$.

We consider the case $r = c = m$, using a reduced-dependence approximation at level k to approximate the likelihood, giving $\tilde{L}^{(k)}_m(\beta) = \tilde{Z}^{(k)}_{m,m}(\theta)^{-1}\exp\{\alpha V_0(y) + \beta V_1(y)\}$. The true likelihood may be computed at cost $O(m^2 2^m)$, and the reduced-dependence approximation at level k at cost $O(m^2 + km2^k)$. The aim is to understand how $k = k_m$ should vary with m to give



(a) b_β^{-1}, for a restricted Ising model with $\alpha = 0$.

(b) Contour plot of $\log\{m^{-1}\delta_m^{(k)}(0.1, 0.3)\}$.

Fig. 2: Inference for an Ising model with a reduced-dependence approximation to the likelihood.

asymptotically valid inference as $m \to \infty$. The error in the log-likelihood,

$$\epsilon_m^{(k)}(\beta) = \tilde{\ell}_m^{(k)}(\beta) - \ell_m(\beta) = \log Z_{m,m}(\beta) - \log \tilde{Z}_{m,m}^{(k)}(\beta),$$

does not depend on the data y, so we do not need to consider the statements in probability: all of the errors are deterministic in this case.

If $\alpha = 0$, and the lattice has periodic boundary conditions, so that the top row of variables are joined to the bottom row, and the left row joined to the right, Kaufman (1949) provides a relatively simple expression for $Z_{r,c}(0, \beta)$, so that it is possible to compute the likelihood exactly, even for large lattices.

We restrict the parameter space to $\alpha = 0$, and assume $\beta \in [0, 0.43]$. This guarantees that $\beta < \beta_c = \log(1 + \sqrt{2})/2 \approx 0.44$, where β_c is a critical value at which the behaviour of the Ising model suddenly changes, so that for $\beta > \beta_c$ large areas of all plus ones or all minus ones are observed. If $\beta_0 = \beta_c$, the maximum likelihood estimator may not have a normal limiting distribution, so our results do not apply to this case.

The information provided by the data about β grows at rate $r_m = m^2$. In the Supplementary Material, we show that if $k \to \infty$ as $m \to \infty$, then $\delta_m^{(k)}(\beta) = O\{m^2 k \exp(-b_\beta k)\} + o(1)$, where $b_\beta = 2\cosh^{-1}\{-1 + \cosh(2\beta)^2/\sinh(2\beta)\}$.

For any choice of k which grows with m, $\sup_{\beta\in[0,0.43]} \delta_m^{(k)}(\beta) = o(m^2)$, so the approximate likelihood estimator will be consistent.

In order to meet the conditions of Theorem 2, $k = k_m$ should be chosen so that $\delta_m^{(k)}\{B_t(\beta_0)\} = o(m)$. For any $t < \beta_c - \beta_0$, for sufficiently large m, $\delta_m^{(k)}\{B_t(\beta_0)\} = \delta_m^{(k)}(\beta_0 + t) = O\{m^2 k \exp(-b_{\beta_0+t} k)\} + o(1)$. Since b_β is a continuous function of β, any k such that $m^2 k \exp(-b_{\beta_0} k) = o(m)$ will meet this condition for sufficiently small t. This may be achieved by taking $k_m = c_{\beta_0} \log m$, for any $c_{\beta_0} > b_{\beta_0}^{-1}$. Figure 2a shows how b_β^{-1} varies with β.

For any $k \to \infty$ as $m \to \infty$, $\gamma_m^{(k)}\{B_t(\beta_0)\} = o(m^2)$, so the reduced-dependence approximation to the likelihood with $k_m = c_{\beta_0} \log m$ will provide asymptotically valid inference for any $c_{\beta_0} > b_{\beta_0}^{-1}$. The cost of computing this approximation is $O(m \log m\, m^{c_{\beta_0} \log 2}) < O(m^{c_{\beta_0} \log 2 + 2})$, polynomial in the size of the model, but increasing rapidly as β_0 approaches the critical value.

If $\alpha \neq 0$, no simple expression for $Z_{r,c}(\alpha, \beta)$ is known. Instead, we investigate the behaviour of $m^{-1}\delta_m^{(k)}(\alpha, \beta)$ numerically, by using $\tilde{\ell}_m^{(K)}$ for some large K as a proxy for the true log-likelihood. Care is needed to choose K sufficiently large to ensure that the results are not sensitive to the choice of K.

Taking $K = 16$, a contour plot of this approximation to $\log\{m^{-1}\delta_m^{(k)}(0.1, 0.3)\}$ against $\log m$ and k is shown in Figure 2b, for $k = 2, \ldots, 12$, $m = 50, \ldots, 300$. The same plots with $K = 15$ and $K = 14$ are overlaid, and the differences are barely visible at this scale. Given this stability, it seems reasonable to assume that a contour plot of the true $\log\{m^{-1}\delta_m^{(k)}(0.1, 0.3)\}$ would look very similar to this. To obtain asymptotically valid inference, k_m should be chosen so that this rescaled error shrinks with m, which seems to occur if k_m grows at rate $c(0.1, 0.3) \log m$, for $c(0.1, 0.3)$ larger than about 1.5. This pattern of behaviour is very similar to the $\alpha = 0$ case. In both cases, reduced-dependence approximations with an appropriate choice of k will give asymptotically valid inference at cost polynomial in the size of the model, in contrast with the exponential cost of computing the likelihood exactly.

4. DISCUSSION

The results obtained here can also be applied to other approximations to the likelihood and to other types of model. The conditions on the approximate likelihood, such as showing that $\delta_n^{\infty}\{B_t(\theta_0)\} = o_p(r_n)$, may be difficult to verify in practice, as the true likelihood is assumed to be unavailable. If the approximation to the likelihood is a truncation of a series expansion for the true likelihood, as is the case for the Laplace approximation, the conditions can be checked by examining the contribution from higher-order terms in the expansion. In other cases, it may only be possible to investigate the size of the errors numerically, by using used a more accurate and expensive approximation to the likelihood as a proxy for the true likelihood, as in the Ising model example with $\alpha \neq 0$.

Many approximation methods have a tuning parameter, k, say, where increasing k allows computation of a more accurate likelihood approximation at increased cost. The reduced-dependence approximation at level k described in Section 3·2 is one example of this. In order for approximate likelihood inference to be close to true likelihood inference, $k = k_n$ should be allowed to vary with n. Given an understanding of how the error in the score function varies with k and n, the results obtained here could be applied to determine how k_n should scale with n. This has the potential to allow the construction of approximate likelihoods which match the inference with the true likelihood closely for all n, but which scale well to large data sizes.

ACKNOWLEDGEMENTS

I am grateful to David Firth, Nancy Reid, Cristiano Varin, the associate editor and two referees for comments which have greatly improved this paper. This work was supported by the U.K. Engineering and Physical Sciences Research Council.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes more detailed justification of claims made in Sections 3·1 and 3·2.

APPENDIX: PROOFS OF RESULTS

Proof of Theorem 1

Proof. We apply Theorem 5.9 of van der Vaart (1998), with $\Psi_n(\theta) = r_n^{-1}\tilde{u}_n(\theta)$ and $\Psi(\theta) = \bar{u}(\theta)$. Then

$$\begin{aligned}\sup_{\theta\in\Theta}\|\Psi_n(\theta) - \Psi(\theta)\| &= \sup_{\theta\in\Theta}\|\bar{u}_n(\theta) + r_n^{-1}\nabla_\theta\epsilon_n(\theta) - \bar{u}(\theta)\| \\ &\leq \sup_{\theta\in\Theta}\|\bar{u}_n(\theta) - \bar{u}(\theta)\| + r_n^{-1}\delta_n \\ &= o_p(1)\end{aligned}$$

The conditions of Theorem 5.9 of van der Vaart (1998) are met, so since $\Psi_n(\tilde{\theta}_n) = 0$, we have that $\tilde{\theta}_n \to \theta_0$ in probability, as $n \to \infty$. □

Proof of Lemma 1

Proof. Taking a Taylor expansion of $\bar{u}_n$ about $\hat{\theta}_n$,

$$\bar{u}_n(\theta) = \bar{u}_n(\hat{\theta}_n) - (\theta - \hat{\theta}_n)^T\bar{J}_n(\theta_n^*) = -(\theta - \hat{\theta}_n)^T\bar{J}_n(\theta_n^*),$$

for some θ_n^* between θ and $\hat{\theta}_n$. Define $\bar{\tilde{u}}_n(\theta) = r_n^{-1}\tilde{u}_n(\theta)$. Then

$$\bar{\tilde{u}}_n(\theta) = \bar{u}_n(\theta) + r_n^{-1}\nabla_\theta\epsilon_n(\theta) = -(\theta - \hat{\theta}_n)^T\bar{J}_n(\theta_n^*) + r_n^{-1}\nabla_\theta\epsilon_n(\theta),$$

so any $\tilde{\theta}_n$ solving $\bar{\tilde{u}}_n(\tilde{\theta}_n) = 0$ solves $\tilde{\theta}_n - \hat{\theta}_n = \bar{J}_n^{-1}(\theta_n^*)r_n^{-1}\nabla_\theta\epsilon_n(\tilde{\theta}_n)$, for some θ_n^* between $\tilde{\theta}_n$ and $\hat{\theta}_n$. But θ_n^* is a consistent estimator of θ, because $\tilde{\theta}_n$ is by Theorem 1, so $\bar{J}_n(\theta_n^*) \to I(\theta_0)$ in probability. So $\tilde{\theta}_n - \hat{\theta}_n = O_p\{r_n^{-1}\delta_n(\tilde{\theta}_n)\}$.

Write A_n for the event $\{\tilde{\theta}_n \in B_t(\theta_0)\}$. Then $\text{pr}(A_n) \to 1$ as $n \to \infty$ since $\tilde{\theta}_n$ is consistent, and conditional on A_n, $\tilde{\theta}_n - \hat{\theta}_n = O_p[r_n^{-1}\delta_n^\infty\{B_t(\theta_0)\}] = o_p(r_n^{-1}a_n)$. For any $\epsilon > 0$,

$$\begin{aligned}&\text{pr}(\|\tilde{\theta}_n - \hat{\theta}_n\| \geq \epsilon a_n r_n^{-1}) \\ &\quad= \text{pr}(\|\tilde{\theta}_n - \hat{\theta}_n\| \geq \epsilon a_n r_n^{-1} \mid A_n)\text{pr}(A_n) + \text{pr}(\|\tilde{\theta}_n - \hat{\theta}_n\| \geq \epsilon a_n r_n^{-1} \mid A_n^C)\text{pr}(A_n^C) \\ &\quad\leq \text{pr}(\|\tilde{\theta}_n - \hat{\theta}_n\| \geq \epsilon a_n r_n^{-1} \mid A_n) + \text{pr}(A_n^C) \\ &\quad\to 0, \quad n \to \infty.\end{aligned}$$

Proof of Theorem 2

Proof. Applying Lemma 1 with $a_n = r_n^{1/2}$, $\tilde{\theta}_n - \hat{\theta}_n = o_p(r_n^{1/2}r_n^{-1}) = o_p(r_n^{-1/2})$. So

$$r_n^{1/2}(\tilde{\theta}_n - \theta_0) = r_n^{1/2}(\hat{\theta}_n - \theta_0) + o_p(1) \to N\{0, I(\theta_0)^{-1}\}, \quad n \to \infty,$$

in distribution, as required. □

Proof of Theorem 3

Proof. We have

$$\begin{aligned}(\tilde{\Lambda}_n - \Lambda_n)/2 &= \{\tilde{\ell}_n(\tilde{\theta}_n) - \tilde{\ell}_n(\tilde{\theta}_n^R)\} - \{\ell_n(\hat{\theta}_n) - \ell_n(\hat{\theta}_n^R)\} \\ &= \{\tilde{\ell}_n(\tilde{\theta}_n) - \tilde{\ell}_n(\hat{\theta}_n)\} + \{\tilde{\ell}_n(\hat{\theta}_n) - \ell_n(\hat{\theta}_n)\} + \{\tilde{\ell}_n(\hat{\theta}_n^R) - \tilde{\ell}_n(\tilde{\theta}_n^R)\} + \{\ell_n(\hat{\theta}_n^R) - \tilde{\ell}_n(\hat{\theta}_n^R)\}.\end{aligned}$$

Taking a Taylor expansion of the first term,

$$\begin{aligned}\tilde{\ell}_n(\hat{\theta}_n) - \tilde{\ell}_n(\tilde{\theta}_n) &= (\hat{\theta}_n - \tilde{\theta}_n)^T\tilde{u}_n(\tilde{\theta}_n) + (\hat{\theta}_n - \tilde{\theta}_n)^T\tilde{J}_n(\bar{\theta}_n)(\hat{\theta}_n - \tilde{\theta}_n) \\ &= (\hat{\theta}_n - \tilde{\theta}_n)^T\tilde{J}_n(\bar{\theta}_n)(\hat{\theta}_n - \tilde{\theta}_n)\end{aligned}$$

for some $\bar{\theta}_n$ between $\hat{\theta}_n$ and $\tilde{\theta}_n$, so $\tilde{\ell}_n(\hat{\theta}_n) - \tilde{\ell}_n(\tilde{\theta}_n) = o_p(1)$, since $\|\hat{\theta}_n - \tilde{\theta}_n\| = o_p(r_n^{-1/2})$ and $\tilde{J}_n(\bar{\theta}_n) = J_n(\bar{\theta}_n) + o_p(r_n) = O_p(r_n)$. Similarly, $\tilde{\ell}_n(\hat{\theta}_n^R) - \tilde{\ell}_n(\tilde{\theta}_n^R) = o_p(1)$, so

$$\begin{aligned}(\tilde{\Lambda}_n - \Lambda_n)/2 &= \epsilon_n(\hat{\theta}_n) - \epsilon_n(\hat{\theta}_n^R) + o_p(1) \\ &= (\hat{\theta}_n - \hat{\theta}_n^R)^T \nabla_\theta \epsilon_n(\theta_n^*) + o_p(1)\end{aligned}$$

for some θ_n^* between $\hat{\theta}_n$ and $\hat{\theta}_n^R$. But $\hat{\theta}_n - \hat{\theta}_n^R = O_p(r_n^{-1/2})$, and, for sufficiently large n, $\|\nabla_\theta \epsilon_n(\theta_n^*)\| \leq \delta_n^\infty\{B_t(\theta_0)\}$, so $\tilde{\Lambda}_n - \Lambda_n = o_p(1)$, as required.

Results needed to prove Theorem 4

In order to prove Theorem 4, it is helpful to first consider properties of inference with the penalized log-likelihood $\ell_n^\pi(\theta) = \ell(\theta) + \log \pi(\theta)$ with log-prior penalty. We write $\hat{\theta}_n^\pi$ for the corresponding penalized likelihood estimator, which is the posterior mode. Similarly, write A^π for the version of the quantity A computed with $\ell_n^\pi(\cdot)$ in place of $\ell_n(\cdot)$, and $\tilde{A}^\pi$ for the approximate version. Since $\epsilon_n^\pi(\theta) = \epsilon_n(\theta)$, all the error terms remain unchanged.

Under the regularity conditions assumed on the model, the penalized likelihood estimator $\hat{\theta}_n^\pi$ will be consistent, and for any $b_n = o_p(r_n)$, the posterior probability that $\theta \in B_{b_n^{-1/2}}(\theta_0)$ will tend to one as $n \to \infty$.

To prove Theorem 4, we will use the following lemma, which says that the error in the penalized log-likelihood ratio may be approximated in terms of the error in the score function.

LEMMA A2. *Suppose that $\delta_n^\infty = o_p(r_n)$, and that there exists $t > 0$ such that $\delta_n^\infty\{B_t(\theta_0)\} = o_p(r_n^{1/2})$. Then $\hat{\theta}_n^\pi$ and $\tilde{\theta}_n^\pi$ are consistent estimators of θ_0, and $\hat{\theta}_n^\pi - \tilde{\theta}_n^\pi = o_p(r_n^{-1/2})$.*
If $\gamma_n^\infty\{B_t(\theta_0)\} = o_p(b_n)$, for some $b_n = o(r_n)$, then

$$\sup_{\theta \in B_{b_n^{-1/2}}(\hat{\theta}_n^\pi)} \left| \left[\{\tilde{\ell}_n^\pi(\tilde{\theta}_n^\pi) - \tilde{\ell}_n^\pi(\theta)\} - \{\ell_n^\pi(\hat{\theta}_n^\pi) - \ell_n^\pi(\theta)\}\right] - (\hat{\theta}_n^\pi - \theta)^T \nabla_\theta \epsilon_n(\theta) \right| = o_p(1).$$

Proof. That $\hat{\theta}_n^\pi - \tilde{\theta}_n^\pi = o_p(r_n^{-1/2})$ follows by a similar argument to the proof of Theorem 2. Write

$$\begin{aligned}C_n(\theta) &= \{\tilde{\ell}_n^\pi(\tilde{\theta}_n^\pi) - \tilde{\ell}_n^\pi(\theta)\} - \{\ell_n^\pi(\hat{\theta}_n^\pi) - \ell_n^\pi(\theta)\} \\ &= \{\tilde{\ell}_n^\pi(\hat{\theta}_n^\pi) - \ell_n^\pi(\hat{\theta}_n^\pi)\} - \{\tilde{\ell}_n^\pi(\theta) - \ell_n^\pi(\theta)\} - \{\tilde{\ell}_n^\pi(\hat{\theta}_n^\pi) - \tilde{\ell}_n^\pi(\tilde{\theta}_n^\pi)\} \\ &= \epsilon_n(\hat{\theta}_n^\pi) - \epsilon_n(\theta) - \{\tilde{\ell}_n^\pi(\hat{\theta}_n^\pi) - \tilde{\ell}_n^\pi(\tilde{\theta}_n^\pi)\}.\end{aligned}$$

Then

$$\tilde{\ell}_n(\hat{\theta}_n^\pi) - \tilde{\ell}_n(\tilde{\theta}_n^\pi) = (\hat{\theta}_n^\pi - \tilde{\theta}_n^\pi)^T \tilde{J}_n^\pi(\bar{\theta}_n)(\hat{\theta}_n^\pi - \tilde{\theta}_n^\pi)$$

for some $\bar{\theta}_n$ between $\hat{\theta}_n^\pi$ and $\tilde{\theta}_n^\pi$, and

$$\epsilon_n(\hat{\theta}_n^\pi) - \epsilon_n(\theta) = (\hat{\theta}_n^\pi - \theta)^T \nabla_\theta \epsilon_n(\theta) + (\hat{\theta}_n^\pi - \theta)^T \nabla_\theta^T \nabla_\theta \epsilon_n\{\theta_n^*(\theta)\}(\hat{\theta}_n^\pi - \theta)$$

for some $\theta_n^*(\theta)$ between $\hat{\theta}_n^\pi$ and θ. Write

$$\begin{aligned}D_n(\theta) &= C_n(\theta) - (\hat{\theta}_n^\pi - \theta)^T \nabla_\theta \epsilon_n(\theta) \\ &= (\hat{\theta}_n^\pi - \theta)^T \nabla_\theta^T \nabla_\theta \epsilon_n\{\theta_n^*(\theta)\}(\hat{\theta}_n^\pi - \theta) - (\hat{\theta}_n^\pi - \tilde{\theta}_n^\pi)^T \tilde{J}_n^\pi(\bar{\theta}_n)(\hat{\theta}_n^\pi - \tilde{\theta}_n^\pi).\end{aligned}$$

Then $\sup_{\theta\in B_{b_n^{-1/2}}(\hat\theta_n^\pi)}\left|D_n(\theta)\right|$ may be expressed as

$$\begin{aligned}
&\sup_{\theta\in B_{b_n^{-1/2}}(\hat\theta_n^\pi)}\left[\left|(\hat\theta_n^\pi-\theta)^T\nabla_\theta^T\nabla_\theta\epsilon_n\{\theta_n^*(\theta)\}(\hat\theta_n^\pi-\theta)-(\hat\theta_n^\pi-\tilde\theta_n^\pi)^T\tilde J_n^\pi(\bar\theta_n)(\hat\theta_n^\pi-\tilde\theta_n^\pi)\right|\right]\\
&\le \sup_{\theta\in B_{b_n^{-1/2}}(\hat\theta_n^\pi)}\left[\left|(\hat\theta_n^\pi-\theta)^T\nabla_\theta^T\nabla_\theta\epsilon_n\{\theta_n^*(\theta)\}(\hat\theta_n^\pi-\theta)\right|\right]+\left|(\hat\theta_n^\pi-\tilde\theta_n^\pi)^T\tilde J_n^\pi(\bar\theta_n)(\hat\theta_n^\pi-\tilde\theta_n^\pi)\right| \qquad 340\\
&\le b_n^{-1}\delta_n^\infty\{B_t(\theta_0)\}+o_p(1)
\end{aligned}$$

for n sufficiently large. But $\delta_n^\infty\{B_t(\theta_0)\}=o_p(b_n)$, so $\sup_{\theta\in B_{b_n^{-1/2}}(\hat\theta_n^\pi)}\left|D_n(\theta)\right|=o_p(1)$, as required. □

Proof of Theorem 4

Proof. The normalized exact and approximate posterior distributions are $\pi(\theta\mid y)=L_n(\theta)\pi(\theta)/Z_n$ and $\tilde\pi(\theta\mid y)=\tilde L_n(\theta)\pi(\theta)/\tilde Z_n$, where $Z_n=\int L_n(\theta)\pi(\theta)d\theta$ and $\tilde Z_n=\int\tilde L_n(\theta)\pi(\theta)d\theta$. 345

First, we find a Laplace approximation $\hat Z_n=(2\pi)^{-p/2}|J_n^\pi(\hat\theta_n^\pi)|^{-1/2}L_n^\pi(\hat\theta_n^\pi)$ to Z_n. Then $\log Z_n-\log\hat Z_n=o_p(1)$, because Z_n is a p-dimensional integral, where p remains fixed as $n\to\infty$. Similarly, $\tilde Z_n$ may be approximated by using Laplace's method. Then

$$\begin{aligned}
\log Z_n-\log\tilde Z_n&=\log\hat Z_n-\log\hat{\tilde Z}_n+o_p(1)\\
&=\{\log|J_n^\pi(\hat\theta_n^\pi)|-\log|\tilde J_n^\pi(\tilde\theta_n^\pi)|\}/2+\ell_n^\pi(\hat\theta_n^\pi)-\tilde\ell_n^\pi(\tilde\theta_n^\pi)+o_p(1). \qquad 350
\end{aligned}$$

Since $\gamma_n^\infty\{B_t(\theta_0)\}=o_p(r_n)$, both $r_n^{-1}J_n^\pi(\hat\theta_n^\pi)$ and $r_n^{-1}\tilde J_n^\pi(\hat\theta_n^\pi)$ converge towards $I(\theta_0)$, so

$$\begin{aligned}
\log|J_n^\pi(\hat\theta_n^\pi)|-\log|\tilde J_n^\pi(\tilde\theta_n^\pi)|&=\log|r_n^{-1}J_n^\pi(\hat\theta_n^\pi)|-\log|r_n^{-1}\tilde J_n^\pi(\tilde\theta_n^\pi)|\\
&=\log|I(\theta_0)+o_p(1)|-\log|I(\theta_0)+o_p(1)|\\
&=o_p(1).
\end{aligned}$$

So $\log Z_n-\log\tilde Z_n=\ell_n^\pi(\hat\theta_n^\pi)-\tilde\ell_n^\pi(\tilde\theta_n^\pi)+o_p(1)$. Since $\delta_n^\infty\{B_t(\theta_0)\}=o_p(r_n^{-1/2})$ and $\gamma_n^\infty\{B_t(\theta_0)\}=$ 355
$o_p(r_n)$, we may choose $b_n=o(r_n)$ such that $\delta_n^\infty\{B_t(\theta_0)\}=o_p(b_n^{-1/2})$ and $\gamma_n^\infty\{B_t(\theta_0)\}=o_p(b_n)$. Writing $S_n=B_{b_n^{-1/2}}(\hat\theta_n^\pi)$,

$$\begin{aligned}
&\sup_{\theta\in S_n}\left\{\left|\log\tilde\pi(\theta\mid y)-\log\pi(\theta\mid y)-(\hat\theta_n^\pi-\theta)^T\nabla_\theta\epsilon_n(\theta)\right|\right\}\\
&\quad=\sup_{\theta\in S_n}\left\{\left|\tilde\ell_n^\pi(\theta)-\log\tilde Z_n-\ell_n^\pi(\theta)+\log Z_n-(\hat\theta_n^\pi-\theta)^T\nabla_\theta\epsilon_n(\theta)\right|\right\}\\
&\quad\le\sup_{\theta\in S_n}\left\{\left|\tilde\ell_n^\pi(\theta)-\tilde\ell_n^\pi(\tilde\theta_n^\pi)-\ell_n^\pi(\theta)+\hat\ell_n^\pi(\hat\theta_n^\pi)-(\hat\theta_n^\pi-\theta)^T\nabla_\theta\epsilon_n(\theta)\right|\right\}+o_p(1) \qquad 360\\
&\quad=o_p(1),
\end{aligned}$$

by Lemma 2. Then

$$\begin{aligned}
2d_{TV}\{\tilde{\pi}(\theta \mid y), \pi(\theta \mid y)\} &= \int_{\Theta} \big|\tilde{\pi}(\theta \mid y) - \pi(\theta \mid y)\big| d\theta \\
&= \int_{S_n} \big|\tilde{\pi}(\theta \mid y) - \pi(\theta \mid y)\big| d\theta + \int_{S_n^C} \big|\tilde{\pi}(\theta \mid y) - \pi(\theta \mid y)\big| d\theta \\
&\leq \int_{S_n} \big|\tilde{\pi}(\theta \mid y) - \pi(\theta \mid y)\big| d\theta + \int_{S_n^C} 2\big|\pi(\theta \mid y)\big| d\theta + \int_{S_n^C} 2\big|\tilde{\pi}(\theta \mid y)\big| d\theta \\
&= \int_{S_n} \left| \frac{\tilde{\pi}(\theta \mid y)}{\pi(\theta \mid y)} - 1 \right| \pi(\theta \mid y) d\theta + o_p(1) \\
&\leq \sup_{\theta \in S_n} \left| \frac{\tilde{\pi}(\theta \mid y)}{\pi(\theta \mid y)} - 1 \right| + o_p(1) \\
&= \sup_{\theta \in S_n} \big| \exp\{\log \tilde{\pi}(\theta \mid y) - \log \pi(\theta \mid y)\} - 1 \big| + o_p(1) \\
&\leq \Big| \exp\Big\{ \sup_{\theta \in S_n} |\log \tilde{\pi}(\theta \mid y) - \log \pi(\theta \mid y)| \Big\} - 1 \Big| + o_p(1).
\end{aligned}$$

But

$$\begin{aligned}
\sup_{\theta \in S_n} |\log \tilde{\pi}(\theta \mid y) - \log \pi(\theta \mid y)| &\leq \sup_{\theta \in S_n} |(\hat{\theta}_n^{\pi} - \theta)^T \nabla_{\theta} \epsilon_n(\theta)| + o_p(1) \\
&= b_n^{-1/2} \delta_n^{\infty}\{B_t(\theta_0)\} + o_p(1) \\
&= o_p(1).
\end{aligned}$$

So $d_{TV}\{\tilde{\pi}(\theta \mid y), \pi(\theta \mid y)\} = o_p(1)$, as claimed. □

REFERENCES

BATES, D., MÄCHLER, M., BOLKER, B. & WALKER, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* **67**, 1–48.

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

FRIEL, N., PETTITT, A., REEVES, R. & WIT, E. (2009). Bayesian inference in hidden Markov random fields for binary data defined on large lattices. *Journal of Computational and Graphical Statistics* **18**, 243–261.

GODAMBE, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Ann. Math. Statist.* **31**, 1208–1211.

JORDAN, M. I. (2004). Graphical models. *Statist. Sci.* **19**, 140–155.

KAUFMAN, B. (1949). Crystal statistics. II. Partition function evaluated by spinor analysis. *Physical Review* **76**, 1232.

LINDSAY, B. G. (1988). Composite likelihood methods. *Contemporary Mathematics* **80**, 221–239.

PAULI, F., RACUGNO, W. & VENTURA, L. (2011). Bayesian composite marginal likelihoods. *Statistica Sinica* **21**, 149–164.

PINHEIRO, J. C. & BATES, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics* **4**, 12–35.

RIBATET, M., COOLEY, D. & DAVISON, A. C. (2012). Bayesian inference from composite likelihoods, with an application to spatial extremes. *Statistica Sinica* **22**, 813–845.

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

TJELMELAND, H. & AUSTAD, H. M. (2012). Exact and approximate recursive calculations for binary markov random fields defined on graphs. *Journal of Computational and Graphical Statistics* **21**, 758–780.

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

VARIN, C., REID, N. & FIRTH, D. (2011). An overview of composite likelihood methods. *Statistica Sinica* **21**, 5–42.

[*Received April* 2012. *Revised September* 2012]