PROBABILITY
AND
MATHEMATICAL STATISTICS
Vol. 36, Fasc. 2 (2016), pp. 295–310

ASYMPTOTICS OF MONTE CARLO MAXIMUM LIKELIHOOD ESTIMATORS*

BY

BŁAŻEJ MIASOJEDOW (WARSZAWA), WOJCIECH NIEMIRO (WARSZAWA),
JAN PALCZEWSKI (LEEDS), AND WOJCIECH REJCHEL (TORUŃ)

Abstract. We describe Monte Carlo approximation to the maximum likelihood estimator in models with intractable norming constants and explanatory variables. We consider both sources of randomness (due to the initial sample and to Monte Carlo simulations) and prove asymptotical normality of the estimator.

2010 AMS Mathematics Subject Classification: Primary: 62F12; Secondary: 60F05.

Key words and phrases: Asymptotic statistics, empirical process, importance sampling, maximum likelihood estimation, Monte Carlo method.

1. INTRODUCTION

Maximum likelihood (ML) is a well-known and often used method in estimation of parameters in statistical models. However, for many complex models exact calculation of such estimators is very difficult or impossible. Such problems arise if considered densities are known only up to intractable norming constants, for instance, in Markov random fields or spatial statistics. The wide range of applications of models with unknown norming constants is discussed, e.g., in [11]. Methods proposed to overcome the problems with computing ML estimates in such models include, among others, maximum pseudolikelihood (MPL) [1] or Monte Carlo maximum likelihood (MCML) [2], [7], [6], [14]. MPL estimators are easy to compute but not efficient. This is demonstrated, e.g., in [17] for an important autologistic spatial model via a simulation study. Comparison of MLP or "coding method" with MCML is also discussed in [8]. In our paper we focus on MCML.

In influential papers [7], [6] the authors prove consistency and asymptotic normality of MCML estimators under the assumption that the initial sample is fixed,

* Work partially supported by Polish National Science Center No. N N201 608 740.

and only the Monte Carlo sample size tends to infinity. Both sources of randomness (one due to the initial sample and the other due to Monte Carlo simulations) are considered in [2], [14], [18]. The authors of the first mentioned paper apply the general importance sampling recipe. They show that for their scheme of simulations, the Monte Carlo sample size has to grow exponentially fast to ensure consistency of the estimator. As the remedy for this problem they propose to use a preliminary estimator which is consistent. Another possibility to overcome this problem is proposed in [14]. The log-likelihood is first decomposed into independent summands and then importance sampling is applied. Papers [2] and [14] describe asymptotic properties of MCML estimators only for models with missing data. The goal of our paper is to consider models with intractable norming constants and explanatory variables. Sometimes a method applicable in one of these models cannot be used in the other. For instance, in missing data models there exists a powerful tool for computing maximum likelihood estimates, namely the EM (expectation-maximization) algorithm [3], [5], [9], [16]. However, this procedure cannot be applied to models with intractable norming constants and observable covariates. In the current paper we adapt the MCML method to our setting. Applying argumentation similar to [14] to our model is not straightforward.

In the paper we consider a parametric model with covariates

$$p(y|x,\theta)=\frac{1}{C(x,\theta)}f(y|x,\theta),$$

where $y\in\mathcal{Y}\subset\mathbb{R}^d$ is a response variable, $x\in\mathcal{X}\subset\mathbb{R}^l$ is a covariate or "explanatory" variable (random or deterministic), $\theta\in\mathbb{R}^p$ is a parameter describing the relation between y and x. The norming constant,

$$C(x,\theta)=\int f(y|x,\theta)dy,$$

is difficult or intractable.

Assume that the data consist of n independent observations $(Y_1,X_1),\ldots,(Y_n,X_n)$. If we regard covariates as random, then we assume that these pairs form an i.i.d. sample from a joint distribution with a density $g(y,x)$. Alternatively, x_i can be regarded as deterministic and then we assume that the random variable Y_i has a probability distribution g_i which depends on x_i. Both cases can be analysed very similarly. For simplicity we focus attention on the model with random covariates. It is not necessary to assume that $g(y|x)=p(y|x,\theta_0)$ for some θ_0. The case when no such θ_0 exists, i.e. the model is misspecified, makes the considerations only slightly more difficult. Thus, let us consider the following log-likelihood:

$$\begin{aligned}(1.1)\qquad \ell_n(\theta)&=\log p(Y_1,\ldots,Y_n|X_1,\ldots,X_n,\theta)\\&=\sum_{i=1}^n\log f(Y_i|X_i,\theta)-\sum_{i=1}^n\log C(X_i,\theta).\end{aligned}$$

The first term in the last line of (1.1) is easy to compute while the second one is approximated by a Monte Carlo (MC) method. Let $h(y)$ be an importance sampling (instrumental) distribution and note that

$$C(x,\theta) = \int f(y|x,\theta)dy = \int \frac{f(y|x,\theta)}{h(y)}h(y)dy = \mathbb{E}_{Y\sim h}\frac{f(Y|x,\theta)}{h(Y)}.$$

Thus, an MC approximation of the log-likelihood $\ell_n(\theta)$ is

$$\ell_n^m(\theta) = \sum_{i=1}^{n} \log f(Y_i|X_i,\theta) - \sum_{i=1}^{n} \log C_m(X_i,\theta), \tag{1.2}$$

where

$$C_m(x,\theta) = \frac{1}{m}\sum_{k=1}^{m} \frac{f(Y^k|x,\theta)}{h(Y^k)},$$

and $Y^1,\ldots,Y^m$ is a sample drawn from h.

Let us note that the general Monte Carlo recipe can also lead to approximation schemes different from (1.2). For instance, we could generate n independent MC samples instead of one, i.e. $Y_i^1,\ldots,Y_i^m \sim h_i, i = 1,\ldots,n$, and use the ith sample to approximate $C(x_i,\theta)$. Using this scenario, one can obtain estimators with better convergence rates, but at the cost of increased computational complexity. Another scheme, proposed in [2], approximates the log-likelihood by

$$\sum_{i=1}^{n} \log f(Y_i|X_i,\theta) - \log \frac{1}{m}\sum_{k=1}^{m}\prod_{i=1}^{n} \frac{f(Y_i^k|X_i,\theta)}{h_i(Y_i^k)}. \tag{1.3}$$

However, this scheme leads to estimators with unsatisfactory asymptotics unless a preliminary estimator is used. Thus, we focus our attention only on (1.2).

Let $\hat{\theta}_n$ be a maximizer of $\ell_n(\theta)$ (a genuine maximum likelihood estimator). It is well known that under some regularity assumptions (see [13], [15])

$$\hat{\theta}_n \sim_{\text{approx.}} \mathcal{N}\left(\theta_*, \frac{1}{n}D^{-1}VD^{-1}\right),$$

where θ_* is a maximizer of $\mathbb{E}_{(Y,X)\sim g} \log p(Y|X,\theta)$, i.e. the Kullback–Leibler projection, $D = \mathbb{E}_{(Y,X)\sim g}\nabla^2 \log p(Y|X,\theta_*)$ and $V = \mathrm{VAR}_{(Y,X)\sim g}\nabla \log p(Y|X,\theta_*)$. Symbols ∇ and ∇^2 denote derivatives with respect to θ, and VAR stands for the variance-covariance matrix. In the main result of the current paper (Theorem 3.1) we prove that the maximizer of (1.2), denoted by $\hat{\theta}_n^m$, satisfies

$$\hat{\theta}_n^m \sim_{\text{approx.}} \mathcal{N}\left(\theta_*, D^{-1}\left(\frac{V}{n} + \frac{W}{m}\right)D^{-1}\right), \tag{1.4}$$

where the matrix W will be given later. Formula (1.4) means that the estimator $\hat{\theta}_n^m$ behaves like a normal vector with the mean θ_* when both the initial sample size n and the Monte Carlo sample size m are large. Note that the first component of the asymptotic variance in (1.4), $D^{-1}VD^{-1}/n$, is the same as the asymptotic variance of the maximum likelihood estimator $\hat{\theta}_n$. The second component, $D^{-1}WD^{-1}/m$, is due to Monte Carlo randomness. Furthermore, if m is large, then asymptotic behaviour of $\hat{\theta}_n^m$ and $\hat{\theta}_n$ is similar. If the model is correctly specified, that is, $g(y|x) = p(y|x, \theta_0)$ for some θ_0, then $\theta_* = \theta_0$ and $D = -V$ (under standard assumptions on passing the derivative under the integral sign).

The choice of the instrumental distribution h affects W and thus the asymptotic efficiency of MCML. In [10], equation (2.11), a formula for optimal h is derived (this h minimizes the trace of W in a model without covariates). This result may be of some theoretical interest, but has a limited practical value because the optimal h can be very difficult to sample from. On the other hand, a more practical approach, suggested by several authors, e.g. [2], [18], is to select some distribution in the underlying parametric family, i.e. to put

$$h(y) = p(y|\psi) = \frac{1}{C(\psi)} f(y|\psi)$$

for some fixed $\psi \in \mathbb{R}^p$ (here we restrict attention to models without covariates). It is natural to guess that a "good choice" of ψ should be close to the target, θ_*. Since θ_* is unknown, one can use a preliminary estimator. Such a choice of h is recommended in [2], [18]. In the first of the cited papers, theoretical results are given which justify using a consistent preliminary estimate of θ_* as ψ, compare [2], Theorems 4 and 7. However, the results are about sampling scheme (1.3). In [18], the sampling scheme (1.2) is considered and the choice of ψ near θ_* is recommended on heuristical grounds. In fact, the intuition behind this choice turns out to be wrong, as demonstrated by the following toy example.

EXAMPLE 1.1. Let $\mathcal{Y} = \{0, 1\}$ and $f(y|\theta) = e^{\theta y}$ for $\theta \in \mathbb{R}$. Of course, the norming constant $C(\theta) = 1 + e^{\theta}$ is easy and there is no need to apply MCML, but the simplicity of this model will allow us to clearly illustrate our point. Assume we have an i.i.d. sample $Y_1, \ldots, Y_n$ from $f(\cdot|\theta_*)/C(\theta_*)$. The ML estimator is $\hat{\theta}_n = \log\big(\bar{Y}_n/(1 - \bar{Y}_n)\big)$, where $\bar{Y}_n = n^{-1}\sum_{i=1}^n Y_i$. Now suppose that we use the MCML approximation (1.2) with $h(y) = f(y|\psi)/C(\psi)$. It can be easily shown that the asymptotic variance W (now a scalar) is the minimum for $\psi_* = 0$, and not for $\psi = \theta_*$! The following direct derivation explains this fact. The formula (1.2) now takes the form

$$\ell_n^m(\theta) = n\theta\bar{Y}_n - n\log\left(\frac{1}{m}\sum_{k=1}^m e^{(\theta-\psi)Y^k}\right) - n\log C(\psi).$$

On noting that

$$\frac{1}{m}\sum_{k=1}^m Y^k e^{(\theta-\psi)Y^k} = \bar{Y}^m e^{\theta-\psi}, \quad \frac{1}{m}\sum_{k=1}^m e^{(\theta-\psi)Y^k} = \bar{Y}^m e^{\theta-\psi} + (1 - \bar{Y}^m),$$

we see that the equation $\nabla \ell_n^m(\theta) = 0$ is equivalent to

$$\bar{Y}_n - \frac{\bar{Y}^m e^{\theta-\psi}}{\bar{Y}^m e^{\theta-\psi} + (1-\bar{Y}^m)} = 0.$$

After elementary computations we see that the solution $\hat{\theta}_n^m$ of this equation is

$$\hat{\theta}_n^m = \log \frac{\bar{Y}_n}{1-\bar{Y}_n} + \psi - \log \frac{\bar{Y}^m}{1-\bar{Y}^m}.$$

Let us rewrite this expression as

$$\hat{\theta}_n^m = \hat{\theta}_n + \psi - \hat{\psi}^m,$$

where the term $\hat{\psi}^m$ is an ML estimate of ψ based on the MC sample. It is clear that $\sqrt{m}(\psi - \hat{\psi}^m) \to_{\mathrm{d}} \mathcal{N}\big(0, e^{-\psi}(1+e^{\psi})^2\big)$, independently of θ. The asymptotic variance of the MC error is the minimum for $\psi_* = 0$. The overall error of MCML is the sum of two independent terms $(\hat{\theta}_n - \theta_*) + (\hat{\psi}^m - \psi_*)$.

Asymptotic properties of MCML estimator (consistency, rates of convergence, asymptotic normality) can be obtained using standard statistical methods from the empirical processes theory [13], [15]. However, these tools should be adjusted to the model with double randomness when both sample sizes n and m tend to infinity simultaneously. This adaptation makes our proofs very arduous and technical despite the fact that the main ideas are rather clear. Therefore, to make the paper more transparent, we present only the proof of asymptotic normality. This result is the most important from a practical point of view. Moreover, the argumentation used in proving this property well illustrates how to adapt standard methods to the double randomness setup. Similar adaptation can be used to obtain consistency and the rate of convergence of the MCML estimator. Since the proof of (1.4) for the model with covariates is rather complicated, we begin in Section 2 with a model without covariates and state Theorem 2.1. It is extended to the general case (Theorem 3.1) in Section 3.

As we have already mentioned, related results on MCML for missing data models can be found in [2], [14]. In particular, our theorems are of similar form to those in [14]. However, models with intractable norming constants and observable covariates, considered in our paper, are more difficult to investigate. Namely, we have to analyse norming constants $C(x, \theta)$ as well as their derivatives $\nabla C(x, \theta)$ and $\nabla^2 C(x, \theta)$, which depend on the parameter θ and the covariate x. To do it, we need some additional assumption (compared to [14]). We discuss it in detail in Remark 3.1 in Section 3.

Note also that in the proof of Theorem 2.3 in [14], the authors used arduous and complicated argumentation relating to weak convergence of stochastic processes and its properties. We are able to give a proof of Theorem 3.1 basing only on elementary methods, if the estimator satisfies an additional and non-restrictive assumption. We explain this idea in Section 3.

2. MODEL WITHOUT COVARIATES

First, we consider a model without covariates

$$p(y|\theta) = \frac{1}{C(\theta)} f(y|\theta)$$

with an intractable norming constant $C(\theta) = \int f(y|\theta)dy$. Assume we have an i.i.d. sample $Y_1, \ldots, Y_n \sim g(y)$. Similarly to the general case, we allow for misspecification of the model, i.e. we do not assume $g(y) = p(y|\theta_0)$ for some θ_0. In what follows, θ_* is a maximizer of $\mathbb{E}_{Y\sim g} \log p(Y|\theta)$, i.e. the Kullback–Leibler projection. The MC approximation (1.2) multiplied by $\frac{1}{n}$ (denoted by $\bar{\ell}_n^m(\theta)$) takes the form

$$\bar{\ell}_n^m(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(Y_i|\theta) - \log \frac{1}{m}\sum_{k=1}^{m} \frac{f(Y^k|\theta)}{h(Y^k)} = \bar{\ell}_n(\theta) - r^m(\theta),$$

where

$$\bar{\ell}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} [\log f(Y_i|\theta) - \log C(\theta)],$$
$$r^m(\theta) = \log \frac{1}{m}\sum_{k=1}^{m} \frac{f(Y^k|\theta)}{h(Y^k)} - \log C(\theta).$$

Now we can state the main result of this section. We often refer to the proof of this theorem while proving the main result of this paper (Theorem 3.1 in Section 3).

THEOREM 2.1. *For some $\delta > 0$ let $U = \{\theta : |\theta - \theta_*| \leqslant \delta\}$ be a neighbourhood of θ_*. Suppose the following assumptions are satisfied*:

1. *second partial derivatives of $f(y|\theta)$ with respect to θ exist and are continuous for all y, and can be passed under the integral sign in $\int f(y|\theta)dy$*;
2. $\sqrt{\min(n,m)}(\hat{\theta}_n^m - \theta_*) = O_p(1)$;
3. *matrices*

$$V = \mathrm{VAR}_{Y\sim g} \nabla \log p(Y|\theta_*), \qquad D = \mathbb{E}_{Y\sim g} \nabla^2 \log p(Y|\theta_*),$$
$$W = \frac{1}{C^2(\theta_*)} \mathrm{VAR}_{Y\sim h} \left[\frac{\nabla f(Y|\theta_*)}{h(Y)} - \frac{\nabla C(\theta_*)}{C(\theta_*)} \frac{f(Y|\theta_*)}{h(Y)} \right]$$

exist, *and D is negative definite*;

4. *the function $D(\theta) = \mathbb{E}_{Y\sim g} \nabla^2 \log p(Y|\theta)$ is continuous at θ_**;
5. $\sup_{\theta\in U} |\nabla^2 \bar{\ell}_n(\theta) - D(\theta)| \to_{\mathrm{p}} 0,\ n \to \infty$;
6. $\sup_{\theta\in U} |\nabla^2 C_m(\theta) - \nabla^2 C(\theta)| \to_{\mathrm{p}} 0,\ m \to \infty$.

Then

$$\left(\frac{V}{n} + \frac{W}{m}\right)^{-1/2} D(\hat{\theta}_n^m - \theta_*) \to_{\mathrm{d}} \mathcal{N}(0, I), \quad n, m \to \infty.$$

Note that 1 and 3 are rather standard regularity assumptions. Condition 2 stipulates the square root consistency of the MCML estimator. It is automatically fulfilled if $\bar{\ell}_n^m(\theta)$ is concave ([12], Theorem 4), in particular for exponential families, i.e. if

$$f(y|\theta) = \exp\big(\theta^T W(y)\big)$$

with $W(y) = \big(W_1(y), \ldots, W_p(y)\big)$. Besides, we show that conditions 4–6 are also satisfied in this example. We can easily see that $\nabla^2 \log p(y|\theta) = -\nabla^2 \log C(\theta)$, so assumptions 4 and 5 are obviously fulfilled. Thus, condition 6 is the last one to establish. The function $\nabla^2 C_m(\theta)$ is matrix-valued, so it is enough to prove that for each component (that is, for each $r, s = 1, \ldots, p$)

$$\sup_{\theta \in U} \big|[\nabla^2 C_m(\theta)]_{rs} - [\nabla^2 C(\theta)]_{rs}\big| \to_{\mathrm{p}} 0, \quad m \to \infty. \tag{2.1}$$

Consider a family of functions

$$\left\{ \left[\frac{\nabla^2 f(y|\theta)}{h(y)}\right]_{rs} = \exp\big(\theta^T W(y)\big)\, \frac{W_r(y) W_s(y)}{h(y)} : \theta \in U \right\}. \tag{2.2}$$

The set U is compact, so to obtain (2.1) it is sufficient to assume that functions in (2.2) are dominated by an integrable function (see [4], Theorem 16(a), and [15], Example 19.8), i.e. for each r, s there is a function η such that $\mathbb{E}_{Y \sim h}\eta(Y) < \infty$ and $\big|[\nabla^2 f(y|\theta)/h(y)]_{rs}\big| \leqslant \eta(y)$ for each θ, y.

Proof of Theorem 2.1. Without loss of generality we can assume that $\theta_* = 0$. First we assume that $\frac{n}{n+m} \to a$ and consider three cases corresponding to rates at which n and m go to infinity: $0 < a < 1$, $a = 0$ and $a = 1$. Once our theorem is proved in these three special cases, standard application of the subsequence principle shows that it is valid in general (for $n \to \infty$ and $m \to \infty$ at arbitrary rates).

We begin with the case $0 < a < 1$. It is well known (see [13], Theorem VII.5) that we need to prove

$$\left(\frac{V}{n} + \frac{W}{m}\right)^{-1/2} \nabla \bar{\ell}_n^m(0) \to_{\mathrm{d}} \mathcal{N}(0, I), \quad n, m \to \infty, \tag{2.3}$$

and for every $M > 0$

$$(n+m) \sup_{|\theta| \leqslant M/\sqrt{n+m}} \left| \bar{\ell}_n^m(\theta) - \bar{\ell}_n^m(0) - \theta^T \nabla \bar{\ell}_n^m(0) - \frac{1}{2}\theta^T D \theta \right| \to_{\mathrm{p}} 0, \quad n, m \to \infty. \tag{2.4}$$

To obtain (2.3) notice that

$$\sqrt{n+m}\, \nabla \bar{\ell}_n^m(0) = \sqrt{\frac{n+m}{n}}\, \sqrt{n}\, \nabla \bar{\ell}_n(0) - \sqrt{\frac{n+m}{m}}\, \sqrt{m}\, \nabla r^m(0) \tag{2.5}$$

and the terms on the right-hand side in (2.5) are independent. We can calculate the gradient

$$\nabla r^m(0) = \frac{\frac{1}{m}\sum_{k=1}^{m}\left[\frac{\nabla f(Y^k|0)}{h(Y^k)} - \frac{\nabla C(0)}{C(0)}\frac{f(Y^k|0)}{h(Y^k)}\right]}{\frac{1}{m}\sum_{k=1}^{m}\frac{f(Y^k|0)}{h(Y^k)}}.$$

By the LLN, CLT and Slutsky's theorem we have $\sqrt{m}\nabla r^m(0) \to_{\mathrm{d}} \mathcal{N}(0,W)$ and $\sqrt{n}\nabla\bar{\ell}_n(0) \to_{\mathrm{d}} \mathcal{N}(0,V)$, which implies

$$\sqrt{n+m}\nabla\bar{\ell}_n^m(0) \to_{\mathrm{d}} \mathcal{N}\big(0, V/a + W/(1-a)\big), \quad n,m\to\infty.$$

Thus, we obtain (2.3) since

$$\sqrt{n+m}\big(V/a + W/(1-a)\big)^{-1/2}\left(V/n + W/m\right)^{1/2} \to I, \quad n,m\to\infty.$$

Now we focus on (2.4). Using the Taylor expansion, the left-hand side of (2.4) can be bounded by

$$\text{(2.6)} \quad \frac{M^2}{2}\Big(\sup_{\theta\in U_n^m}|\nabla^2\bar{\ell}_n(\theta) - D(\theta)| + \sup_{\theta\in U_n^m}|D(\theta) - D(0)| + \sup_{\theta\in U_n^m}|\nabla^2 r^m(\theta)|\Big)$$

for $U_n^m = \{\theta : |\theta| \leqslant M/\sqrt{n+m}\}$. The first two terms in (2.6) tend to zero in probability by assumptions 4 and 5. We prove that assumption 6 implies convergence to zero in probability of the third term in (2.6). Calculating the second derivative, we get

$$\nabla^2 r^m(\theta) = \frac{\nabla^2 C_m(\theta)}{C_m(\theta)} - \frac{\nabla C_m(\theta)\nabla^T C_m(\theta)}{C_m^2(\theta)} - \frac{\nabla^2 C(\theta)}{C(\theta)} + \frac{\nabla C(\theta)\nabla^T C(\theta)}{C^2(\theta)}.$$

Therefore,

$$\begin{aligned}\text{(2.7)} \quad \sup_{\theta\in U}|\nabla^2 r^m(\theta)| &\leqslant \sup_{\theta\in U}\frac{|\nabla^2 C_m(\theta)|\,|C_m(\theta) - C(\theta)|}{C_m(\theta)C(\theta)} \\ &\quad + \sup_{\theta\in U}\frac{|\nabla^2 C_m(\theta) - \nabla^2 C(\theta)|}{C(\theta)} + \sup_{\theta\in U}\frac{|\nabla C_m(\theta)|^2\,|C_m^2(\theta) - C^2(\theta)|}{C_m^2(\theta)C^2(\theta)} \\ &\quad + \sup_{\theta\in U}\frac{|\nabla C_m(\theta)\nabla^T C_m(\theta) - \nabla C(\theta)\nabla^T C(\theta)|}{C^2(\theta)}.\end{aligned}$$

Note that continuous functions $C(\theta), |\nabla C(\theta)|, |\nabla^2 C(\theta)|$ are bounded on the compact set U, in particular, the function $C(\theta)$ is separated from zero, i.e. there exist positive constants α, K such that

$$\alpha \leqslant C(\theta) \leqslant K, \quad |\nabla C(\theta)| \leqslant K, \quad |\nabla^2 C(\theta)| \leqslant K$$

for every $\theta \in U$. Therefore, all we need is assumption 6 and

$$\sup_{\theta\in U} |C_m(\theta) - C(\theta)| \to_{\mathrm{p}} 0, \quad m \to \infty, \tag{2.8}$$

$$\sup_{\theta\in U} |\nabla C_m(\theta) - \nabla C(\theta)| \to_{\mathrm{p}} 0, \quad m \to \infty. \tag{2.9}$$

Indeed, by these conditions we infer that for arbitrary $\varepsilon, \eta > 0$ and sufficiently large m the events

$$\begin{aligned} |C_m(\theta) - C(\theta)| &\leqslant \varepsilon \quad \text{for all } \theta \in U, \\ |\nabla C_m(\theta) - \nabla C(\theta)| &\leqslant \varepsilon \quad \text{for all } \theta \in U, \\ |\nabla^2 C_m(\theta) - \nabla^2 C(\theta)| &\leqslant \varepsilon \quad \text{for all } \theta \in U \end{aligned}$$

have probability at least $1 - \eta$. Therefore, we get the following bounds for every $\theta \in U$:

$$\alpha/2 \leqslant C_m(\theta) \leqslant K + \alpha/2, \quad |\nabla C_m(\theta)| \leqslant K + \varepsilon, \quad |\nabla^2 C_m(\theta)| \leqslant K + \varepsilon$$

(to be precise, they hold with probability at least $1 - \eta$ if m is sufficiently large). Thus, we can prove that every expression that bounds $\sup_{\theta\in U} |\nabla^2 r^m(\theta)|$ in (2.7) is arbitrarily small, for instance, considering the first one on the right-hand side of (2.7), we have

$$\sup_{\theta\in U} \frac{|\nabla^2 C_m(\theta)|\, |C_m(\theta) - C(\theta)|}{C_m(\theta)C(\theta)} \leqslant \frac{2(K+\varepsilon)\varepsilon}{\alpha^2}.$$

Besides, notice that uniform convergence in (2.8) and (2.9), that we have just used, easily follows from the Taylor expansion, LLN and assumption 6. For instance, for some $\theta' \in (0, \theta)$

$$\nabla C_m(\theta) - \nabla C(\theta) = \nabla C_m(0) - \nabla C(0) + [\nabla^2 C_m(\theta') - \nabla^2 C(\theta')]\theta,$$

so

$$\sup_{\theta\in U} |\nabla C_m(\theta) - \nabla C(\theta)| \leqslant |\nabla C_m(0) - \nabla C(0)| + \delta \sup_{\theta\in U} |\nabla^2 C_m(\theta) - \nabla^2 C(\theta)|.$$

Thus, the proof in the case $0 < a < 1$ is completed. For $a = 0$ or $a = 1$ we proceed similarly. For example, if $a = 0$, then we should prove an analog of (2.4), namely, for every $M > 0$

$$n \sup_{|\theta| \leqslant M/\sqrt{n}} \left| \bar{\ell}_n^m(\theta) - \bar{\ell}_n^m(0) - \theta^T \nabla \bar{\ell}_n^m(0) - \frac{1}{2}\theta^T D\theta \right| \to_{\mathrm{p}} 0, \quad n, m \to \infty.$$

Argumentation is almost the same as in the proof of (2.4). To obtain (2.3) in this case note that

$$\sqrt{n}\nabla\bar{\ell}_n^m(0) = \sqrt{n}\nabla\bar{\ell}_n(0) - \sqrt{\frac{n}{m}}\,\sqrt{m}\nabla r^m(0). \tag{2.10}$$

Therefore, the expression (2.10) tends in distribution to $\mathcal{N}(0, V)$. Moreover,

$$\sqrt{n}V^{-1/2}\,(V/n + W/m)^{1/2} \to I, \quad n, m \to \infty,$$

which completes the proof. ■

3. MODEL WITH COVARIATES

Let us return to the general case and state the main theorem of the paper. We need the following new notation:

$$\phi(y|x) = \left[\frac{\nabla f(y|x,\theta_*)}{h(y)} - \frac{\nabla C(x,\theta_*)}{C(x,\theta_*)}\frac{f(y|x,\theta_*)}{h(y)}\right]\frac{1}{C(x,\theta_*)},$$

$$r_n^m(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left[\log\frac{1}{m}\sum_{k=1}^{m}\frac{f(Y^k|X_i,\theta)}{h(Y^k)} - \log C(X_i,\theta)\right].$$

Then $\bar{\ell}_n^m(\theta) = \bar{\ell}_n(\theta) - r_n^m(\theta)$.

THEOREM 3.1. *For some $\delta > 0$ let $U = \{\theta : |\theta - \theta_*| \leqslant \delta\}$ be a neighbourhood of θ_*. Suppose the following assumptions are satisfied*:

1. *second partial derivatives of $f(y|x,\theta)$ with respect to θ exist and are continuous for all y and x, and may be passed under the integral sign in $\int f(y|x,\theta)dy$ for fixed x*;
2. $\sqrt{\min(n,m)}(\hat{\theta}_n^m - \theta_*) = O_p(1)$;
3. *matrices*

$$V = \mathrm{VAR}_{(Y,X)\sim g}\nabla\log p(Y|X,\theta_*), \qquad D = \mathbb{E}_{(Y,X)\sim g}\nabla^2\log p(Y|X,\theta_*)$$

and the expectation $\tilde{W} = \mathbb{E}_{Y\sim h, X\sim g}\big|\phi(Y|X)\big|^2$ exist, and D is negative definite;

4. *the function $D(\theta) = \mathbb{E}_{(Y,X)\sim g}\nabla^2\log p(Y|X,\theta)$ is continuous at θ_**;
5. $\sup_{\theta\in U}|\nabla^2\bar{\ell}_n(\theta) - D(\theta)| \to_{\mathrm{p}} 0,\ n \to \infty$;
6. *it follows that*:

(a) $\sup\limits_{x\in\mathcal{X}}|C_m(x,\theta_*) - C(x,\theta_*)| \to_{\mathrm{p}} 0,\ m \to \infty$;

(b) $\sup\limits_{x\in\mathcal{X}}|\nabla C_m(x,\theta_*) - \nabla C(x,\theta_*)| \to_{\mathrm{p}} 0,\ m \to \infty$;

(c) $\sup\limits_{\theta\in U, x\in\mathcal{X}}|\nabla^2 C_m(x,\theta) - \nabla^2 C(x,\theta)| \to_{\mathrm{p}} 0,\ m \to \infty$;

7. *there exist constants* $\alpha > 0$, $K > 0$ *such that for each* $x \in \mathcal{X}$ *and* $\theta \in U$

$$\alpha \leqslant C(x,\theta) \leqslant K, \quad |\nabla C(x,\theta)| \leqslant K, \quad |\nabla^2 C(x,\theta)| \leqslant K.$$

Then the matrix

$$W = \mathrm{VAR}_{Y\sim h}\, \mathbb{E}_{X\sim g}\, \phi(Y|X)$$

is finite and

$$\left(\frac{V}{n} + \frac{W}{m}\right)^{-1/2} D(\hat{\theta}_n^m - \theta_*) \to_{\mathrm{d}} \mathcal{N}(0,I), \quad n,m \to \infty.$$

Some comments on the relation of Theorem 3.1 to the results in [14] are given in Remark 3.1 after the proof of this theorem.

Proof of Theorem 3.1. Without loss of generality we can assume that $\theta_* = 0$.

Similarly to the proof of Theorem 2.1 we consider three cases: $0 < a < 1$, $a = 0$ and $a = 1$, where $\frac{n}{n+m} \to a$. Finally, we complete the proof by using the subsequence principle.

We focus on the case $0 < a < 1$ because for $a = 0$ or $a = 1$ we proceed in a similar way (cf. the proof of Theorem 2.1). It is well known (see [13], Theorem VII.5) that we need to prove that for every $M > 0$

(3.1)
$$(n+m) \sup_{|\theta|\leqslant M/\sqrt{n+m}} \left| \bar{\ell}_n^m(\theta) - \bar{\ell}_n^m(0) - \theta^T \nabla \bar{\ell}_n^m(0) - \frac{1}{2}\theta^T D\theta \right| \to_{\mathrm{p}} 0, \quad n,m \to \infty,$$

and

$$\left(\frac{V}{n} + \frac{W}{m}\right)^{-1/2} \nabla \bar{\ell}_n^m(0) \to_d \mathcal{N}(0,I), \quad n,m \to \infty. \tag{3.2}$$

We start with (3.1). Using the Taylor expansion, the left-hand side of (3.1) can be bounded by

$$\frac{M^2}{2}\big(\sup_{\theta\in U_n^m} |\nabla^2 \bar{\ell}_n(\theta) - D(\theta)| + \sup_{\theta\in U_n^m} |D(\theta) - D(0)| + \sup_{\theta\in U_n^m} |\nabla^2 r_n^m(\theta)|\big) \tag{3.3}$$

for $U_n^m = \{\theta : |\theta| \leqslant M/\sqrt{n+m}\}$. The first two terms in (3.3) tend to zero in probability by assumptions 4 and 5. We prove that assumptions 6 and 7 imply convergence to zero in probability of the third term in (3.3). Calculating the second derivative of $r_n^m(\theta)$, we get

$$\nabla^2 r_n^m(\theta) = \frac{1}{n}\sum_{i=1}^n \left[\frac{\nabla^2 C_m(X_i,\theta)}{C_m(X_i,\theta)} - \frac{\nabla C_m(X_i,\theta)\nabla^T C_m(X_i,\theta)}{C_m^2(X_i,\theta)} \right.$$
$$\left. - \frac{\nabla^2 C(X_i,\theta)}{C(X_i,\theta)} + \frac{\nabla C(X_i,\theta)\nabla^T C(X_i,\theta)}{C^2(X_i,\theta)} \right].$$

Therefore,

$$\begin{aligned}
\text{(3.4)}\quad \sup_{\theta\in U}|\nabla^2 r_n^m(\theta)| \leqslant{} & \sup_{\theta\in U, x\in\mathcal{X}} \frac{|\nabla^2 C_m(x,\theta)|\,|C_m(x,\theta)-C(x,\theta)|}{C_m(x,\theta)C(x,\theta)}\\
&+\sup_{\theta\in U, x\in\mathcal{X}} \frac{|\nabla^2 C_m(x,\theta)-\nabla^2 C(x,\theta)|}{C(x,\theta)}\\
&+\sup_{\theta\in U, x\in\mathcal{X}} \frac{|\nabla C_m(x,\theta)|^2\,|C_m^2(x,\theta)-C^2(x,\theta)|}{C_m^2(x,\theta)C^2(x,\theta)}\\
&+\sup_{\theta\in U, x\in\mathcal{X}} \frac{|\nabla C_m(x,\theta)\nabla^T C_m(x,\theta)-\nabla C(x,\theta)\nabla^T C(x,\theta)|}{C^2(x,\theta)}.
\end{aligned}$$

To prove that every term on the right-hand side of (3.4) tends to zero in probability, we proceed analogously to the proof of Theorem 2.1. Namely, we strengthen convergence in assumptions 6(a) and 6(b) to hold uniformly over $\theta\in U$. Then, using these arguments, assumptions 6(c) and 7, we obtain uniform bounds (over $\theta\in U$ and $x\in\mathcal{X}$) for $C_m(x,\theta)$, $|\nabla C_m(x,\theta)|$ and $|\nabla^2 C_m(x,\theta)|$ that hold with probability at least $1-\eta$ if m is sufficiently large. Hence, the same reasoning as in the analogous part of the proof of Theorem 2.1 gives convergence of every expression on the right-hand side of (3.4) to zero in probability.

The last step is proving (3.2). First, notice that

$$\begin{aligned}
\sqrt{n+m}\nabla\bar{\ell}_n^m(0) &= \sqrt{\frac{n+m}{n}}\,\sqrt{n}\nabla\bar{\ell}_n(0) - \sqrt{\frac{n+m}{m}}\,\sqrt{m}\nabla r_n^m(0)\\
&= A_1 + A_2,
\end{aligned}$$

where

$$\text{(3.5)}\qquad A_1 = \left[\sqrt{\frac{n+m}{n}}\,\sqrt{n}\nabla\bar{\ell}_n(0) - \sqrt{\frac{n+m}{m}}\frac{1}{\sqrt{m}}\sum_{k=1}^m \bar{\phi}(Y^k)\right],$$

$$\text{(3.6)}\qquad A_2 = \sqrt{\frac{n+m}{m}}\left[\frac{1}{\sqrt{m}}\sum_{k=1}^m \bar{\phi}(Y^k) - \sqrt{m}\nabla r_n^m(0)\right],$$

and $\bar{\phi}(y)=\mathbb{E}_{X\sim g}\phi(y|X)$. By the CLT, the expression (3.5) tends in distribution to $\mathcal{N}\big(0, V/a + W/(1-a)\big)$ since the Monte Carlo sample is independent of the observation. To show that the term (3.6) tends to zero in probability, we prove that

$$\text{(3.7)}\qquad \sqrt{m}\nabla r_n^m(0) - \frac{1}{n}\sum_{i=1}^n \frac{1}{\sqrt{m}}\sum_{k=1}^m \phi(Y^k|X_i)$$

and

$$\text{(3.8)}\qquad \frac{1}{n}\sum_{i=1}^n \frac{1}{\sqrt{m}}\sum_{k=1}^m \phi(Y^k|X_i) - \frac{1}{\sqrt{m}}\sum_{k=1}^m \bar{\phi}(Y^k)$$

tend to zero in probability. We start with (3.7) and calculate

$$\nabla r_n^m(0) = \frac{1}{n}\sum_{i=1}^{n} \frac{\frac{1}{m}\sum_{k=1}^{m} \phi(Y^k|X_i)\, C(X_i,0)}{C_m(X_i,0)}.$$

Therefore, by the Cauchy–Schwarz inequality, the expression (3.7) is bounded by

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n} \frac{[C_m(X_i,0)-C(X_i,0)]^2}{C_m^2(X_i,0)}}\,\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|\frac{1}{\sqrt{m}}\sum_{k=1}^{m}\phi(Y^k|X_i)\right|^2}. \tag{3.9}$$

By assumptions 6(a) and 7 we again infer that for arbitrary $\varepsilon > 0, \eta > 0$ and sufficiently large m with probability at least $1-\eta$ for every $x \in \mathcal{X}$

$$|C_m(x,0) - C(x,0)| \leqslant \varepsilon \quad \text{and} \quad C_m(x,0) \geqslant \alpha/2.$$

Therefore, the term under the first square root in (3.9) tends in probability to zero because with probability at least $1-\eta$

$$\frac{1}{n}\sum_{i=1}^{n}\frac{[C_m(X_i,0)-C(X_i,0)]^2}{C_m^2(X_i,0)} \leqslant \sup_{x\in\mathcal{X}} \frac{[C_m(x,0)-C(x,0)]^2}{C_m^2(x,0)} \leqslant \frac{4\varepsilon^2}{\alpha^2}$$

if m is sufficiently large. Using Markov's inequality and assumption 3, we see that the second square root in (3.9) is bounded in probability, since

$$\begin{aligned}\mathbb{E}_{X_i\sim g, Y^k\sim h}\frac{1}{n}\sum_{i=1}^{n}\left|\frac{1}{\sqrt{m}}\sum_{k=1}^{m}\phi(Y^k|X_i)\right|^2 &= \mathbb{E}_{X\sim g, Y^k\sim h}\left|\frac{1}{\sqrt{m}}\sum_{k=1}^{m}\phi(Y^k|X)\right|^2\\ &= \mathbb{E}_{X\sim g, Y\sim h}\left|\phi(Y|X)\right|^2 = \tilde{W} < \infty,\end{aligned}$$

where we use the fact that $\mathbb{E}_{Y\sim h}\phi(Y|x) = 0$ for fixed x.

Now consider (3.8). Change the order of summation and notice that

$$\begin{aligned}&\mathbb{E}_{X_i\sim g, Y^k\sim h}\left|\frac{1}{\sqrt{m}}\sum_{k=1}^{m}\left[\frac{1}{n}\sum_{i=1}^{n}\phi(Y^k|X_i) - \bar{\phi}(Y^k)\right]\right|^2\\ &= \mathbb{E}_{X_i\sim g, Y\sim h}\left|\frac{1}{n}\sum_{i=1}^{n}\phi(Y|X_i) - \bar{\phi}(Y)\right|^2 = \frac{1}{n}\mathbb{E}_{X\sim g, Y\sim h}\left|\phi(Y|X) - \bar{\phi}(Y)\right|^2,\end{aligned}$$

so (3.8) tends to zero in L^2, hence, in probability. ∎

REMARK 3.1. To explain the relation between the results in [14] concerning models with missing data and our results for models with intractable constants, we compare our methods and those in [14]. Describing the missing data model, we borrow the notation from [14].

In the missing data model one approximates the observed data likelihood, that is

$$\textstyle\int f_\theta(x,y)dx,$$

where $f_\theta(x,y)$ is the joint density of observed y and missing x. Therefore, one estimates the integral of the *density*. In the model with intractable constants and covariates, which is considered in this paper, one approximates

$$\textstyle\int f(y|x,\theta)dy,$$

i.e. the integral of a function that is not a density but only a *nonnegative* function. These facts have significant impact on the proofs of Theorem 2.3 in [14] and Theorem 3.1 in the current paper. Namely, in [14] one considers limits of sequences of random variables:

$$\frac{1}{m}\sum_{i=1}^{m}\frac{f_\theta(X_i|y)}{h(X_i)},\quad \frac{1}{m}\sum_{i=1}^{m}\frac{\nabla f_\theta(X_i|y)}{h(X_i)},\quad \frac{1}{m}\sum_{i=1}^{m}\frac{\nabla^2 f_\theta(X_i|y)}{h(X_i)}, \tag{3.10}$$

where $X_1,\ldots,X_m$ are i.i.d. from a density h, and y,θ are fixed. These limits are one, zero and zero, respectively, so they do not depend on y and θ. In our problem we consider the expressions

$$\frac{1}{m}\sum_{k=1}^{m}\frac{f(Y^k|x,\theta)}{h(Y^k)},\quad \frac{1}{m}\sum_{k=1}^{m}\frac{\nabla f(Y^k|x,\theta)}{h(Y^k)},\quad \frac{1}{m}\sum_{k=1}^{m}\frac{\nabla^2 f(Y^k|x,\theta)}{h(Y^k)}, \tag{3.11}$$

where $Y^1,\ldots,Y^m$ are i.i.d. from a density h, and x,θ are fixed. The limits of expressions in (3.11) are $C(x,\theta),\nabla C(x,\theta)$ and $\nabla^2 C(x,\theta)$, respectively. Thus, they do depend on x and θ. In fact, the proofs of Theorem 2.3 in [14] and our Theorem 3.1 are based on uniform convergence of quantities in (3.10) and (3.11). Thus, dependence on x,θ of limits of terms in (3.11) makes the proof of Theorem 3.1 more difficult than Theorem 2.3 in [14], because we also have to investigate functions $C(x,\theta),\nabla C(x,\theta)$ and $\nabla^2 C(x,\theta)$ in $x\in\mathcal{X}$ and $\theta\in U$. To do it, we need the additional condition (assumption 7) that helps us to control $C(x,\theta)$, $|\nabla C(x,\theta)|$ and $|\nabla^2 C(x,\theta)|$ uniformly over x and θ.

The considerable difference between [14] and our paper can be also found while comparing argumentation used to obtain the convergence

$$\left(\frac{V}{n}+\frac{W}{m}\right)^{-1/2}\nabla\bar{\ell}_n^m(0)\to_{\mathrm{d}}\mathcal{N}(0,I),\quad n,m\to\infty, \tag{3.12}$$

that is needed in the proof of Theorem 3.1 and its analog in the proof of Theorem 2.3 in [14]. Namely, to prove the analog of (3.12) in [14] one uses an arduous and complicated method based on weak convergence of stochastic processes and its properties (see [14], Lemma A.4). Our analysis relates only to elementary tools, for instance, the CLT. The price that we pay for this significant simplification of the proof is the additional assumption 2 in Theorem 3.1. However, this price is low, because assumption 2 is not restrictive. Indeed, it is automatically fulfilled if $\bar{\ell}_n^m(\theta)$ is concave in θ (see [12], Theorem 4). In particular, $\bar{\ell}_n^m(\theta)$ is concave for models with densities belonging to the exponential family, for instance, the autologistic model [8].

Note that in assumptions 5 and 6 in Theorem 2.1 and assumptions 5 and 6 in Theorem 3.1 we need uniform convergence in probability. In the corresponding conditions (4), (5) and (7) in Theorem 2.3 in [14] one uses almost sure convergence. Moreover, comparing assumption 6(b) in Theorem 3.1 to assumption (6) in Theorem 2.3 in [14], we see that the condition contained in the current paper is also weaker, because we replace the Donsker class by the Glivenko–Cantelli class (in probability).

Finally, we discuss assumptions in Theorem 3.1. Note that conditions 1–3 are similar to their analogs in Theorem 2.1. Therefore, we briefly comment on the others. Consider the exponential family with

$$f(y|x,\theta) = \exp\big(\theta^T W(y,x)\big),$$

where $W(y,x) = \big(W_1(y,x),\dots,W_p(y,x)\big)$, the set $\mathcal{X}$ is compact and the function $W(y,x)$ is continuous with respect to the variable x. For simplicity we restrict attention to a finite (but very large) space $\mathcal{Y}$, so that

$$C(x,\theta) = \sum_{y\in\mathcal{Y}} \exp\big(\theta^T W(y,x)\big).$$

The autologistic model [8] that is very popular in spatial statistics belongs to this family. We can calculate that

$$\begin{aligned}
\nabla C(x,\theta) &= \sum_{y\in\mathcal{Y}} \exp\big(\theta^T W(y,x)\big) W(y,x),\\
\nabla^2 C(x,\theta) &= \sum_{y\in\mathcal{Y}} \exp\big(\theta^T W(y,x)\big) W(y,x) W^T(y,x),\\
\nabla^2 \log p(y|x,\theta) = -\nabla^2 \log C(x,\theta) &= -\frac{\nabla^2 C(x,\theta)}{C(x,\theta)} + \frac{\nabla C(x,\theta)\nabla^T C(x,\theta)}{C^2(x,\theta)}.
\end{aligned}$$

Since the function $W(y,x)$ is continuous with respect to x, functions $C(x,\theta)$, $\nabla C(x,\theta), \nabla^2 C(x,\theta)$ are continuous with respect to both variables on the compact set $\mathcal{X}\times U$, so assumption 7 is satisfied. Besides, the function $\nabla^2 \log p(y|x,\theta)$ is also continuous, which implies condition 4 to be fulfilled. The uniform convergence in assumptions 5 and 6 follows from Theorem 16(a) in [4] or Example 19.8 in [15] if we again use compactness of sets $\mathcal{X}, U$ and continuity of considered functions.

REFERENCES

[1] J. Besag, *Spatial interaction and the statistical analysis of lattice systems*, J. R. Stat. Soc. Ser. B. Stat. Methodol. 36 (1974), pp. 192–236.

[2] O. Cappé, R. Douc, E. Moulines, and C. Robert, *On the convergence of the Monte Carlo maximum likelihood method for latent variable models*, Scand. J. Stat. 29 (2002), pp. 615–635.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, J. R. Stat. Soc. Ser. B. Stat. Methodol. 39 (1977), pp. 1–38.

[4] T. S. Ferguson, *A Course in Large Sample Theory*, Chapman and Hall, London 2010.

[5] G. Fort and E. Moulines, *Convergence of the Monte Carlo expectation maximization for curved exponential families*, Ann. Statist. 31 (2003), pp. 1220–1259.

[6] C. J. Geyer, *On the convergence of Monte Carlo maximum likelihood calculations*, J. R. Stat. Soc. Ser. B. Stat. Methodol. 56 (1994), pp. 261–274.

[7] C. J. Geyer and E. A. Thompson, *Constrained Monte Carlo maximum likelihood for dependent data*, J. R. Stat. Soc. Ser. B. Stat. Methodol. 54 (1992), pp. 657–699.

[8] F. W. Huffer and H. Wu, *Markov chain Monte Carlo for autologistic regression models with application to the distribution of plant species*, Biometrics 54 (1998), pp. 509–524.

[9] R. A. Levine and G. Casella, *Implementations of the Monte Carlo EM algorithm*, J. Comput. Graph. Statist. 10 (2001), pp. 422–439.

[10] B. Miasojedow, W. Niemiro, J. Palczewski, and W. Rejchel, *Adaptive Monte Carlo maximum likelihood*, in: *Challenges in Computational Statistics and Data Mining*, S. Matwin and J. Mielniczuk (Eds.), Stud. Comput. Intell., Vol. 605, Springer, 2016, pp. 247–270.

[11] J. Møller, A. N. Pettitt, R. Reeves, and K. K. Berthelsen, *An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants*, Biometrika 93 (2006), pp. 451–458.

[12] W. Niemiro, *Asymptotics for M-estimators defined by convex minimization*, Ann. Statist. 20 (1992), pp. 1514–1533.

[13] D. Pollard, *Convergence of Stochastic Processes*, Springer, New York 1984.

[14] Y. J. Sung and C. J. Geyer, *Monte Carlo likelihood inference for missing data models*, Ann. Statist. 35 (2007), pp. 990–1011.

[15] A. W. van der Vaart, *Asymptotic Statistics*, Cambridge University Press, Cambridge 1998.

[16] G. C. G. Wei and M. A. Tanner, *A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms*, J. Amer. Statist. Assoc. 85 (1990), pp. 699–704.

[17] H. Wu and F. W. Huffer, *Modeling the distribution of plant species using the autologistic regression model*, Environ. Ecol. Stat. 4 (1997), pp. 49–64.

[18] M. Zalewska, W. Niemiro, and B. Samoliński, *MCMC imputation in autologistic model*, Monte Carlo Methods Appl. 16 (2010), pp. 421–438.

Błażej Miasojedow
Institute of Applied Mathematics and Mechanics
University of Warsaw
Banacha 2
02-097 Warsaw, Poland
E-mail: B.Miasojedow@mimuw.edu.pl

Wojciech Niemiro
Institute of Applied Mathematics and Mechanics
University of Warsaw
Banacha 2
02-097 Warsaw, Poland
E-mail: wniem@mimuw.edu.pl

Jan Palczewski
School of Mathematics
University of Leeds
Woodhouse Lane
Leeds LS2 9JT, United Kingdom
E-mail: J.Palczewski@leeds.ac.uk

Wojciech Rejchel
Faculty of Mathematics and Computer Science
Nicolaus Copernicus University
Chopina 12/18
87-100 Toruń, Poland
E-mail: wrejchel@gmail.com

Received on 19.5.2015;
revised version on 27.1.2016