Institute of Mathematical Statistics

LECTURE NOTES — MONOGRAPH SERIES

EFFICIENCY OF THE PSEUDO-LIKELIHOOD ESTIMATE IN A ONE-DIMENSIONAL LATTICE GAS

J. L. Jensen
University of Aarhus
Denmark

Abstract

For a simple one-dimensional lattice gas we consider the efficiency properties of the maximum pseudo-likelihood estimate. We show that the pseudo-likelihood estimating function is not optimal within a natural class of estimating functions, although numerical investigations show that it is very close to being optimal. We also show that the pseudo-likelihood is far from being efficient when there is strong dependence in the model.

Key words: Efficiency; estimating function; Gibbs model; pseudo-likelihood.

1 Introduction

In the field of stochastic processes it is often not possible to give the likelihood function in an explicit form. Instead one uses estimating functions, and it seems natural to look for an optimal estimating function within a class of such functions. A theory for this has been developed in Heyde (1988). For martingale estimating functions an application of these ideas can be found in Bibby and Sørensen (1996).

In this paper we will try to use these ideas in the setting of Gibbs lattice models. Such models are defined through interactions between neighbouring points and typically there is a norming constant in the distribution that cannot be calculated explicitly. For the lattice $\mathbf{Z}^d$, $d > 1$, there is also the possibility of phase transitions and the maximum likelihood estimate need not be asymptotically normally distributed. Due to these problems other estimating procedures have been considered. Besag (1975) introduced the pseudo-likelihood function, which only uses local conditional distributions. It has been shown recently (Guyon and Künsch 1992; Jensen and Künsch 1994) that the maximum pseudo-likelihood estimate admits a random norming so that the limiting distribution is normal. The efficiency of the maximum pseudo-likelihood estimate seems largely not to have been investigated. The

paper of Guyon and Künsch (1992) contains a comparison of five estimators, including the pseudo-likelihood estimator, for a model slightly different from the model considered in this paper. What we propose to do here is to view the pseudo-likelihood estimating function as one in a class of estimating functions, and then to find the optimal choice within this class.

The term pseudo-likelihood is here adopted from Besag (1975) although the contrast function used has no direct relation to a proper likelihood function. Also the reference to martingale estimating functions (Heyde, 1988) is indirect. The estimating function here is not a martingale, but it has the property that the individual terms have conditional mean zero, where the conditioning set consists of all but one variable. As in the martingale case this property ensures the consistency of the estimate. However, the conditional-mean-zero property is not something that is shared with the true likelihood.

To be able to perform all the calculations we will only consider a one-dimensional Gibbs model. We can then find the optimal estimating function within the class considered, and it turns out that the pseudo-likelihood is not optimal. However, for the model considered, the maximum pseudo-likelihood estimate is very close to being optimal. For the simple model considered it is also possible to compare the pseudo-likelihood with the true likelihood. When the interaction in the model is strong we find that the efficiency of the pseudo-likelihood estimate is very poor. Furthermore, an attempt to improve on this by extending the pseudo-likelihood idea turns out to give only a minor improvement. It is an open problem whether these conclusions carry over to higher dimensions.

2 The Gibbs Model

Let $X_i \in \{-1, 1\}$, $i \in \mathbf{Z}$, be a lattice gas, where X_i interacts with the four nearest neighbours $(X_{i-2}, X_{i-1}, X_{i+1}, X_{i+2})$. The conditional specifications are given by

$$P(X_i = x_i \mid (X_{i-2}, X_{i-1}, X_{i+1}, X_{i+2}) = (x_{i-2}, x_{i-1}, x_{i+1}, x_{i+2})) \tag{2.1}$$

$$= \{2\cosh[\beta(x_{i-2} + x_{i-1} + x_{i-1} + x_{i+2})]\}^{-1} \exp\{\beta x_i(x_{i-2} + x_{i-1} + x_{i+1} + x_{i+2})\}$$

and for $l > k$

$$P\left((X_k, X_{k+1}, ..., X_l) = (x_k, x_{k+1}, ..., x_l) \,\middle|\, \begin{array}{c}(X_{k-2}, X_{k-1}, X_{l+1}, X_{l+2}) = \\ (x_{k-2}, x_{k-1}, x_{l+1}, x_{l+2})\end{array}\right) \tag{2.2}$$

$$= Z_{l-k}(\beta; x_{k-2}, x_{k-1}, x_{l+1}, x_{l+2})^{-1} \exp\left\{\beta\left[\sum_{i=k}^{l+1} x_i(x_{i-1} + x_{i-2}) + x_{l+2}x_l\right]\right\}$$

With $n = l-k+1$ the evaluation of Z_{l-k} involves a sum with 2^n terms, and so for n large it is not feasible to evaluate Z_{l-k}. Instead of using the likelihood function for estimating β we use the pseudo-likelihood function $\exp(pl_n(\beta))$. The latter is given as the product of the conditional densities in (2.1), see Besag (1975). Let the observations be $x_{-1}, x_0, x_1, \ldots, x_n, x_{n+1}, x_{n+2}$, then

$$pl_n(\beta) = \sum_{i=1}^{n} \{\beta x_i s_i - \log[2\cosh(\beta s_i)]\},$$

where

$$s_i = x_{i-2} + x_{i-1} + x_{i+1} + x_{i+2}.$$

From this we find

$$pl_n'(\beta) = \sum_{i=1}^{n} [x_i - \tanh(\beta s_i)] s_i, \tag{2.3}$$

$$-pl_n''(\beta) = \sum_{i=1}^{n} \cosh(\beta s_i)^{-2} s_i^2. \tag{2.4}$$

The question we want to investigate is whether the estimating equation (2.3) has some optimality properties? The form of (2.3) suggests a slightly more general class of estimating functions, namely

$$\Gamma_n(\beta; g) = \sum_{i=1}^{n} [x_i - \tanh(\beta s_i)] g(s_i; \beta). \tag{2.5}$$

This class of estimating functions has the important property that each term in the sum (2.5) has conditional mean zero, that is

$$E([X_i - \tanh(\beta S_i)] g(S_i; \beta) \mid X_j : j \neq i) = 0. \tag{2.6}$$

This property is essential when proving asymptotic normality of the estimate in related, but more complicated models, where one has the possibility of phase transitions (see Jensen and Künsch, 1994). What we want to consider is whether

$$g(s_i; \beta) = s_i$$

is the optimal choice of g in (2.5). Optimality is here defined in terms of having the smallest asymptotic variance of the estimate.

For $2 < i < n-1$ we have from (2.6) that

$$E\{[X_i - \tanh(\beta S_i)] g(S_i; \beta) \Gamma_n(\beta; g)\} = EY_i\{Y_{i-2} + Y_{i-1} + Y_i + Y_{i+1} + Y_{i+2}\},$$

where $Y_i = [X_i - \tanh(\beta S_i)] g(S_i; \beta)$. We therefore find that

$$H(g) = \lim_{n\to\infty} \operatorname{Var}\left(\frac{1}{\sqrt{n}} \Gamma_n(\beta; g)\right) = EY_1\{Y_1 + 2Y_2 + 2Y_3\}. \tag{2.7}$$

Minus the derivative of (2.5) with respect to β is

$$j_n(\beta; g) = \sum_{i=1}^{n} \left\{ \cosh(\beta s_i)^{-2} s_i g(s_i; \beta) - [x_i - \tanh(\beta s_i)] \frac{\partial g(s_i; \beta}{\partial \beta} \right\}.$$

We therefore have from (2.6) that

$$J(g) = \lim_{n\to\infty} E\left(\frac{1}{n} j_n(\beta; g)\right) = E\frac{S_1 g(S_1; \beta)}{\cosh(\beta S_1)^2}. \tag{2.8}$$

Combining (2.7) and (2.8) we find

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{\sim} N\left(0, H(g)/J(g)^2\right),$$

where $\hat{\beta}$ is the solution to $\Gamma_n(\beta; g) = 0$.

We want to minimize $H(g)/J(g)^2$ with respect to g. For symmetry reasons in (2.5) we only consider functions with $g(-s_i; \beta) = -g(s_i; \beta)$ for $s_i \neq 0$. We now argue that the smallest variance is obtained with $g(0; \beta) = 0$. Intuitively this is clear from (2.5) as those terms with $s_i = 0$ give no information on β. More precisely, we can argue as follows. Write $U_i = Y_i 1(S_i \neq 0)$ and $V_i = Y_i 1(S_i = 0)$. Both these terms have conditional mean zero and therefore

$$\begin{aligned}
\lim_{n\to\infty} \mathrm{Var}\left(\frac{1}{\sqrt{n}}\Gamma_n(\beta; g)\right) &= \lim_{n\to\infty} \mathrm{Var}\left(\frac{1}{\sqrt{n}}\sum_1^n U_i + \frac{1}{\sqrt{n}}\sum_1^n V_i\right) \\
&= \lim_{n\to\infty} \left\{ \mathrm{Var}\left(\frac{1}{\sqrt{n}}\sum_1^n U_i\right) + \mathrm{Var}\left(\frac{1}{\sqrt{n}}\sum_1^n V_i\right) \right. \\
&\qquad \left. + \mathrm{Cov}\left(\frac{1}{\sqrt{n}}\sum_1^n U_i, \frac{1}{\sqrt{n}}\sum_1^n V_i\right) \right\} \\
&= EU_1(U_1 + 2U_2 + 2U_3) + EV_1(V_1 + 2V_2 + 2V_3) \\
&\qquad + EV_3(U_1 + U_2 + U_4 + U_5). \qquad (2.9)
\end{aligned}$$

The conditional specification in (2.2) implies that $-X \sim X$, that is, inverting the signs of all the X_i's does not change the distribution. Since $U_i(-X) = U_i(X)$ and $V_i(-X) = -V_i(X)$ we find that the third term in (2.9) is zero. The second term in (2.9) is proportional to $g(0; \beta)^2$ and so we get the minimum for $g(0; \beta) = 0$. Since also $J(g)$ does not depend on $g(0; \beta)$ we get that $H(g)/J(g)^2$ is minimized for $g(0; \beta) = 0$.

Normalizing g by setting $g(4; \beta) = 4$ we end up considering the class of estimating functions (2.5) with g belonging to

$$\{g : \{-4, -2, 0, 2, 4\} \to \mathbf{R} \mid g(0) = 0, g(-s) = -g(s), g(4) = 4\}. \tag{2.10}$$

We therefore only have one free parameter $g(2; \beta)$ that we want to choose so that $H(g)/J(g)^2$ is minimized.

3 A partial evaluation of the limiting variance

In this section we give a partial derivation of $H(g)$ and $J(g)$ from (2.7) and (2.8), respectively. We will perform the calculation by conditioning on (x_{-1}, x_0, x_4, x_5). The conditional distribution has 8 states with probabilities given by (2.2), and the conditional means of the terms in (2.7) and (2.8) can be written explicitly. There are 16 possibilities for (x_{-1}, x_0, x_4, x_5), but these can be paired two by two using that $-X \sim X$. We therefore basically have 8 different conditional distributions to consider.

As an example consider the case $(x_{-1}, x_0, x_4, x_5) = (1, 1, 1, 1)$. With $\xi = g(2; \beta)$, $r = \tanh(2\beta)$, and $t = \tanh(4\beta)$ the conditional mean of $Y_1(Y_1 + 2Y_2 + 2Y_3)$ is

$$\left\{e^{9\beta} + 3e^{\beta} + 4e^{-3\beta}\right\}^{-1} \left\{80e^{9\beta}(1-t)^2 + e^{\beta}[16(1+t)^2 + 6(1-r)^2\xi^2 - 32\xi(1-r)(1+t)] + 6e^{-3\beta}(1+r)^2\xi^2\right\}.$$

Calculating all the conditional means and using the notation $P(ij; kl) = P(X_{-1} = i, X_0 = j, X_4 = k, X_5 = l)$ we obtain

$$\begin{aligned}
H(g) &= \frac{2P(11;11)}{e^{9\beta} + 3e^{\beta} + 4e^{-3\beta}} \left\{80e^{9\beta}(1-t)^2 + 16e^{\beta}(1+t)^2 \right. \\
&\quad \left. +\xi\left[-32e^{\beta}(1-r)(1+t)\right] + \xi^2\left[6e^{\beta}(1-r)^2 + 6e^{-3\beta}(1+r)^2\right]\right\} \\
&+ \frac{2P(-11;11)}{e^{7\beta} + e^{3\beta} + 5e^{-\beta} + e^{-5\beta}} \left\{\xi\left[16e^{7\beta}(1-t)(1-r)\right] + \xi^2\left[e^{7\beta}(1-r)^2\right.\right. \\
&\quad \left.\left. +e^{3\beta}((1+r)^2 - 4(1-r)(1+r)) + e^{-\beta}(1-r)^2 + 5e^{-5\beta}(1+r)^2\right]\right\} \\
&+ \frac{2P(11;1\text{-}1)}{e^{7\beta} + e^{3\beta} + 5e^{-\beta} + e^{-5\beta}} \left\{48e^{7\beta}(1-t)^2 + 16e^{-\beta}(1+t)^2 \right. \\
&\quad +\xi\left[8e^{7\beta}(1-t)(1-r) - 16e^{-\beta}(1-r)(1+t)\right] + \xi^2\left[e^{3\beta}(3(1-r)^2\right. \\
&\quad \left.\left. -2(1-r)(1+r)) + e^{-\beta}((1-r)^2 + (1+r)^2) + 5e^{-5\beta}(1+r)^2\right]\right\} \\
&+ \frac{2P(1\text{-}1;11)}{e^{5\beta} + 4e^{\beta} + 3e^{-3\beta}} \left\{\xi\left[8e^{5\beta}(1-t)(1-r)\right] + \xi^2\left[3e^{5\beta}(1-r)^2\right.\right. \\
&\quad \left.\left. +e^{\beta}((1+r)^2 + 3(1-r)^2 - 2(1-r)(1+r)) + 3e^{-3\beta}(1+r)^2\right]\right\} \\
&+ \frac{2P(11;\text{-}11)}{e^{5\beta} + 4e^{\beta} + 3e^{-3\beta}} \left\{16e^{5\beta}(1-t)^2 + 16e^{-3\beta}(1+t)^2 \right. \\
&\quad +\xi\left[16e^{5\beta}(1-r)(1-t)\right] \\
&\quad \left. +\xi^2\left[e^{\beta}(2(1-r)^2 - 4(1+r)(1-r)) + 6e^{-3\beta}(1+r)^2\right]\right\} \\
&+ \frac{2P(11;\text{-}1\text{-}1)}{4e^{3\beta} + 2e^{-\beta} + 2e^{-5\beta}} \left\{16e^{3\beta}(1-t)^2 + 16e^{-5\beta}(1+t)^2\right.
\end{aligned} \tag{3.1}$$

$$
\begin{aligned}
&+\xi\left[8e^{3\beta}(1-r)(1-t)+16e^{-5\beta}(1+t)(1+r)\right]+\xi^2\left[e^{3\beta}(1-r)^2\right.\\
&\left.+e^{-\beta}((1-r)^2+3(1+r)^2-6(1-r)(1+r))+e^{-5\beta}(1+r)^2\right]\Big\}\\
+\;&\frac{2P(1\text{-}1;1\text{-}1)}{2e^{3\beta}+6e^{-\beta}}\left\{\xi^2\left[10e^{3\beta}(1-r)^2+2e^{-\beta}(1+r)^2\right]\right\}\\
+\;&\frac{2P(1\text{-}1;-11)}{e^{5\beta}+4e^{\beta}+3e^{-3\beta}}\left\{\xi\left[8e^{5\beta}(1-t)(1-r)\right]+\xi^2\left[3e^{5\beta}(1-r)^2\right.\right.\\
&\left.\left.+e^{\beta}(3(1-r)^2+(1+r)^2-2(1-r)(1+r))+3e^{-3\beta}(1+r)^2\right]\right\}
\end{aligned}
$$

A similar calculation, with $\phi = 16/\cosh(4\beta)^2$ and $\psi = 2/\cosh(2\beta)^2$, gives the formula

$$
\begin{aligned}
J(g) \;=\;&\frac{2P(11;11)}{e^{9\beta}+3e^{\beta}+4e^{-3\beta}}\left\{\phi\left[e^{9\beta}+e^{\beta}\right]+\xi\psi\left[2e^{\beta}+2e^{-3\beta}\right]\right\} \qquad (3.2)\\
+\;&\frac{2P(11;1\text{-}1)}{e^{7\beta}+e^{3\beta}+5e^{-\beta}+e^{-5\beta}}\left\{\phi\left[e^{7\beta}+e^{-\beta}\right]+\xi\psi\left[e^{3\beta}+2e^{-\beta}+e^{-5\beta}\right]\right\}\\
+\;&\frac{2P(\text{-}11;11)}{e^{7\beta}+e^{3\beta}+5e^{-\beta}+e^{-5\beta}}\left\{\xi\psi\left[e^{7\beta}+e^{3\beta}+e^{-\beta}+e^{-5\beta}\right]\right\}\\
+\;&\frac{2P(11;\text{-}11)}{e^{5\beta}+4e^{\beta}+3e^{-3\beta}}\left\{\phi\left[e^{5\beta}+e^{-3\beta}\right]+\xi\psi\left[2e^{\beta}+2e^{-3\beta}\right]\right\}\\
+\;&\frac{2P(1\text{-}1;11)}{e^{5\beta}+4e^{\beta}+3e^{-3\beta}}\left\{\xi\psi\left[e^{5\beta}+2e^{\beta}+e^{-3\beta}\right]\right\}\\
+\;&\frac{2P(11;\text{-}1\text{-}1)}{4e^{3\beta}+2e^{-\beta}+2e^{-5\beta}}\left\{\phi\left[e^{3\beta}+e^{-5\beta}\right]+\xi\psi\left[e^{3\beta}+2e^{-\beta}+e^{-5\beta}\right]\right\}\\
+\;&\frac{2P(1\text{-}1;1\text{-}1)}{2e^{3\beta}+6e^{-\beta}}\left\{\xi\psi\left[2e^{3\beta}+2e^{-\beta}\right]\right\}\\
+\;&\frac{2P(1\text{-}1;\text{-}11)}{e^{5\beta}+4e^{\beta}+3e^{-3\beta}}\left\{\xi\psi\left[e^{5\beta}+2e^{\beta}+e^{-3\beta}\right]\right\}.
\end{aligned}
$$

Writing $H(g) = a_0 + a_1\xi + a_2\xi^2$ and $J(g) = c_0 + c_1\xi$ we find that the asymptotic variance $H(g)/J(g)^2$ is minimized for

$$\tilde{\xi} = \frac{2a_0c_1 - a_1c_0}{2a_2c_0 - a_1c_1}. \qquad (3.3)$$

For $\beta = 0$ the X_i's are independent and $P(\cdot\cdot;\cdot\cdot) = 1/16$. From (3.1) and (3.2) we find $H(g) = \frac{7}{2} + \frac{1}{2}\xi + \frac{7}{8}\xi^2$ and $J(g) = 1 + \frac{1}{2}\xi$. The optimal value of ξ is then $\tilde{\xi} = 2$, that is, the estimating function (2.3) is optimal in the class (2.10) when $\beta = 0$.

Let us now evaluate (3.3) in the limit $\beta \to \infty$. In the limit $\beta \to \infty$ the distribution of X becomes concentrated in sequences with no change of sign along the sequence. One change of sign reduces the probability by a factor

$\exp(-6\beta)$. By such intuitive reasoning we find that as $\beta \to \infty$

$$P(11;11) \to \tfrac{1}{2},$$
$$P(11;1\text{–}1) = P(\text{–}11;11) \sim \tfrac{1}{2}\mathrm{e}^{-6\beta}, \qquad P(11;\text{–}1\text{–}1) \sim 2\mathrm{e}^{-6\beta},$$
$$\max\{P(11;\text{–}11), P(1\text{–}1;11), P(1\text{–}1;1\text{–}1), P(1\text{–}1;\text{–}11)\} = o(\mathrm{e}^{-6\beta}).$$

These formulas can be proved directly from the results of the next section. Using these asymptotic relations we find from (3.1) and (3.2) that

$$a_0 \sim 64\mathrm{e}^{-8\beta}, \quad a_1 \sim -128\mathrm{e}^{-12\beta}, \quad a_2 \sim 16\mathrm{e}^{-10\beta},$$
$$c_0 \sim 64\mathrm{e}^{-8\beta}, \quad c_1 \sim 16\mathrm{e}^{-10\beta},$$

and therefore from (3.3)

$$\tilde{\xi} \to 1.$$

We can therefore conclude that for β large the pseudo-likelihood estimating equation (2.3) is not optimal in the class (2.10). In the next section we evaluate precisely the difference between the optimal choice and the pseudo-likelihood.

4 The equivalent Markov chain

For a one-dimensional lattice gas of finite range it is possible to express the distribution through a stationary Markov chain. We will use this to evaluate the probabilities $P(\cdot\cdot;\cdot\cdot)$ in (3.1) and (3.2).

In (2.2) the interaction of x_i with previous values is through $x_i(x_{i-2} + x_{i-1})$. If we pair the variables $Y_i = (Y_{i1}, Y_{i2}) = (X_{2i-1}, X_{2i})$ the interaction of Y_i with previous values is through $f(y_{i-1}, y_i) = y_{i1}(y_{i-1,1} + y_{i-1,2}) + y_{i2}(y_{i-1,2} + y_{i1})$. To get the Gibbs model in (2.2) we need a Markov chain for which the product of the transition probabilities equals $\exp\left(\sum_1^k f(y_{i-1}, y_i)\right)$ except for an initial term and a final term. This is exactly the construction used in large deviation theory for Markov chains, see e.g. Jensen (1991). We number the possible values of Y_i in the order $(1,1), (1,-1), (-1,1)$ and $(-1,-1)$. We let Q be the matrix with entries $\exp\{\beta f(y_{i-1}, y_i)\}$, that is,

$$Q = \begin{pmatrix} \mathrm{e}^{4\beta} & 1 & \mathrm{e}^{-2\beta} & \mathrm{e}^{-2\beta} \\ 1 & 1 & \mathrm{e}^{-2\beta} & \mathrm{e}^{2\beta} \\ \mathrm{e}^{2\beta} & \mathrm{e}^{-2\beta} & 1 & 1 \\ \mathrm{e}^{-2\beta} & \mathrm{e}^{-2\beta} & 1 & \mathrm{e}^{4\beta} \end{pmatrix}$$

Let λ be the largest eigenvalue of Q and let $e_r = (1, v, v, 1)$ be a corresponding right eigenvector and $e_l = (1, u, u, 1)$ a corresponding left eigenvector. We

then define the matrix P by $P(y,z) = \lambda^{-1} e_r(y)^{-1} Q(y,z) e_r(z)$,

$$P = \frac{1}{\lambda} \begin{pmatrix} e^{4\beta} & v & ve^{-2\beta} & e^{-2\beta} \\ v^{-1} & 1 & e^{-2\beta} & v^{-1}e^{2\beta} \\ v^{-1}e^{2\beta} & e^{-2\beta} & 1 & v^{-1} \\ e^{-2\beta} & ve^{-2\beta} & v & e^{4\beta} \end{pmatrix},$$

and define the vector μ by

$$\mu = \frac{(1, uv, uv, 1)}{2 + 2uv}.$$

With these definitions we get

$$\prod_{i=1}^{k} P(y_{i-1}, y_i) = \lambda^{-k} e_r(y_0)^{-1} e_r(y_k) \exp\left\{\beta \sum_{i=1}^{k} f(y_{i-1}, y_i)\right\}, \quad (4.1)$$

which gives the same model as in (2.2). The stationary distribution of Y_i is given by μ. It is not difficult to show that

$$\lambda = \frac{1}{2}\left\{e^{4\beta} + 1 + 2e^{-2\beta} + \sqrt{e^{8\beta} - 2e^{4\beta} + 4e^{2\beta} + 9 + 4e^{-2\beta}}\right\} \quad (4.2)$$

$$u = \frac{\lambda - e^{4\beta} - e^{-2\beta}}{1 + e^{2\beta}}, \quad \text{and} \quad v = \frac{\lambda - e^{4\beta} - e^{-2\beta}}{1 + e^{-2\beta}}.$$

Let now

$$q_{ij} = P(i,1)P(1,j) + P(i,2)P(2,j) + P(i,3)P(3,j) + P(i,4)P(4,j).$$

Then we have for $P(\cdot\cdot;\cdot\cdot)$ in (3.1) and (3.2)

$$\begin{aligned} P(11;11) &= \mu(1)\{q_{11}[P(1,1) + P(1,2)] + q_{13}[P(3,1) + P(3,2)]\}, \\ P(11;\text{-}11) &= \mu(1)\{q_{12}[P(2,1) + P(2,2)] + q_{14}[P(4,1) + P(4,2)]\} \\ P(\text{-}11;11) &= \mu(3)\{q_{31}[P(1,1) + P(1,2)] + q_{33}[P(3,1) + P(3,2)]\}, \end{aligned}$$

and so forth. We are therefore now in a position to calculate (3.1) and (3.2) and the optimal value (3.3). In Table 1 we have given $\tilde{\xi}$ and the asymptotic variance with the optimal choice $\tilde{\xi}$ and with $\xi = 2$ corresponding to the pseudo-likelihood estimate. It is clear from the table that although the pseudo-likelihood is not optimal in the class (2.10) it is very close to being so, and the difference has no practical importance.

Because of the representation of our model as a Markov chain we can find the limiting variance of the maximum likelihood estimate $\hat{\beta}_0$. From (4.1) we can find the observed information based on $X_1, \ldots, X_n$, $n = 2k$. Dividing

β	$\tilde{\xi}$	$\xi = \tilde{\xi}$	$\xi = 2$	mle	double
0.0	2.00	0.50	0.50	0.50	0.51
0.2	2.04	0.33	0.33	0.29	0.33
0.4	2.03	0.46	0.46	0.29	0.44
0.6	1.86	1.67	1.67	0.69	1.53
0.8	1.64	8.36	8.41	2.39	7.86
1.0	1.46	43.32	43.78	8.76	41.96
1.5	1.19	2496.61	2518.04	205.55	2487.62

Table 1: The column $\tilde{\xi}$ gives the optimal value (3.3) of ξ. The other columns give the asymptotic variance of different estimates. The column $\xi = \tilde{\xi}$ is the estimate obtained from (2.5) with the optimal choice of g from (2.10), $\xi = 2$ is the pseudo-likelihood estimate obtained from (2.3), "mle" is the maximum likelihood estimate, and "double" is the extended pseudo-likelihood estimate obtained from (5.1).

this by n and taking the limit $n \to \infty$ we get $\frac{1}{2}\frac{d^2}{d\beta^2}\log\lambda(\beta)$, with λ given in (4.2), and therefore

$$\sqrt{n}(\hat{\beta}_0 - \beta) \xrightarrow{\sim} N\left(0, 2\left\{\frac{d^2}{d\beta^2}\log\lambda(\beta)\right\}^{-1}\right).$$

In Table 1 the limiting variance of the maximum likelihood estimate has been included. As can be seen for $\beta = 0$ the pseudo-likelihood estimate is fully efficient, whereas for large values of β the efficiency is quite poor. Actually, for $\beta \to \infty$ the ratio of the limiting variance of the pseudo-likelihood estimate to the variance of the maximum likelihood estimate is $9\exp(2\beta)/16$.

In statistical applications models of the form considered here are often used in situations with a weak interaction, that is, with small values of β, see e.g. Besag (1974). In such cases we can expect the efficiency of the pseudo-likelihood to be acceptable.

5 An extended pseudo-likelihood

The pseudo-likelihood considered in Section 2 is based on the one-dimensional conditional distribution (2.1). It is therefore not surprising that the efficiency of the pseudo-likelihood estimate is poor when β is large. When β is large the dependency in the chain is very strong and this can not be seen efficiently from the one-dimensional conditional distributions. It seems of interest then to investigate what improvement we get by extending the pseudo-likelihood idea to consider the conditional distributions of two variables, say, given the remaining variables. Precisely, we base a new pseudo-likelihood $\exp(\tilde{pl}_n)$ on

(2.2) with $l = k + 1$,

$$\tilde{pl}_n(\beta) = \sum_{i=1}^{n} \left\{ \beta \left[x_i s_i^1 + x_{i+1} s_i^2 + x_i x_{i+1} \right] - \log \left[2e^{\beta} \cosh(\beta(s_i^1 + s_i^2)) + 2e^{-\beta} \cosh(\beta(s_i^1 - s_i^2)) \right] \right\}, \quad (5.1)$$

where

$$s_i^1 = x_{i-2} + x_{i-1} + x_{i+2} \quad \text{and} \quad s_i^2 = x_{i-1} + x_{i+2} + x_{i+3}.$$

As for $pl_n(\beta)$ we can calculate the limiting variance of $\tilde{pl}_n'(\beta)$ and the limiting mean of $\tilde{pl}_n''(\beta)$ to get the limiting variance of the estimate $\tilde{\beta}$ satisfying $\tilde{pl}_n'(\tilde{\beta}) = 0$. The result can be seen in Table 1 in the column with the heading "double". As can be seen, for this one-dimensional lattice gas, the improvement is quite small and of no practical importance. It seems important to investigate if this conclusion is also true for higher dimensions. Intuitively, one feels that in two dimensions, say, much more about directional differences in the interactions can be learned from a set of nine points, say, than from a single point.

To summarize, the pseudo-likelihood is based on local information and this is the basis for simple formulas and for simple asymptotic properties. However, using only local information is not a very efficient procedure when there is strong interaction. It therefore seems likely that if one can construct a more efficient class of estimating functions, one will be faced with complicated asymptotic properties.

References

Besag (1974): Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. Roy. Statist. Soc. B*, **36**, 192-236.

Besag, J. (1975): Statistical analysis of non-lattice data. *The Statistician*, **24**, 179-195.

Bibby, B.M. and Sørensen, M. (1996): Martingale estimation function for discretely observed diffusion processes. *Bernoulli*, **1**, 17-39.

Guyon, X. and Künsch, H.R. (1992): Asymptotic comparison of estimators in the Ising model. In Stochastic Models, Statistical Methods and Algorithms in Image Analysis. Eds. P.Barone, A. Frigessi and M. Piccioni, *Lecture Notes in Statist.*, **74**, 177-198, Springer, New York.

Heyde, C.C. (1988): Fixed sample and asymptotic optimality for classes of estimating functions. *Contemporary Mathematics*, **80**, 241-244.

Jensen J.L. (1991): Saddlepoint expansions for sums of Markov dependent variables on a continuous state space. *Probab. Th. Rel. Fields*, **89**, 181-199.

Jensen, J.L. and Künsch, H.R. (1994): On asymptotic normality of pseudo likelihood estimates for pairwise interaction processes. *Ann. Inst. Statist. Math.*, **46** 475-486.