A note on pseudolikelihood constructed from marginal densities

D.R.Cox
Nuffield College, Oxford, OX1 1NF, UK
david.cox@nuf.ox.ac.uk
and
N.Reid
Department of Statistics, University of Toronto, Toronto M5S 3GS, Canada
reid@utstat.utoronto.ca

SUMMARY

For likelihood-based inference involving distributions in which high-dimensional dependencies are present it may be useful to use approximate likelihoods based, for example, on the univariate or bivariate marginal distributions. The asymptotic properties of formal maximum likelihood estimators in such cases are outlined. In particular, applications in which only a single $q \times 1$ vector of observations is observed are examined. Conditions under which consistent estimators of parameters result from the approximate likelihood using only pairwise joint distributions are studied. Some examples are analysed in detail.

Some key words: Component of variance; Composite likelihood; Consistency of estimation; Dichotomised Gaussian distribution; Generalized estimating equations; Genetic statistics; Maximum likelihood; Pseudolikelihood.

1 Pseudolikelihood and score function

While the likelihood function has a central place in the formal theory of statistical inference for particular models, there are a number of situations where some modification of the likelihood is needed perhaps for robustness or perhaps because of the complexity of the full likelihood. In this paper we examine a special form of pseudolikelihood that is potentially useful when complex interdependencies are involved in the full likelihood.

Suppose that Y is a $q \times 1$ vector random variable with density $f(y;\theta)$, where θ is an unknown parameter which initially we take to be one-dimensional. From independent observations $Y^{(1)}, \ldots, Y^{(n)}$ we may find the maximum likelihood estimator of θ which has under the usual regularity conditions an asymptotically normal distribution with mean θ and variance the inverse of the Fisher information

$$nE\{-\frac{\partial^2 \log f(Y;\theta)}{\partial\theta^2}\}.$$

Now suppose that it is difficult to specify the full q-dimensional distribution in convenient form but that it is possible to specify all one-, two-, ... dimensional distributions up to some order. Here we concentrate on just the one- and two-dimensional ones; that is, we can specify for all $s, t = 1, \ldots, q$ the univariate and bivariate densities $f_s(y_s;\theta), f_{st}(y_s, y_t;\theta)$ for $s \neq t$. Thus from one vector Y we may form the first- and second-order loglikelihood contributions

$$\begin{aligned} \ell_1(\theta;Y) &= \Sigma_s \log f(Y_s;\theta), \\ \ell_2(\theta;Y) &= \Sigma_{s>t} \log f(Y_s, Y_t;\theta) - aq\ell_1(\theta;Y), \end{aligned}$$

where a is a constant to be chosen. Note that taking $a = 0$ corresponds to taking all possible bivariate distributions whereas $a = 1/2$ corresponds in

effect to taking all possible conditional distributions of one component given another; this is the pseudolikelihood suggested by Besag (1974) for analysis of spatial data. It may happen that, say, $\ell_1(\theta; Y)$ is in fact independent of θ, i.e. that the one-dimensional marginal distributions contain no information about θ.

For n independent, identically distributed vectors we define corresponding pseudo loglikelihoods by addition:

$$\begin{aligned} \ell_1(\theta; Y^{(1)}, \ldots, Y^{(n)}) &= \Sigma_i \ell_1(\theta; Y^{(i)}), \\ \ell_2(\theta; Y^{(1)}, \ldots, Y^{(n)}) &= \Sigma_i \ell_2(\theta; Y^{(i)}). \end{aligned}$$

We define pseudo score functions by loglikelihood derivatives in the usual way:

$$\begin{aligned} U_1(\theta; Y) &= \partial \ell_1(\theta; Y)/\partial\theta = \Sigma_s U_{1s}(\theta), \\ U_2(\theta; Y) &= \partial \ell_2(\theta; Y)/\partial\theta = \Sigma_{s>t} U_{2st}(\theta) - aq\Sigma_s U_{1s}(\theta), \\ U_1(\theta; Y^{(1)}, \ldots, Y^{(n)}) &= \partial \ell_1(\theta; Y^{(1)}, \ldots, Y^{(n)})/\partial\theta = \Sigma U_1(\theta; Y^{(i)}), \\ U_2(\theta; Y^{(1)}, \ldots, Y^{(n)}) &= \partial \ell_2(\theta; Y^{(1)}, \ldots, Y^{(n)})/\partial\theta = \Sigma U_2(\theta; Y^{(i)}). \end{aligned}$$

The functions ℓ_1 and ℓ_2 are examples of composite likelihood functions, studied in generality in Lindsay (1988). The first term in ℓ_2 is called the pairwise likelihood. The estimating equations

$$U_\nu(\tilde{\theta}; Y^{(1)}, \ldots, Y^{(n)}) = 0$$

for $\nu = 1, 2$ are under the usual regularity conditions unbiased, provided of course that the relevant pseudo loglikelihoods depend on θ and so do not make the U identically zero. The resulting estimator is for large n asymptotically normal with mean θ and variance

$$[E\{-U'_\nu(\theta)\}]^{-2} E\{U^2_\nu(\theta)\}. \tag{1}$$

Furthermore, $E(U_\nu^2)$ can be estimated by $n^{-1}\Sigma_i U_\nu^2(\tilde{\theta}; Y^{(i)})$ and $E\{-U_\nu'(\theta; Y)\}$ by $-\ell_\nu''(\tilde{\theta})/n$.

An example is given by the symmetric normal distribution. We assume for each i that the components of $Y^{(i)}$ follow a standard normal distribution and that $\mathrm{corr}(Y_r^{(i)}, Y_s^{(i)}) = \rho$. There is no information about ρ in the univariate marginal densities. The second-order pseudolikelihood is

$$\begin{aligned} l_2(\rho; Y^{(1)}, \ldots, Y^{(n)}) &= -\frac{nq(q-1)}{4}\log(1-\rho^2) - \frac{q-1+\rho}{2(1-\rho^2)}SS_W \\ &\quad -\frac{(q-1)(1-\rho)}{2(1-\rho^2)}\frac{SS_B}{q}, \end{aligned}$$

where

$$SS_W = \Sigma_{i=1}^n \Sigma_{r=1}^q (Y_r^{(i)} - \bar{Y}_\cdot^{(i)})^2, \qquad SS_B = \Sigma_{i=1}^n Y_\cdot^{(i)2}.$$

The associated score function is

$$\begin{aligned} U_2(\rho; Y^{(1)}, \ldots, Y^{(n)}) &= \frac{nq(q-1)\rho}{2(1-\rho^2)} - \frac{1+\rho^2+2(q-1)\rho}{2(1-\rho^2)^2}SS_W \\ &\quad -\frac{(q-1)(1-\rho)^2}{2(1-\rho^2)^2}\frac{SS_B}{q} \end{aligned}$$

and the asymptotic variance of $\tilde{\rho}$ is

$$\mathrm{avar}(\tilde{\rho}) = \frac{2}{nq(q-1)}\frac{(1-\rho)^2 c(q,\rho)}{(1+\rho^2)^2},$$

where

$$c(q,\rho) = (1-\rho)^2(3\rho^2+1) + q\rho(-3\rho^3+8\rho^2-3\rho+2) + q^2\rho^2(1-\rho)^2.$$

This may be compared to the variance of the maximum likelihood estimator using the full model,

$$\mathrm{avar}(\hat{\rho}) = \frac{2}{nq(q-1)}\frac{\{1+(q-1)\rho\}^2(1-\rho)^2}{\{1+(q-1)\rho^2\}}.$$

This ratio is 1 for $q = 2$, as expected, and is also 1 if $\rho = 0$ or 1, for any value of q. Figure 1 illustrates the loss of information with increasing q.

2 Estimating equations: large q

In the previous section we considered fixed q as n increases. We now look at the problem where a small number n of individually large sequences is available, i.e. we let q increase for fixed n. This includes the possibility of observing a single replicate of a process in which substantial and possibly complicated internal dependencies are present. The case that n and q increase simultaneously, for example in a fixed ratio, may also be of interest.

While the estimating equation $U_\nu(\tilde{\theta}; Y) = 0$ is unbiased, this no longer implies satisfactory properties of the resulting estimator.

Consider first the estimating equation $U_1(\tilde{\theta}; Y) = 0$, still assuming for simplicity that θ is a scalar. We expand formally around θ to obtain, to the first order,

$$q^{-1}\Sigma U_{1s}(\theta) + q^{-1}(\tilde{\theta} - \theta)\Sigma U'_{1s}(\theta) = 0.$$

The second random sum is typically $O_p(1)$, whereas the first random sum has zero mean and variance

$$q^{-2}\{\Sigma \mathrm{var}(U_{1s}) + 2\Sigma_{s>t}\mathrm{cov}(U_{1s}, U_{1t})\}$$

and, depending on the form of the covariance terms, this may be $O(q^{k-2})$ for $1 \leq k \leq 2$, so that the first random sum is $O_p(q^{(k/2)-1})$. There are two main possibilities. First, if $k = 2$ the two random terms in the expansion of the estimating equation are of the same order in probability and this suggests that $\tilde{\theta}$ will not be a consistent estimator of θ as q increases; there is too much internal correlation present. On the other hand, if $k < 2$ then expansion shows that $q^{1-k/2}(\tilde{\theta} - \theta)$ is asymptotically normal with mean zero and variance that can be derived from the expansion. Of course if k is close to 2 then convergence may be very slow.

To illustrate the general discussion, if the components of Y have marginally any exponential family distribution with mean parameter θ and are arbitrarily correlated, then if the correlation is that of short-range dependent stationary time series, $k = 1$ and convergence of the overall sample mean to θ will be at the usual rate, that is $1/\sqrt{q}$, whereas if the correlation is that of a long-range dependent process with Hurst coefficient $H > 1/2$ then $k = 2H$ and convergence will be slower. Finally, if all pairs are equally correlated then $k = 2$ and the mean of a single realization is not a consistent estimator of θ. These results are clear from first principles.

A very similar if more complicated discussion applies to the use of pairwise dependencies. The estimating equation $U_2(\tilde{\theta}; Y) = 0$ has the expansion

$$q^{-2}\{\Sigma_{s>t}U_{2st}(\theta) - aq\Sigma_r U_{1r}(\theta)\} + q^{-2}(\tilde{\theta} - \theta)\{\Sigma_{s>t}U'_{2st}(\theta) - aq\Sigma_s U'_{1s}\theta\} = 0.$$

The second term is typically $O_p(1)$, whereas the first term has mean zero and variance

$$q^{-4}[\mathrm{var}\{\Sigma_{s>t}U_{2st}(\theta)\} - 2aq\mathrm{cov}\{\Sigma_{s>t}U_{2st}(\theta), \Sigma_r U_{1r}(\theta)\} + (aq)^2\mathrm{var}\{\Sigma_s U_{1s}(\theta)\}]. \quad (2)$$

This can be calculated as a function of expected products of U's of various kinds:

$$\begin{aligned}
\mathrm{var}\{\Sigma_{s>t}U_{2st}(\theta)\} &= \frac{1}{2}q(q-1)E(U_{2st}^2) + q(q-1)(q-2)E(U_{2st}U_{2sv}) \\
&\quad + \frac{1}{4}q(q-1)(q-2)(q-3)E(U_{2st}U_{2vw}), \\
\mathrm{cov}\{\Sigma_{s>t}U_{2st}, \Sigma U_{1r}\} &= \frac{1}{2}q(q-1)E(U_{2st}U_{1s}) + \frac{1}{2}q(q-1)(q-2)E(U_{2st}U_{1v}), \\
\mathrm{var}\{\Sigma U_{1s}\} &= qE(U_{1s}^2) + q(q-1)E(U_{1s}U_{1t}),
\end{aligned}$$

where $s \neq t \neq v \neq w$. The leading term in q is

$$\{E(U_{2st}U_{2vw}) + a^2E(U_{1s}U_{1t})\}$$

and thus the first random sum may be of comparable order to the second random sum, which suggests that the estimating equation will not usually lead to a consistent estimator of θ. This is the case in the example above of the normal correlation coefficient when $n = 1$; the asymptotic variance of $\tilde{\rho}$ is $O(1)$ in q.

When $n > 1$ the usual asymptotic theory again applies in n, and the variance of $\tilde{\theta}$ is given by (1) with $\nu = 2$. In this case if both the univariate and bivariate distributions provide information on θ it should be possible to choose a to maximize the information provided.

In the above development we have assumed that each pair of observations (Y_s, Y_t) has the same bivariate distribution, so there is no need to distinguish between $U_{2st}U_{2sv}$ and $U_{2st}U_{2vs}$, for example. In principle however the pseudolikelihood can be defined for the case where the pairs have different distributions, in which case in the expression for the variance of the first random term in the expansion of the estimating equation we have

$$\begin{aligned}\mathrm{var}\{\Sigma_{s>t}U_{2st}(\theta)\} = \Sigma_{s>t}i_{2st} + \Sigma_{s>t,s>v}i_{st,sv} + \Sigma_{s>t>v}i_{st,tv} \\ +\Sigma_{s>t,w>s}i_{st,ws} + \Sigma_{s>t,w>t}i_{st,wt} + \Sigma_{s>t,w>v}i_{st,wv}\end{aligned}$$

where, for example,

$$i_{st,tv} = E\{U_{2st}(\theta)U_{2tv}(\theta)\} = E[\{\partial \log f(Y_s, Y_t; \theta)/\partial\theta\}\{\partial \log f(Y_t, Y_v; \theta)/\partial\theta\}].$$

3 Vector θ

Essentially the same arguments apply when the parameter θ is a vector of length d. The formal likelihood derivatives are defined as before component by component and the expansion of the formal estimating equation can be written

$$U_\nu(\tilde{\theta}; Y^{(1)}, \ldots Y^{(n)}) \;\; = \;\; 0$$

$$
\begin{aligned}
&= U_\nu(\theta; Y^{(1)}, \ldots Y^{(n)}) + (\tilde{\theta} - \theta)^{\mathrm{T}} U'_\nu(\theta; Y^{(1)}, \ldots Y^{(n)}) \\
&= \Sigma_{i=1}^n U_\nu(\theta; Y^{(i)}) + (\tilde{\theta} - \theta)^{\mathrm{T}} \Sigma_{i=1}^n U'_\nu(\theta; Y^{(i)}),
\end{aligned}
$$

so that, asymptotically in n,

$$
\mathrm{acov}(\tilde{\theta}) = n^{-1} E\{-U'_\nu(\theta; Y)\}^{-1} E\{U_\nu(\theta; Y) U_\nu^{\mathrm{T}}(\theta; Y)\} E\{-U'_\nu(\theta; Y)\}^{-1}
$$

and furthermore $E(-U_\nu U_\nu^{\mathrm{T}})$ can be estimated by $\Sigma_i U_\nu(\tilde{\theta}; Y^{(i)}) U^{\mathrm{T}}(\tilde{\theta}; Y^{(i)})$ and $E\{U'_\nu(\theta; Y)\}$ by $-\partial^2 l_\nu(\tilde{\theta})/\partial\theta\partial\theta^{\mathrm{T}}$.

When $n = 1$ we can obtain a consistent estimator using l_2 only if $E(U_2 U_2^T)$ is $O(q^3)$ not $O(q^4)$. The analogue of (2) is

$$
\begin{aligned}
E(U_2 U_2^T) = q^{(2)} K_{st,st}/2 + q^{(3)} K_{st,sw} + q^{(4)} K_{st,vw}/2 \\
-2qa(q^{(2)} K_{st,s} + q^{(3)} K_{st,v}) \\
+q^2 a^2 (q K_{s,s} + q^{(2)} K_{s,t}),
\end{aligned} \tag{3}
$$

where $K_{st,vw} = E(U_{2st} U_{2vw}^T)$, $K_{st,v} = E(U_{2st} U_{1v}^T)$, and so on, and s, t, v, w are all different.

Thus a necessary and sufficient condition for an asymptotic theory in q to hold for fixed n and in particular for $n = 1$ is that there be a real root of the equation

$$
K_{st,vw} - 2aK_{st,v} + 2a^2 K_{s,t} = 0 \tag{4}
$$

in a; note that the K's are square matrices of size equal to the dimension of θ. In some situations consistent estimation would be confined to certain components of θ and then a more complicated condition would be involved.

4 Examples

Example 1: one-way random effects

A simple illustration of these results can be obtained from a single group of q observations taken from a one-way normal-theory random effects model, which can then be reparameterised in various ways; that is, a component $Y_s^{(i)}$ of the ith vector has the form $Y_s^{(i)} = \mu + \xi^{(i)} + \epsilon_{is}$, where $\xi^{(i)}$ and the ϵ_{is} are independently normally distributed with zero mean and variances respectively τ_ξ and τ_ϵ. When $n = 1$ it is clear that only τ_ϵ can be well estimated. The problem may be reformulated by writing Y as multivariate normal with components having mean μ and variance $\sigma^2 = \tau_\xi + \tau_\epsilon$ and with any two components of the same vector having correlation $\rho = \tau_\xi/(\tau_\xi + \tau_\epsilon)$. The example given in §1 is a special case with $\mu = 0$, $\sigma^2 = 1$.

Example 2: dichotomised normal

Suppose that V follows a q-variate normal distribution with correlation matrix R, and that $Y_1, \ldots, Y_q$ are binary variables produced by dichotomising the unobserved components $V_1, \ldots, V_q$. Without loss of generality we can take the mean of the V's to be zero and the variance one. Let $r_{st} = \mathrm{corr}(V_s, V_t)$ and denote the points of dichotomy by $\gamma_1, \ldots, \gamma_q$; that is $Y_s = 0$ or $Y_s = 1$ according as $V_s \leq \gamma_s$ or $V_s > \gamma_s$. We simplify the discussion by supposing the γ_s known; an important special case is median dichotomy when $\gamma_s = 0$. The marginal distributions of the Y_s provide no information so that we use the bivariate pairs: the pseudo loglikelihood based on these pairs is

$$\begin{aligned} \ell_2(R) &= \Sigma_{s>t}\{y_s y_t \log p_{11}^{st} + y_s(1-y_t)\log p_{10}^{st} + (1-y_s)y_t \log p_{01}^{st} \\ &\quad + (1-y_s)(1-y_t)\log p_{00}^{st}\}, \end{aligned}$$

where

$$\begin{aligned} p_{11}^{st} &= \int_{\gamma_t}^{\infty} \left[1 - \Phi\left\{\frac{\gamma_s - r_{st}v}{\sqrt{(1-r_{st}^2)}}\right\}\right]\phi(v)dv \\ p_{10}^{st} &= 1 - \Phi(\gamma_s) - p_{11}^{st} \\ p_{01}^{st} &= \Phi(\gamma_s) - p_{00}^{st} = \Phi(\gamma_t) - p_{11}^{st} \end{aligned}$$

$$p_{00}^{st} = \Phi(\gamma_s) - p_{01}^{st} = p_{11}^{st} + \Phi(\gamma_s) - \Phi(\gamma_t).$$

Note that the full likelihood analysis would involve the q-dimensional normal integral and that quite apart from any computational difficulties there might be fears about the robustness of the specification in so far as it involves high order integrals.

For numerical illustration we consider $\gamma_s = 0$, $r_{st} = \rho$, in which case p_{ij} does not depend on s and t, and we have $p_{00} = p_{11}, p_{10} = p_{01} = (1/2) - p_{11}$, $p_{11} = \sin^{-1}(\rho)/(2\pi) - (1/4)$ and

$$\begin{aligned}
\ell_2(p_{11}) &= \Sigma_{s>t}\{y_s y_t \log p_{11} + y_s(1-y_t)\log\{(1/2) - p_{11}\} \\
&\quad + (1-y_s)y_t \log\{(1/2) - p_{11}\} + (1-y_s)(1-y_t)\log p_{11}\} \\
&= \Sigma_{s>t}[(2y_s y_t - y_s - y_t)\log[p_{11}/\{(1/2) - p_{11}\}] + \{q(q-1)/2\}\log(p_{11})] \\
&= t\log[p_{11}/\{(1/2) - p_{11}\}] + \{q(q-1)/2\}\log p_{11},
\end{aligned}$$

where

$$t = \Sigma_{s>t}(2y_s y_t - y_s - y_t).$$

Then

$$E\{-\ell_2''(\rho)\} = \frac{q(q-1)}{p_{11}(1-2p_{11})}\{p_{11}'(\rho)\}^2,$$

$$\begin{aligned}
\mathrm{var}\{\ell_2'(\rho)\} &= \mathrm{var}\{\ell_2'(p_{11})\}\{p_{11}'(\rho)\}^2 \\
&= \left(\frac{1}{p_{11}} + \frac{1}{(1/2) - p_{11}}\right)^2 \mathrm{var}(t)\{p_{11}'(\rho)\}^2,
\end{aligned}$$

where

$$\begin{aligned}
\mathrm{var}(t) &= 4\mathrm{var}(\Sigma_{s>t} y_s y_t) + (q-1)^2 \mathrm{var}(\Sigma y_s) - 4(q-1)\mathrm{cov}(\Sigma_{s>t} y_s y_t, \Sigma y_s) \\
&= q^4(p_{1111} - 2p_{111} + 2p_{11} - p_{11}^2 - 1/4) \\
&+ q^3(-6p_{1111} + 12p_{111} - 9p_{11} + 2p_{11}^2 + 1) \\
&+ q^2(11p_{1111} - 22p_{111} + 14p_{11} - p_{11}^2 - 5/4) \\
&+ q(-6p_{1111} + 12p_{111} - 7p_{11} + 1/2),
\end{aligned}$$

and

$$\begin{aligned} p_{1111} &= \text{pr}(U_s > 0, U_t > 0, U_v > 0, U_w > 0) \\ p_{111} &= \text{pr}(U_s > 0, U_t > 0, U_v > 0). \end{aligned}$$

In the repeated sampling context as q tends to infinity for fixed n

$$\begin{aligned} \text{avar}(\tilde{\rho}) &= \frac{1}{n} E\{-\ell_2''(\rho)\}^{-2} \text{var}\{\ell_2'(\rho)\} \\ &= \frac{1}{n} \frac{4\pi^2(1-\rho^2)}{q^2(q-1)^2} \text{var}(t); \end{aligned}$$

if $n = 1$ this variance is $O(1)$ as expected on general grounds.

To compare this to the full likelihood calculation we need the full Fisher information. The full likelihood is

$$\begin{aligned} \log \text{pr}(Y_1 = y_1, \ldots, Y_q = y_q) &= (y_1 \ldots y_q) \log p_{11\ldots1}^q + (1-y_1) y_2 \ldots y_q \log p_{01\ldots1}^{q-1} + \\ &\quad + y_1(1-y_2) y_3 \ldots y_q \log p_{101\ldots1}^{q-1} + \\ &\quad \ldots + y_1 \ldots y_{q-1}(1-y_q) \log p_{1\ldots10}^{q-1} \\ &+ \ldots + (1-y_1)\ldots(1-y_q) \log p_{0\ldots0}^0, \end{aligned}$$

where $p_{0\ldots01\ldots1}^i$ is the probability of i ones and $q-i$ zeroes. Then

$$\begin{aligned} \ell'(\rho) &= \Sigma_{i=0}^q t_i (d/d\rho) \log(p_{0\ldots01\ldots1}^{q-i}), \\ \ell''(\rho) &= \Sigma_{i=0}^q t_i (d^2/d\rho^2) \log(p_{0\ldots01\ldots1}^{q-i}), \end{aligned}$$

where t_i is the sum of $\binom{q}{i}$ sequences with $q-i$ ones and i zeroes, and has expected value $\binom{q}{i} p_{0\ldots01\ldots1}^i$. From this the expected Fisher information, and hence the asymptotic variance of $\hat{\rho}$, is readily obtained.

The p's can be evaluated using a simplification given in Tong (1990, p.192) as

$$p_{0\ldots01\ldots1}^i = \int_{-\infty}^{\infty} \left[1 - \Phi\left\{\frac{t\sqrt{\rho}}{\sqrt{(1-\rho)}}\right)\}\right]^i \left[\Phi\left\{\frac{t\sqrt{\rho}}{\sqrt{(1-\rho)}}\right\}\right]^{q-i} \phi(t) dt.$$

Table 1 shows the asymptotic relative efficiency of $\tilde{\rho}$ for $q = 10$ for selected values of ρ. As in Example 1, the loss of efficiency is relatively small, having in this case a maximum of about 15%. Calculations not shown here for smaller values of q show a maximum loss of efficiency of 8% for $q = 8$ and 3% for $q = 5$.

A more general version of the dichotomized normal, in which a latent multivariate normal random vector is categorized into several classes, is treated by the method of pairwise likelihood in deLeon (2003), extending a method proposed by Anderson and Pemberton (1985).

Example 3: 2×2 *tables*

Suppose $X^{(1)}, \ldots, X^{(N)}$ are unobserved binary vectors of length q, and for each pair of positions (s, t) with $s, t = 1, \ldots, q$, we observe the multinomial vector $y = (y_{11}^{st}, y_{10}^{st}, y_{01}^{st}, y_{00}^{st})$, where

$$y_{ij}^{st} = \Sigma_{k=1}^{N} 1\{x_s^{(k)} = i, x_t^{(k)} = j\}, \quad \{i, j\} \in \{0, 1\}$$

and $1\{A\}$ is the indicator function for the event A. The pseudo loglikelihood based on the vector of counts is

$$\ell_2(\rho; y) = \Sigma_{s>t}(y_{11}^{st} \log p_{11}^{st} + y_{10}^{st} \log p_{10}^{st} + y_{01}^{st} \log p_{01}^{st} + y_{00}^{st} \log p_{00}^{st}),$$

where we suppose that $p_{ij}^{st} = p_{ij}^{st}(\rho)$. This is a simplified version of the pairwise loglikelihood discussed in Fearnhead (2003), where s and t represent loci on each of N chromosomes. A specific model for $p_{ij}^{st}(\rho)$ is discussed in McVean et al. (2002). Fearnhead (2003) shows that the estimator of ρ based on a truncated version of $\ell_2(\rho)$ is consistent as $q \to \infty$ but that an approximation to $\ell_2(\rho)$ used in McVean et al. (2002) does not give consistent estimation of ρ. In the latter case consistency fails because the score function from the approximation to $\ell_2(\rho)$ does not have mean 0. In contrast the score function for $\ell_2(\rho)$ does have mean 0: at issue is whether or not under suitable models

for $p_{ij}^{st}(\rho)$ the variance is $O(1)$ or $O(1/q)$ as $q \to \infty$. In an application in which N is the number of individuals, an asymptotic theory in N and q simultaneously would be relevant.

5 Discussion

In many applications, such as the genetics example above, it is prohibitively difficult to compute the full likelihood function. A similar situation arises in the analysis of spatial data, and Besag's pseudolikelihood is obtained from ℓ_2 by choosing $a = 1/2$. Pairwise likelihood for the analysis of spatial data is discussed in Nott and Ryden (1999) and Heagerty and Lele (1998), building on unpublished work by Hjort. These papers also consider the possibility of unequal weighting of the contributions of different pairs. Henderson and Shimakura (2003) use the pairwise likelihood in a model for serially correlated count data for which the full likelihood is intractable.

In other applications the pseudolikelihood function based on pairs of observations may provide a useful model in settings where it is difficult to construct the full joint distribution. One possible application is to the study of multivariate extremes, where models for bivariate extremes such as discussed in Coles (2001, Ch. 8) could be used to construct a pseudolikelihood. Parner (2001) considers modelling pairs of failure times in the context of survival data.

In the examples in which the main parameter of interest is the correlation between two elements of the vector there is no information in the univariate margins. The expression for the variance of $\Sigma_{s>t} U_{2st}$ given in §3 suggests that, if the parameter of interest appears in both the bivariate and univariate margins, it might be possible by suitable choice of a to eliminate the leading term in q in the variance of the score.

The pseudolikelihood for the dichotomised normal example is the likelihood for the quadratic exponential distribution, which has been proposed for the analysis of multivariate binary data (Cox, 1972; Prentice & Zhao, 1990; Cox & Wermuth, 1994). The score function from the quadratic exponential is one version of a generalised estimating equation, as discussed for example in Liang et al. (1992). A feature of generalised estimating equations is that they lead to consistent estimators of parameters in the mean function, even if the covariances are misspecified. Similarly we might expect that use of the pseudolikelihood ℓ_2, for example, would lead to consistent estimators of correlation parameters under a range of possible models for higher-order dependency as incorporated into the full joint distributions. Pairwise likelihood methods for correlated binary data are discussed in Kuk and Nott (2000) and LeCessie and van Houwelingen (1994).

Acknowledgement

The authors are grateful to Anthony Davison for suggesting the application to extreme values and to Gil McVean for details on the genetics application. The research was partially supported by the Natural Sciences and Engineering Research Council of Canada.

References

Anderson, J.A. and Pemberton, J.D. (1985). The grouped continuous model for multivariate ordered categorical variables and covariate adjustment. *Biometrics* **41**, 875–85.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with Discussion). *J. R. Statist. Soc.* B **36**, 192–236.

Coles, S. (2001). *An Introduction to Statistical Modelling of Extreme Values.* New York: Springer-Verlag.

Cox, D.R. (1972). The analysis of multivariate binary data. *Appl. Statist.* **21**, 113–20.

Cox, D.R. & Wermuth, N. (1994). A note on the quadratic exponential binary distribution. *Biometrika* **81**, 403–8.

Fearnhead, P. (2003). Consistency of estimators of the population-scaled recombination rate. *Theor. Pop. Biol.* **64**, 67–79.

Heagerty, P.J. & Lele, S. R. (1998). A composite likelihood approach to binary spatial data. *J. Am. Statist. Assoc.* **93**, 1099–111.

Henderson, R. & Shimakura, S. (2003). A serially correlated gamma frailty model for longitudinal count data. *Biometrika* **90**, 355–66.

Kuk, A.Y.C. & Nott, D.J. (2000). A pairwise likelihood approach to analyzing correlated binary data. *Stat. Prob. Lett.* **47**, 329–35.

LeCessie, S. & van Houwelingen, J.C. (1994). Logistic regression for correlated binary data. *Appl. Statist.* **43**, 95–108.

deLeon, A.R. (2003). Pairwise likelihood approach to grouped continuous model and its extension. *Stat. Prob. Lett.*, to appear.

Liang, K.-Y., Zeger, S.L. & Qaqish, B. (1992). Multivariate regression analysis for categorical data (with Discussion). *J. R. Statist. Soc.* B **54**, 3–40.

Lindsay, B.L. (1988). Composite likelihood methods. *Contemporary Mathematics* **80**, 221–39.

McVean, G., Awadalla, P. & Fearnhead, P. (2002). A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* **160**, 1231–41.

Nott, D.J. and Rydén, T. (1999). Pairwise likelihood methods for inference in image models. *Biometrika* **86**, 661–76.

Parner, E.T. (2001). A composite likelihood approach to multivariate survival data. *Scand. J. Statist.* **28**, 295–302.

Tong, Y.L. (1990). *Multivariate Normal Distribution.* New York: Springer-Verlag.

Zhao, L.P. & Prentice, R.L. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika* **77**, 642–8.

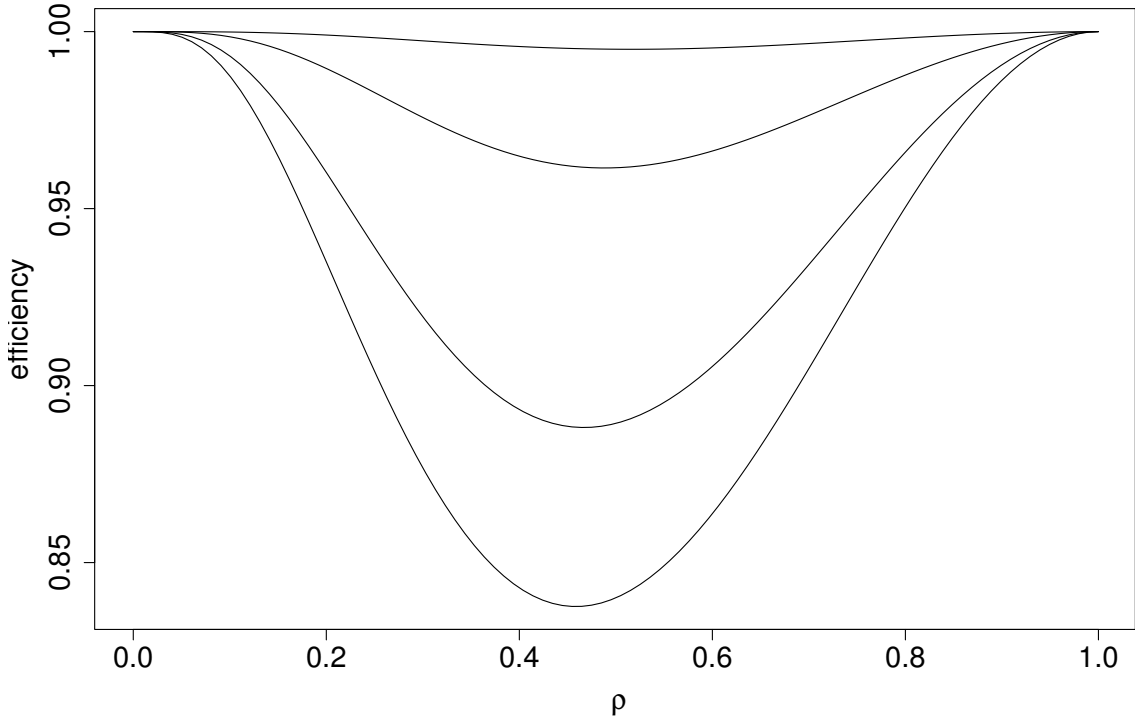

Figure 1: Ratio of asymptotic variance of $\hat{\rho}$ to $\tilde{\rho}_2$, as a function of ρ, for fixed q. At $q = 2$ the ratio is identically 1. The lines shown are for $q = 3, 5, 8, 10$ (descending).

Table 1: Asymptotic relative efficiency (ARE) of $\tilde{\rho}$ for selected values of ρ, with $q = 10$, computed as the ratio of the asymptotic variances.

ρ	0.02	0.05	0.12	0.20	0.40	0.50
avar$\tilde{\rho}$	0.066	0.084	0.128	0.178	0.270	0.285
avar$\hat{\rho}$	0.066	0.084	0.128	0.177	0.261	0.272
ARE	0.998	0.995	0.992	0.968	0.953	0.968
ρ	0.60	0.70	0.80	0.90	0.95	0.98
avar$\tilde{\rho}$	0.161	0.074	0.031	0.009	0.274	0.235
avar$\hat{\rho}$	0.145	0.065	0.027	0.007	0.257	0.212
ARE	0.953	0.903	0.900	0.874	0.867	0.850