Biometrika (2009), **96**, 4, *pp*. 975–982
doi: 10.1093/biomet/asp056
Advance Access publication 29 October 2009

Printed in Great Britain

Maximum likelihood estimation using composite likelihoods for closed exponential families

BY KANTI V. MARDIA, JOHN T. KENT

Department of Statistics, University of Leeds, Leeds, LS2 9JT, U.K.

k.v.mardia@leeds.ac.uk j.t.kent@leeds.ac.uk

GARETH HUGHES

Novartis International AG, CH-4002 Basel, Switzerland

ghughes@live.co.uk

AND CHARLES C. TAYLOR

Department of Statistics, University of Leeds, Leeds, LS2 9JT, U.K.

c.c.taylor@leeds.ac.uk

SUMMARY

In certain multivariate problems the full probability density has an awkward normalizing constant, but the conditional and/or marginal distributions may be much more tractable. In this paper we investigate the use of composite likelihoods instead of the full likelihood. For closed exponential families, both are shown to be maximized by the same parameter values for any number of observations. Examples include log-linear models and multivariate normal models. In other cases the parameter estimate obtained by maximizing a composite likelihood can be viewed as an approximation to the full maximum likelihood estimate. An application is given to an example in directional data based on a bivariate von Mises distribution.

Some key words: Bivariate von Mises distribution; Closed exponential family; Fisher information; Log-linear model; Maximum likelihood; Multivariate normal distribution; Pseudolikelihood.

1. INTRODUCTION

Consider a statistical model $f(x;\theta)$, which can be viewed as a density in x for fixed θ or a likelihood in θ for fixed x. The usual maximum likelihood estimate $\hat{\theta}_{\mathrm{FL}}$ is obtained by maximizing the full likelihood $f(x;\theta)$ over θ. In this paper we look at an alternative to the full likelihood called a composite likelihood.

If x can be partitioned into three pieces x_A, x_B, x_C, say, where B or C may be empty, then the conditional density $f(x_A \mid x_B;\theta)$, or marginal density if $B = \emptyset$, continues to depend on at least part of θ. Given a collection of such partitions, the conditional densities can be multiplied together to yield a composite likelihood, whose maximum over θ can be denoted $\hat{\theta}_{\mathrm{CL}}$.

The purpose of this paper is two-fold: to identify statistical models and types of composite likelihood for which the composite likelihood estimator is identical to the maximum likelihood estimator, and to investigate the use of the $\hat{\theta}_{\mathrm{CL}}$ as a tractable approximation to $\hat{\theta}_{\mathrm{FL}}$ when the full likelihood involves a difficult normalization constant.

The history of composite likelihoods, also called pseudolikelihoods, goes back at least to Besag (1974) who, in the terminology developed below, used a full conditional composite likelihood based on one observation from a binary spatial process. Further details for Gaussian processes were given there and in Besag (1975, 1977).

Later work has often focused on the case of n independent identically distributed observations from a multivariate distribution, especially issues of consistency and asymptotic normality. Lindsay (1988) and Arnold & Strauss (1991a) discuss the general class of composite likelihoods considered here and give many examples. Arnold & Strauss (1991b) and Arnold et al. (2001) investigated the construction of joint distributions given the conditionals and discussed various properties of such distributions. Cox & Reid (2004) focused on the special case of marginal composite likelihoods. Varin (2008) is a recent review paper with a wealth of examples. The main emphasis in the present paper is on n independent identically distributed multivariate observations.

2. CLOSED EXPONENTIAL FAMILIES

Let x be a multivariate quantity that can be split into p pieces, $x = (x_1, \ldots, x_p)$, and consider a product base measure $\mu(dx) = \mu(dx_1) \cdots \mu(dx_p)$. Typically, the jth component of x will contain n observations on variable j. Write $P = \{1, \ldots, p\}$ for the full set of indices. Also, let $t = t(x)$ be a q-dimensional sufficient statistic as a function of x and consider the canonical exponential family

$$f(x;\theta) = \exp\{\theta^{\mathrm{T}} t - c(\theta)\} \tag{1}$$

with respect to this base measure. Here θ is a q-dimensional parameter vector and $c(\theta)$ determines the normalizing constant.

In certain cases an exponential family is closed under marginalization in the following sense.

DEFINITION 1. *A canonical exponential family for x, with sufficient statistic t is said to be closed if, for all subsets $B \subset P$, the marginal distribution of x_B follows a canonical exponential family with sufficient statistic t_B, where t_B is the subset of components of t that depend just on x_B.*

To understand the closure property in more detail, let (A, B) be a partition of P, so $A \cap B = \emptyset$, $A \cup B = P$. Let $t_{A;B}$ denote those components of t depending on x_A including those which also depend on x_B, and let t_B denote those components of t depending just on x_B. Then t and θ can be partitioned as $t = (t_{A;B}^{\mathrm{T}}, t_B^{\mathrm{T}})^{\mathrm{T}}$ and $\theta = (\theta_{A;B}^{\mathrm{T}}, \theta_B^{\mathrm{T}})^{\mathrm{T}}$. The joint distribution of x can always be split into a product of a conditional and a marginal distribution,

$$f(x;\theta) = f(x_A \mid x_B;\theta) f_B(x_B;\theta). \tag{2}$$

Since f is a canonical exponential family, the first factor depends on θ only through $\theta_{A;B}$. Further, under the closure assumption, the marginal density of x_B takes the form

$$f_B(x_B;\theta) = \exp\bigl\{\theta_B^{*\mathrm{T}} t_B - c_B(\theta_B^*)\bigr\},$$

where $\theta_B^* = \phi_B(\theta)$ depends on the full parameter θ through some function $\phi_B(\cdot)$. In the language of Barndorff-Nielsen & Cox (1994, p. 38), t_B is said to be a cut; it is S-ancillary for $\theta_{A;B}$ and S-sufficient for θ_B^*.

The conditional density can be written as

$$f(x_A \mid x_B;\theta) = \exp\bigl\{\theta_{A;B}^{\mathrm{T}} t_{A;B} - d_{A;B}(x_B, \theta_{A;B})\bigr\}. \tag{3}$$

By writing $f(x_A \mid x_B;\theta) = f(x;\theta)/f_B(x_B;\theta)$, the normalizing constant can be expressed as

$$d(x_B, \theta_{A;B}) = \{\phi_B(\theta) - \theta_B\}^{\mathrm{T}} t_B + c(\theta) - c_B\{\phi_B(\theta)\};$$

from (3), it depends on θ only through $\theta_{A;B}$.

In passing, note that we have proved the integral representation

$$\int \exp\bigl(\theta_{A;B}^{\mathrm{T}} t_{A;B}\bigr)\mu(dx_A) = \exp\{\psi_B(\theta_{A;B})^{\mathrm{T}} t_B - g_{A;B}(\theta_{A;B})\},$$

where the functions $\psi_B(\cdot)$ and $g_{A;B}(\cdot)$ are related to $\phi_B(\cdot)$ and $c_B(\cdot)$ by

$$\theta_B^* = \phi_B(\theta) = \theta_B + \psi_B(\theta_{A;B}), \quad c_B(\theta_B^*) = c(\theta) + g_{A;B}(\theta_{A;B}).$$

Note that $c_B(\theta_B^*)$ depends on θ only through $\theta_B^* = \phi_B(\theta)$.

As a simple example let x be an $n \times 2$ matrix representing n observations from a bivariate normal distribution, $N_2(0, \Sigma)$, and let j index the jth variable, $j = 1, 2$. Also, denote the 2×2 sample sum of squares and products matrix about the origin by S. The distinct elements of $-S/2$ form the sufficient statistic in this case. With the partition $A = \{1\}$ and $B = \{2\}$, the sufficient statistic t is partitioned as

$$t_{A;B} = -\frac{1}{2}(s_{11}, 2s_{12})^{\mathrm{T}}, \quad t_B = -\frac{1}{2}s_{22},$$

with canonical parameters $\theta_{A;B} = (\sigma^{11}, \sigma^{12})$, $\theta_B = \sigma^{22}$, in terms of the elements of the inverse covariance matrix $\Sigma^{-1} = (\sigma^{ij})$. In this case $\theta_B^* = \sigma_{22}^{-1} = \{\sigma^{22} - (\sigma^{12})^2/\sigma^{11}\}$.

The joint likelihood (1) can be maximized by maximizing separately the two terms on the right-hand side of (2). More specifically, if $\hat{\theta} = (\hat{\theta}_{A;B}^{\mathrm{T}}, \hat{\theta}_B^{\mathrm{T}})^{\mathrm{T}}$ maximizes the left-hand side, then $\hat{\theta}_{A;B}$ maximizes the first term on the right-hand side and $\hat{\theta}_B^* = \hat{\theta}_B + \psi_B(\hat{\theta}_{A;B})$ maximizes the second term.

Further, because the joint density, the conditional density and the marginal density are all exponential families, these maximum likelihood estimators uniquely match population and sample moments. That is, writing T as a random sufficient statistic with realized value t in the dataset,

$$E_{\hat{\theta}}(T) = t, \tag{4}$$

$$E_{\hat{\theta}_{A;B}, x_B}(T_{A;B}) = t_{A;B}, \tag{5}$$

$$E_{\hat{\theta}_B^*}(T_B) = t_B, \tag{6}$$

where the three expectations are taken over the joint density, depending on θ, the conditional density of x_A given x_B, depending on θ only through $\theta_{A;B}$, and the marginal density of x_B, depending on θ only through $\theta_B^* = \phi_B(\theta)$, respectively.

Another property of a closed exponential family is that the closure property is preserved under marginalization. That is, if $D \subset P$ is a subset of indices, then the marginal distribution of x_D is also closed. To prove this claim, let $B \subset D$. From the definition of closure, the marginal distribution of x_B is a canonical exponential family with sufficient statistic t_B, which is a subset of t, the sufficient statistic for x. Since the components of x_B are a subset of those of x_D, the components of t_B also comprise the subset of the components of t_D, which depend just on x_B. Hence the marginals from x_D satisfy the condition of Definition 1.

This section focused on partitions (A, B). However, many of the results continue to hold when $A \cup B$ is a proper subset of P, in particular, the representation (3) for the conditional density $f(x_A \mid x_B; \theta)$ and the moment results (5) and (6).

3. Composite likelihoods

Consider now a collection of partitions $\{A(\ell), B(\ell), C(\ell)\}$ $(\ell = 1, \ldots, m)$, where $B(\ell)$ or $C(\ell)$ may be empty. When $C(\ell) = \emptyset$, $\{A(\ell), B(\ell)\}$ is a partition; when $B(\ell) = \emptyset$, $f(x_{A(\ell)} \mid x_{B(\ell)}; \theta)$ becomes a marginal density. Define the composite loglikelihood by

$$l_{\mathrm{CL}}(\theta) = \sum_{\ell=1}^{m} \log f\big(x_{A(\ell)} \mid x_{B(\ell)}, \theta\big). \tag{7}$$

Here are several common composite likelihoods.

Full likelihood. For completeness we include the usual full likelihood in this list, with $m = 1$, $A(1) = P$, $B(1) = \emptyset$, $C(1) = \emptyset$.

Full conditional composite likelihood. The simplest partitions to use are the full conditionals, with

$$A(\ell) = \{\ell\}, \quad B(\ell) = \{1, \ldots, p\} \backslash \{\ell\} \quad (\ell = 1, \ldots, p),$$

so that each variable is conditioned on the rest in turn.

Pairwise conditional composite likelihood. This composite likelihood takes the form

$$\text{PCCL}(\theta) = \prod_{i \neq j} f(x_i \mid x_j; \theta),$$

where the product contains $p(p-1)$ factors.

Pairwise marginal composite likelihood. This composite likelihood takes the form

$$\text{PMCL}(\theta) = \prod_{i<j} f\{(x_i, x_j); \theta\},$$

where the product contains $p(p-1)/2$ factors.

The pairwise composite likelihoods are of most interest for exponential families with only first-order interactions, i.e. each component of t depends on at most two components of x.

Denote the parameter value maximizing the composite loglikelihood (7) by $\hat{\theta}_{\text{CL}}$, which may or may not be unique. The following results give conditions to ensure that $\hat{\theta}_{\text{CL}}$ is unique and satisfies $\hat{\theta}_{\text{CL}} = \hat{\theta}_{\text{FL}}$.

THEOREM 1. *Consider data x from a closed exponential family model. If $C(\ell) = \emptyset$ for all ℓ in* (7) *and if each component of t is contained in $t_{A(\ell);B(\ell)}$ for at least one ℓ, then $\hat{\theta}_{\text{CL}}$ is unique and satisfies $\hat{\theta}_{\text{CL}} = \hat{\theta}_{\text{FL}}$.*

Proof. From the decomposition in (2), the ℓth term of the log composite likelihood is maximized by $\hat{\theta}_{\text{FL}}$. Hence the log composite likelihood is maximized by $\hat{\theta}_{\text{FL}}$. All we need to confirm is that $\hat{\theta}_{\text{FL}}$ is unique.

The ℓth term of the composite loglikelihood depends on θ only through $\theta_{A(\ell);B(\ell)}$, and by (5) is maximized if and only if $(\hat{\theta}_{\text{CL}})_{A(\ell);B(\ell)} = (\hat{\theta}_{\text{FL}})_{A(\ell);B(\ell)}$. Combining these results over all ℓ yields $\hat{\theta}_{\text{CL}} = \hat{\theta}_{\text{FL}}$. □

Theorem 1 covers the important case of the full conditional composite likelihood. However, more care is needed for the pairwise composite likelihoods. First, since for $p > 2$ the subsets of indices $C(\ell)$ are not empty, these cases are not covered in Theorem 1. And second, it is necessary to make an assumption that the sufficient statistic t involves only first-order interactions, i.e. each component of t depends on at most two components of x. The following theorem covers the cases of pairwise composite likelihoods.

THEOREM 2. *Consider data x from a closed exponential family model for which only first-order interactions are present. For the pairwise conditional and pairwise marginal composite likelihoods, $\hat{\theta}_{\text{CL}}$ is unique and satisfies $\hat{\theta}_{\text{CL}} = \hat{\theta}_{\text{FL}}$.*

Proof. Fix two indices $i \neq j$, let $B = \{i, j\}$ and consider the bivariate model $f(x_B; \theta)$, which depends on θ only through $\phi_B(\theta)$, in the notation of § 2. Using properties of the full likelihood estimator for the pairwise marginal case, and calling on Theorem 1 for the pairwise conditional case, it follows that $\phi_B(\hat{\theta}_{\text{CL}}) = \phi_B(\hat{\theta}_{\text{FL}})$. Hence from equation (6), it follows that $E_{\hat{\theta}_{\text{CL}}}(T_B) = t_B$.

Each component of t depends on at most two components of x, i and j, say. Letting i, j range through all pairs of indices implies $E_{\hat{\theta}_{\text{CL}}}(T) = t$, and hence from equation (4) that $\hat{\theta}_{\text{CL}}$ is unique and equals $\hat{\theta}_{\text{FL}}$. □

It might be thought that the condition $C(\ell) = \emptyset$ for all ℓ could be dropped in Theorem 1, but the theorem is not true at this level of generality, as the following counterexample shows.

Let the $p \times 1$ vector x denote a single observation from a multivariate normal distribution $N_p(\mu, \Sigma)$, where the unknown mean vector μ is the parameter to estimate, and the covariance matrix Σ is assumed known. Let $A(\ell) = \{\ell\}$ $(\ell = 1, \ldots, p)$, equal each coordinate in turn and let $B(\ell)$ be a subset of the remaining coordinates. Then the composite loglikelihood takes the form

$$-\frac{1}{2}\sum_{\ell=1}^{p}\left\{x_\ell - \mu_\ell - \beta_\ell^{\mathrm{T}}\left(x_{B(\ell)} - \mu_{B(\ell)}\right)\right\}^2 / \sigma^2_{\ell\ell\cdot B(\ell)} = -\frac{1}{2}\sum\left\{\gamma_\ell^{\mathrm{T}}(x-\mu)\right\}^2, \quad \text{say.}$$

Here $\beta_\ell = \Sigma^{-1}_{B(\ell)B(\ell)}\Sigma_{B(\ell)\ell}$ denotes the regression coefficient of x_ℓ on $x_{B(\ell)}$, and $\sigma^2_{\ell\ell\cdot B(\ell)}$ is the residual variance. The coefficient vectors γ_ℓ, treated as row vectors, can be stacked together into a $p \times p$ matrix Γ.

Clearly, the composite loglikelihood is maximized by the usual maximum likelihood estimator $\hat{\mu}_{\mathrm{FL}} = x$. Further, it is easy to see that $\mu = x$ uniquely maximizes the composite loglikelihood if and only if Γ is nonsingular. More specifically, if Γ is singular, then the composite loglikelihood is constant when $x - \mu$ lies in any direction perpendicular to the rows of Γ.

To develop a counterexample, let $p = 3, \sigma_{11} = \sigma_{22} = \sigma_{33} = 1, \sigma_{12} = \sigma_{13} = 3/4, \sigma_{23} = 2/9$ and $B_1 = \{3\}$, $B_2 = \emptyset$, $B_3 = \{1, 2\}$. It turns out that $\det(\Gamma) = 0$.

4. Behaviour of nonclosed models

If a statistical model is not closed, it is of interest to ask how close $\hat{\theta}_{\mathrm{CL}}$ is to $\hat{\theta}_{\mathrm{FL}}$. This question can be addressed most simply in the asymptotic setting as $n \to \infty$ for n independent identically distributed observations from a model with p variables and q parameters.

The standard theory of consistency and asymptotic normality for estimating equations can be applied here. Subject to mild regularity conditions and the identifiability of the parameters, the composite likelihood estimate will be consistent with asymptotic normal distribution $n^{1/2}(\hat{\theta}_{\mathrm{CL}} - \theta) \sim N_q(0, I_{\mathrm{CL}}^{-1})$, where the information matrix takes the sandwich form $I_{\mathrm{CL}} = H^{-1} J H^{-1}$ in terms of matrices $J = E(UU^{\mathrm{T}})$ and $H = -E(\partial U / \partial \theta^{\mathrm{T}})$. Here $U_{\mathrm{CL}}(x; \theta) = \partial \log l_{\mathrm{CL}}(x; \theta)/\partial\theta$ denotes the score vector. For the full likelihood, $H = J = I$, say, reduces to the usual Fisher information matrix. See, e.g. Godambe (1960), Kent (1982), Lindsay (1988) and Varin & Vidoni (2005).

Thus the key question is identifiability, and in general this property does not hold unless the composite likelihood is rich enough to include all the information about θ. For example, for the full conditional composite likelihood, identifiability always holds by the Brook (1964) expansion, which demonstrates that the full conditionals always determine the joint distribution, subject to a mild positivity regularity condition. On the other hand, for the partial conditional and partial marginal composite likelihoods, identifiability only holds for models involving at most first-order interactions. Such models include the multivariate normal distribution, but not a log-linear model for a three-way contingency table, where all the parameters are unknown. This log-linear model allows arbitrary bivariate distributions for the two-way margins, but it also includes second-order interactions, which cannot be identified from the pairwise marginals.

The efficiency of a maximum composite likelihood estimator relative to the maximum full likelihood estimator can be summarized by

$$(|I_{\mathrm{CL}}| / |I|)^{1/q}; \tag{8}$$

see, for example, Davison (2003, p. 113).

5. Examples

5·1. *Contingency tables*

Consider an $m_1 \times m_2$ contingency table for two random variables X and Y with joint probability distribution

$$P(X = i, Y = j) = p_{ij} = \exp\left\{\sum_{k=1}^{m_1}\sum_{l=1}^{m_2} \theta_{kl} I(k = i, l = j) - c(\theta)\right\}$$
$$(i = 1, \ldots, m_1;\ j = 1, \ldots, m_2),$$

where $\theta_{ij} = \log p_{ij}$ and $c(\theta) = \log\{\sum\sum \exp(\theta_{ij})\}$. Here the model is overparameterized since adding a constant to all the θ_{ij} does not change the model. The usual solution is to set one parameter equal to zero, e.g. $\theta_{11} = 0$, with the other parameters unconstrained. The marginal probabilities for Y can be written in the form

$$P(Y = j_0) = p^*_{j_0} = \sum_{i=1}^{m_1} p_{ij_0} = \exp\left\{\sum_j \theta^*_j I(j = j_0) - c^*_B(\theta^*)\right\},$$

say, and hence this canonical exponential family model is closed. A similar argument works for multi-way tables.

5·2. *Unrestricted normal models*

Let the $n \times p$ matrix x denote n observations from a p-dimensional normal model, $N_p(0, \Sigma)$ or $N_p(\mu, \Sigma)$, i.e. with or without a mean parameter, where the covariance matrix is unconstrained. This model is closed since all the marginals are also normal. Further, since the normal model involves only first-order interactions, all the maximum composite likelihood estimators suggested above are identical to the maximum full likelihood estimator.

5·3. *Unrestricted equicovariance normal model*

In the above model, with zero mean for simplicity, suppose $\Sigma = \Sigma_e = \sigma^2(I + \rho L)$ is an equicovariance matrix, where I is the $p \times p$ identity matrix and L is a $p \times p$ matrix whose diagonal entries are all 0 and off-diagonal elements are all 1, with $-1/(p-1) < \rho < 1$. This example does not fall within the setting of Theorem 1 because the sufficient statistics, proportional to

$$s_{\mathrm{d}} = \frac{1}{p}\sum_{j=1}^{p} s_{jj}, \quad s_{\mathrm{o}} = \frac{1}{p(p-1)}\sum_{i \neq j} s_{ij},$$

where $S = x^{\mathrm{T}}x$ denotes the $p \times p$ sample sum of squares and products matrix, do not depend on just one or two components of x, but depend on all p components of x. However, it is still the case that all the composite likelihood estimators suggested above are identical to the full likelihood estimator.

To verify this claim, let equi(S) denote the matrix whose diagonal elements all equal s_{d} and whose off-diagonal elements equal s_{o}. It is easily checked that each of the composite loglikelihoods, including the full loglikelihood, depend on S only through s_{d} and s_{o}. In particular, for any of the loglikelihoods under consideration, $l(\Sigma_e; S) = l\{\Sigma_e; \mathrm{equi}(S)\}$, where Σ_e denotes a covariance matrix with the equicovariance property and the dependence of the likelihood on the data is made explicit. Since the normal family is closed, for each of the composite likelihoods under consideration, the right-hand side is maximized over all unrestricted covariance matrices by $\hat{\Sigma} = \mathrm{equi}(S)/n$. Since this matrix has the equicovariance property, it is also the maximum over restricted covariance matrices Σ_e. Hence all the likelihoods under consideration are maximized by $\hat{\Sigma}_e = \mathrm{equi}(S)/n$.

5·4. *The restricted equicovariance normal model*

As an example where the different composite likelihoods are not equivalent, consider the p-dimensional normal distribution with mean 0 and an equicovariance matrix with common known variances $\sigma^2 = 1$. Hence there is one parameter ρ to estimate. The fact that one of the parameters is known means that this model is no longer a canonical exponential family. In this setting it turns out that there are three distinct estimators: the full likelihood estimator; the pairwise marginal estimator, which is identical to the pairwise conditional estimator; and the full conditional estimator, which is different from the pairwise estimator.

Cox & Reid (2004) studied the pairwise estimator in this context and investigated its asymptotic efficiency. Further theoretical and numerical investigation is given in an unpublished University of Leeds research report by Mardia, Taylor and Hughes. In summary the pairwise and full conditional composite estimators differ for $p \geqslant 3$, but neither estimator dominates the other for all values of ρ and p.

6. THE BIVARIATE VON MISES DISTRIBUTION

The bivariate von Mises distribution on the torus (Singh et al., 2002) is an interesting example of a nonclosed exponential family model which can be regarded as approximately closed. For angular variables $X, Y \in [-\pi, \pi)$, the joint probability density function is given by

$$f(x, y) \propto \exp\{\kappa_1 \cos(x - \mu_1) + \kappa_2 \cos(y - \mu_2) + \lambda \sin(x - \mu_1)\sin(y - \mu_2)\}. \tag{9}$$

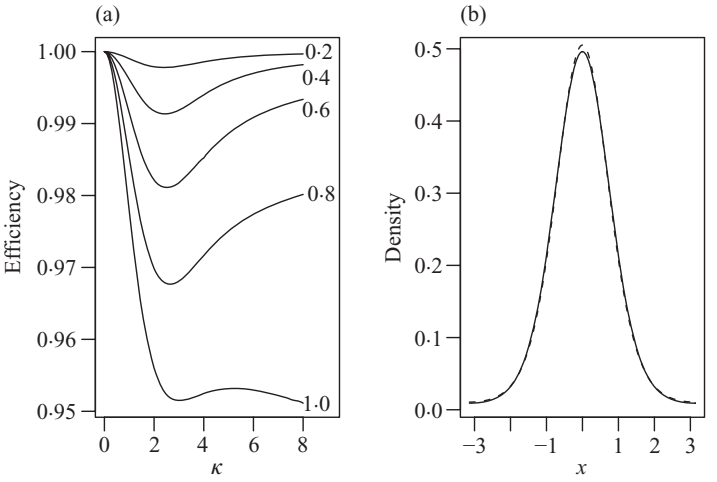

Fig. 1. Bivariate von Mises example. (a) Efficiency of composite likelihood with κ lying between 0 and 8, and $\rho = \lambda/\kappa = 0{\cdot}2, 0{\cdot}4, 0{\cdot}6, 0{\cdot}8, 1{\cdot}0$. (b) Comparison of exact, solid line, and approximate, dashed line, marginal densities for x with parameters $\kappa = 2$, $\rho = 0{\cdot}8$.

Under high concentration, the distribution of $(X - \mu_1, Y - \mu_2)$ is asymptotically bivariate normal with mean 0 and inverse covariance matrix

$$\Sigma^{-1} = \begin{pmatrix} \kappa_1 & -\lambda \\ -\lambda & \kappa_2 \end{pmatrix}, \tag{10}$$

where in general (9) defines a unimodal density on the torus provided (10) is positive definite.

There are five parameters in this exponential family distribution. It is not canonical due to the presence of μ_1 and μ_2 in the interaction term in the exponent, but if μ_1 and μ_2 are known, it becomes a 3-parameter canonical exponential family. The conditional distributions lie in the von Mises family with

$$f(y \mid x) = \{2\pi I_0(\Delta)\}^{-1} \exp\{\kappa_2 \cos(y - \mu_2) + \lambda \sin(x - \mu_1) \sin(y - \mu_2)\},$$

where $I_0(\cdot)$ is a modified Bessel function and $\Delta = \Delta(x; \kappa_2, \lambda, \mu_1) = \{\kappa_2^2 + \lambda^2 \sin^2(x - \mu_1)\}^{1/2}$. However, for $\lambda \neq 0$, the marginal distribution of X is not in the von Mises family, as would be required if the bivariate von Mises family were closed. Instead it has a Bessel density

$$f(x) \propto I_0(\Delta) \exp\{\kappa_1 \cos(x - \mu_1)\}.$$

Hence even if μ_1 and μ_2 are known, this exponential family model is not closed.

The normalizing constant for the bivariate model (9) can be awkward to compute for large parameters, but the Bessel function arising in the conditional von Mises distribution is straightforward to compute quickly in all cases. Hence we investigate the use of the composite likelihood as a substitute for the usual full likelihood.

For illustrative purposes, we limit attention to the case $\mu_1 = \mu_2 = 0$ and $\kappa_1 = \kappa_2 = \kappa$, say. Thus there are two parameters, κ and λ. Efficiency is measured using (8) and is summarized in Fig. 1(a). These results are based on extensive algebraic computations, which are set out in the aforementioned unpublished University of Leeds research report. The efficiency is plotted against κ for various values of $\rho = \lambda/\kappa$, interpreting $\lambda = 0$ whenever $\kappa = 0$. For $\lambda = 0$ the efficiency is always 1, since in this case $f(x, y) = f(x \mid y) f(y \mid x)$. Also, the efficiency tends to 1 in the limiting normal case $\kappa \to \infty$, provided $\rho < 1$.

Note the high efficiency in all cases, except to some extent in the limiting case for a unimodal density, $\rho = 1$. The reason for the high efficiency seems to be that this bivariate exponential family model is approximately closed. For example, Fig. 1(b) shows a plot of the marginal density of X for $x \in [-\pi, \pi]$,

with parameter values, $\kappa = 2$, $\rho = 0{\cdot}8$, together with a von Mises density matched to have the same first cosine moment. It can be seen that the two densities are very close, though further work is needed.

The bivariate von Mises model can be easily extended to a higher-dimensional torus (Mardia et al., 2008). The computational advantages of the full conditional composite likelihood become even more pronounced in this case. Such models have become important in bioinformatics for the modelling of correlated conformational angles in protein structure prediction (Mardia et al., 2007; Boomsma et al., 2008).

The bivariate von Mises example highlights two open issues left for further research. These are to find further examples of closed exponential family models and to formalize the notion of approximately closed.

ACKNOWLEDGEMENT

Gareth Hughes was supported by a U.K. Engineering and Physical Sciences Research Council studentship. We are grateful to Andy Wood for comments that helped to clarify the underlying issues and to the referees for suggestions that improved the presentation.

REFERENCES

ARNOLD, B. C., CASTILLO, E. & SARABIA, J. M. (2001). Conditionally specified distributions: an introduction. *Statist. Sci.* **16**, 249–65.

ARNOLD, B. C. & STRAUSS, D. J. (1991a). Pseudolikelihood estimation: some examples. *Sankhyā* B **53**, 233–43.

ARNOLD, B. C. & STRAUSS, D. J. (1991b). Bivariate distributions with conditionals in prescribed exponential families. *J. R. Statist. Soc.* B **53**, 365–75.

BARNDORFF-NIELSEN, O. E. & COX, D. R. (1994). *Inference and Asymptotics*. London: Chapman and Hall.

BESAG, J. E. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. R. Statist. Soc.* B **34**, 192–236.

BESAG, J. E. (1975). Statistical analysis of non-lattice data. *Statistician* **24**, 179–95.

BESAG, J. E. (1977). Efficiency of pseudolikelihood estimation for simple Gaussian fields. *Biometrika* **64**, 616–18.

BOOMSMA, W., MARDIA, K. V., TAYLOR, C. C., FERKINGHOFF-BORG, J., KROGH, A. & HAMELRYCK, T. (2008). A generative, probabilistic model of local protein structure. *Proc. Nat. Acad. Sci.* **105**, 8932–37.

BROOK, D. (1964). On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems. *Biometrika* **51**, 481–83.

COX, D. R. & REID, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika* **91**, 729–37.

DAVISON, A. C. (2003). *Statistical Models*. Cambridge: Cambridge University Press.

GODAMBE, V. P. (1960). An optimum property of regular maximum likelihood equation. *Ann. Math. Statist.* **31**, 1208–11.

KENT, J. T. (1982). Robust properties of likelihood ratio tests. *Biometrika* **69**, 19–27.

LINDSAY, B. G. (1988). Composite likelihood methods. *Contemp. Math.* **80**, 221–39.

MARDIA, K. V., HUGHES, G., TAYLOR, C. C. & SINGH, H. (2008). Multivariate von Mises distribution with applications to bioinformatics. *Can. J. Statist.* **36**, 99–109.

MARDIA, K. V., TAYLOR, C. C. & SUBRAMANIAM, M. (2007). Protein bioinformatics and mixtures of bivariate von Mises distributions for angular data. *Biometrics* **63**, 505–12.

SINGH, H., HNIZDO, V. & DEMCHUK, E. (2002). Probabilistic model for two dependent circular variables. *Biometrika* **89**, 719–23.

VARIN, C. (2008). On composite marginal likelihoods. *Adv. Statist. Anal.* **92**, 1–28.

VARIN, C. & VIDONI, P. (2005). A note on composite likelihood inference and model selection. *Biometrika* **92**, 519–28.

[*Received April* 2008. *Revised May* 2009]