AN OVERVIEW OF COMPOSITE LIKELIHOOD METHODS

Cristiano Varin, Nancy Reid and David Firth

Ca' Foscari University of Venice, University of Toronto and University of Warwick

Abstract: A survey of recent developments in the theory and application of composite likelihood is provided, building on the review paper of Varin (2008). A range of application areas, including geostatistics, spatial extremes and space-time models as well as clustered and longitudinal data and time series are considered. The important area of applications to statistical genetics is omitted, in light of Larribe and Fearnhead (2010). Emphasis is given to the development of the theory, and the current state of knowledge on efficiency and robustness of composite likelihood inference.

Key words and phrases: copulas, generalized estimating equations, geostatistics, Godambe information, longitudinal data, multivariate binary data, pseudo-likelihood, quasi-likelihood, robustness, spatial extremes, time series

1. Introduction

Composite likelihood is an inference function derived by multiplying a collection of component likelihoods; the particular collection used is often determined by the context. Because each individual component is a conditional or marginal density, the resulting estimating equation obtained from the derivative of the composite log-likelihood is an unbiased estimating equation. Because the components are multiplied, whether or not they are independent, the inference function has the properties of likelihood from a misspecified model. This paper reviews recent work in the area of composite likelihood, reviews the contributions presented at a workshop on composite likelihood held at the University of Warwick in April, 2008, and presents an overview of developments since then. It complements and extends the review of Varin (2008); in particular adding more details on various types of composite likelihood, constructed from marginal and conditional inference, adding yet more application areas, and considering spatial aspects in greater detail. A review of composite likelihood in statistical genetics is given in Larribe and Fearnhead (2010).

In Section 2 we give an overview of the main inferential results for composite likelihood, all based on the asymptotic theory of estimating equations and mis-specified models. Section 3 surveys the wide range of application areas where composite likelihood has been proposed, often under names such as pseudo-likelihood or quasi-likelihood, and Section 4 concentrates on a number of theoretical issues. In Section 5 we consider some of the computational aspects of both construction of, and inference from, composite likelihood, and conclude in Section 6 with a summary of unresolved issues.

2. Composite likelihood inference

2.1 Definitions and notation

Consider an m-dimensional vector random variable Y with probability density function $f(y;\theta)$, for some unknown p-dimensional parameter vector $\theta \in \Theta$. Denote by $\{\mathcal{A}_1, \ldots, \mathcal{A}_K\}$ a set of marginal or conditional events with associated likelihoods $\mathcal{L}_k(\theta; y) \propto f(y \in \mathcal{A}_k; \theta)$. Following Lindsay (1988) a composite likelihood is the weighted product

$$\mathcal{L}_{\mathrm{C}}(\theta; y) = \prod_{k=1}^{K} \mathcal{L}_k(\theta; y)^{w_k},$$

where w_k are nonnegative weights to be chosen. If the weights are all equal then they can be ignored: selection of unequal weights to improve efficiency is discussed in the context of particular applications in §3 and §4.

Although the above definition allows for combinations of marginal and conditional densities (Cox and Reid, 2004), composite likelihoods are typically distinguished in conditional and marginal versions.

Composite conditional likelihoods Perhaps the precedent of composite likelihood is the pseudolikelihood proposed by Besag (1974) for approximate inference in spatial processes. This pseudolikelihood is the product of the conditional densities of a single observation given its neighbours,

$$\mathcal{L}_{\mathrm{C}}(\theta; y) = \prod_{r=1}^{m} f(y_r | \{y_s : y_s \,\text{is neighbour of}\, y_r\}; \theta).$$

More recent variants of Besag's proposal involve using blocks of observations in both conditional and conditioned events, see Vecchia (1988) and Stein et al. (2004).

Liang (1987) studied composite conditional likelihoods of type

$$\mathcal{L}_{\mathrm{C}}(\theta; y) = \prod_{r=1}^{m-1} \prod_{s=r+1}^{m} f(y_r | y_r + y_s; \theta), \tag{2.1}$$

and applied them to stratified case-control studies. Further work on this proposal may be found in Hanfelt (2004), Wang and Williamson (2005) and Fujii and Yanagimoto (2005).

Molenberghs and Verbeke (2005) in the context of longitudinal studies, and Mardia et al. (2008) in bioinformatics, constructed composite likelihoods by pooling pairwise conditional densities

$$\mathcal{L}_{\mathrm{C}}(\theta; y) = \prod_{r=1}^{m} \prod_{s=1}^{m} f(y_r | y_s; \theta),$$

or by pooling full conditional densities

$$\mathcal{L}_{\mathrm{C}}(\theta; y) = \prod_{r=1}^{m} f(y_r | y_{(-r)}; \theta),$$

where $y_{(-r)}$ denotes the vector of all the observations but y_r.

Composite marginal likelihoods The simplest composite marginal likelihood is the pseudolikelihood constructed under working independence assumptions,

$$\mathcal{L}_{\mathrm{ind}}(\theta; y) = \prod_{r=1}^{m} f(y_r; \theta),$$

sometimes referred to in the literature as the independence likelihood (Chandler and Bate, 2007). The independence likelihood permits inference only on marginal parameters. If parameters related to dependence are also of interest it is necessary to model blocks of observations, as in the pairwise likelihood (Cox and Reid, 2004; Varin, 2008)

$$\mathcal{L}_{\mathrm{pair}}(\theta; y) = \prod_{r=1}^{m-1} \prod_{s=r+1}^{m} f(y_r, y_s; \theta), \tag{2.2}$$

and in its extensions constructed from larger sets of observations, see Caragea and Smith (2007).

For continuous symmetric responses with inference focused on the dependence structure, Curriero and Lele (1999) and Lele and Taper (2002) proposed composite marginal likelihoods based on pairwise differences,

$$\mathcal{L}_{\text{diff}}(\theta; y) = \prod_{r=1}^{m-1} \prod_{s=r+1}^{m} f(y_r - y_s; \theta). \tag{2.3}$$

Terminology Composite likelihoods are referred to with several different names, including pseudolikelihood (Molenberghs and Verbeke, 2005), approximate likelihood (Stein et al., 2004), and quasi-likelihood (Hjort and Omre, 1994; Glasbey, 2001; Hjort and Varin, 2008). The first two are too generic to be informative, and the third is a possible source of misunderstanding as it overlaps with a well established alternative (McCullagh, 1983; Wedderburn, 1974). Composite marginal likelihoods in time series are sometimes called split-data likelihoods (Rydén, 1994; Vandekerkhove, 2005). In the psychometric literature, methods based on composite likelihood are called limited information methods. We will consistently use the phrase composite (marginal/conditional) likelihood in this review, and use the notation $\mathcal{L}_{\text{C}}(\cdot)$ and $c\ell(\cdot)$ for the likelihood and log-likelihood function, respectively. If needed we will distinguish marginal, $\mathcal{L}_{\text{MC}}$, and conditional, $\mathcal{L}_{\text{CC}}$, composite likelihoods.

2.2 Derived quantities

The maximum composite likelihood estimator $\hat{\theta}_{\text{CL}}$ locates the maximum of the composite likelihood, or equivalently of the composite log-likelihood $c\ell(\theta; y) = \sum_{k=1}^{K} \ell_k(\theta; y) w_k$, where $\ell_k(\theta; y) = \log \mathcal{L}_k(\theta; y)$. In standard problems $\hat{\theta}_{\text{CL}}$ may be found by solving the composite score function $u(\theta; y) = \nabla_\theta c\ell(\theta; y)$ which is a linear combination of the scores associated with each log-likelihood term $\ell_k(\theta; y)$.

Composite likelihoods may be seen as misspecified likelihoods, where misspecification occurs because of the working independence assumption among the likelihood terms forming the pseudolikelihood. Consequently, the second Bartlett identity does not hold, and we need to distinguish between the sensitivity matrix

$$\text{H}(\theta) = \text{E}_\theta \left\{ -\nabla_\theta \, u(\theta; Y) \right\} = \int \left\{ -\nabla_\theta \, u(\theta; y) \right\} f(y; \theta) \text{d}y$$

and the variability matrix

$$\text{J}(\theta) = \text{var}_\theta \left\{ u(\theta; Y) \right\},$$

and the Fisher information needs to be substituted by the Godambe information matrix (Godambe, 1960)

$$\mathrm{G}(\theta) = \mathrm{H}(\theta)\mathrm{J}(\theta)^{-1}\mathrm{H}(\theta), \tag{2.4}$$

also referred to as the sandwich information matrix. We will reserve the notation $\mathrm{I}(\theta) = \mathrm{var}_\theta\{\nabla_\theta \log f(Y;\theta)\}$ for the expected Fisher information; if $c\ell(\theta)$ is a true log-likelihood function then $\mathrm{G} = \mathrm{H} = \mathrm{I}$. An estimating equation $u(\theta; y)$ which has $\mathrm{H}(\theta) = \mathrm{J}(\theta)$ for all θ is called information unbiased, after Lindsay (1982).

2.3 Asymptotic theory

In the case of n independent and identically distributed observations $Y_1, \ldots, Y_n$ from the model $f(y;\theta)$ on $\mathbb{R}^m$, and $n \to \infty$ with m fixed, some standard asymptotic results are available from Kent (1982), Lindsay (1988) and Molenberghs and Verbeke (2005, Ch. 9), which we now summarize. Since

$$\mathcal{L}_\mathrm{C}(\theta; y) = \prod_{i=1}^{n} \mathcal{L}_\mathrm{C}(\theta; y_i), \text{ and } c\ell(\theta; y) = \sum_{i=1}^{n} c\ell(\theta; y_i),$$

under regularity conditions on the component log-densities we have a central limit theorem for the composite likelihood score statistic, leading to the result that the composite maximum likelihood estimator, $\hat{\theta}_\mathrm{CL}$ is asymptotically normally distributed:

$$\sqrt{n}(\hat{\theta}_\mathrm{CL} - \theta) \xrightarrow{d} N_p\{0, \mathrm{G}^{-1}(\theta)\},$$

where $N_p(\mu, \Sigma)$ is the p-dimensional normal distribution with mean and variance as indicated, and $\mathrm{G}(\theta)$ is the Godambe information matrix in a single observation, defined at (2.4).

The ratio of $\mathrm{G}(\theta)$ to the expected Fisher information $\mathrm{I}(\theta)$ determines the asymptotic efficiency of $\hat{\theta}_\mathrm{CL}$ relative to the maximum likelihood estimator from the full model. If θ is a scalar this can be assessed or plotted over the range of θ; see, for example, Cox and Reid (2004, Fig. 1).

Suppose scientific interest is in a q-dimensional subvector ψ of the parameter $\theta = (\psi, \tau)$. Composite likelihood versions of Wald and score statistics for testing null hypothesis $\mathrm{H}_0 : \psi = \psi_0$ are easily constructed, and have the usual asymptotic χ^2_q distribution, see Molenberghs and Verbeke (2005). The Wald-type statistic is

$$W_e = n(\hat{\psi}_{CL} - \psi_0)^\mathrm{T} \mathrm{G}_{\psi\psi}(\hat{\theta}_{CL})(\hat{\psi}_{CL} - \psi_0),$$

where $\mathrm{G}_{\psi\psi}$ is the $q \times q$ submatrix of the Godambe information pertaining to ψ. The score-type statistic is

$$W_u = \frac{1}{n} u_\psi \left\{\psi_0, \hat{\tau}_{\mathrm{CL}}(\psi_0)\right\}^{\mathrm{T}} \widetilde{\mathrm{H}}^{\psi\psi} \widetilde{\mathrm{G}}_{\psi\psi} \widetilde{\mathrm{H}}^{\psi\psi} u_\psi \left\{\psi_0, \hat{\tau}_{\mathrm{CL}}(\psi_0)\right\},$$

where $\mathrm{H}^{\psi\psi}$ is the $q \times q$ submatrix of the inverse of $\mathrm{H}(\theta)$ pertaining to ψ, and $\widetilde{\mathrm{H}} = \mathrm{H}\{\psi_0, \hat{\tau}_{\mathrm{CL}}(\psi_0)\}$. As in ordinary likelihood inference W_e and W_u suffer from practical limitations: W_e is not invariant to reparametrization, while W_u may be numerically unstable. In addition, estimates of the variability and sensitivity matrices $\mathrm{H}(\theta)$ and $\mathrm{J}(\theta)$ are needed. While they can sometimes be evaluated explicitly, it is more usual to use empirical estimates. As $\mathrm{H}(\theta)$ is a mean, its empirical estimation is straightforward, but the empirical estimation of $\mathrm{J}(\theta)$ requires some internal replication; see §5.

The composite likelihood ratio statistic

$$W = 2\left[c\ell(\hat{\theta}_{\mathrm{CL}}; y) - c\ell\left\{\psi_0, \hat{\tau}_{\mathrm{CL}}(\psi_0); y\right\}\right] \tag{2.5}$$

seems preferable, but it has the drawback of a non-standard asymptotic distribution

$$W \stackrel{d}{\to} \sum_{j=1}^{q} \lambda_j Z_j^2,$$

where $Z_1, \ldots, Z_q$ are independent normal variates and $\lambda_1, \ldots, \lambda_q$ are the eigenvalues of the matrix $(\mathrm{H}^{\psi\psi})^{-1} G^{\psi\psi}$. This result may be derived under the general framework of misspecified likelihoods, see Kent (1982) and the book length exposition of White (1994).

Geys et al. (1999) proposed the adjusted composite likelihood ratio statistic $W' = W/\bar{\lambda}$ with an approximate χ^2_q distribution, where $\bar{\lambda}$ denotes the average of the eigenvalues λ_j; Rotnitzky and Jewell (1990) suggested this for the independence likelihood. The mean of W' coincides with that of its asymptotic χ^2_q distribution, but higher order moments differ. A better solution is provided by a Satterthwaite (1946) adjustment $W'' = \nu W/(q\bar{\lambda})$ with approximate χ^2_ν distribution, where the rescaling and the effective degrees of freedom $\nu = (\sum_{j=1}^q \lambda_j)^2 / \sum_{j=1}^q \lambda_j^2$ are chosen so that the mean and the variance of W'' coincide with that of the approximate distribution (Varin, 2008; Lindsay et al., 2000).

Chandler and Bate (2007) proposed a different type of adjustment for the independence likelihood: essentially stretching the composite log-likelihood on

the θ-axis, about $\hat{\theta}_{\mathrm{CL}}$ to ensure, at least approximately, that the second Bartlett identity holds, and thus that the usual χ^2_q approximation can be used. Vertical rescaling is another possibility, discussed briefly in Chandler and Bate (2007, §6), and extended to composite likelihood in Pace et al. (2010). In the scalar parameter case, vertical rescaling is the same as dividing the composite log-likelihood ratio statistic by $\mathrm{J}^{-1}\mathrm{H}$.

Saddlepoint approximations for quadratic forms are derived in Kuonen (1999), and seem directly applicable to W, but we are not aware of detailed discussion of this application.

The computational simplicity of composite likelihood functions in typical situations allows use of the parametric bootstrap. This has the advantage of working also in non-standard settings, such as when the parameter under the null hypothesis lies on the boundary of the parametric space (Bellio and Varin, 2005), but it has the drawback of requiring the complete specification of a joint model for the data, thus losing in model robustness.

Analogues of the Akaike (AIC) and the Bayesian (BIC) information criteria for model selection are easily derived in the framework of composite likelihoods. They have usual forms $\mathrm{AIC} = -2c\ell(\hat{\theta}_{\mathrm{CL}}; y) + 2\dim(\theta)$ and $\mathrm{BIC} = -2c\ell(\hat{\theta}_{\mathrm{CL}}; y) + \dim(\theta)\log n$, where $\dim(\theta)$ is an effective number of parameters, estimated from the sensitivity matrix and the Godambe information: $\dim(\theta) = \mathrm{tr}\left\{\mathrm{H}(\theta)\mathrm{G}(\theta)^{-1}\right\}$. Formal derivation of these information criteria may be found in Varin and Vidoni (2005) for the composite AIC criterion and in Gao and Song (2009) for the composite BIC criterion.

These criteria may be used for model averaging (Claeskens and Hjort, 2008), or for selection of tuning parameters in shrinkage methods. See Gao and Song (2009) for examples of the Lasso penalty with composite marginal likelihoods.

The inference in the previous section follows directly from the usual asymptotic theory, under standard regularity conditions. It is also of interest to consider the case where n is fixed and m increases, as in the case of a single ($n = 1$) long time series or a spatial dataset. In this case the asymptotic theory depends on the availability of internal replication: for example in an autoregressive model of small-ish order, there is sufficient independence in a single long series to obtain a central limit result.

The asymptotic variance for pairwise likelihood and a modified version of it was treated in Cox and Reid (2004), using Taylor series expansions. Since the validity of these expansions depends on the consistency of θ, which does not hold in general for $m \to \infty$, the argument was purely informal, and a more rigorous treatment is needed. Cox and Reid (2004) also suggested that it may be possible to choose $a \neq 0$ in the composite log-likelihood $\ell_{\mathrm{C}}(\theta) = \ell_{\mathrm{pair}}(\theta) - am\ell_{\mathrm{ind}}(\theta)$ to ensure consistency as $m \to \infty$ for fixed n, but to our knowledge no examples of this strategy have been investigated.

3. Applications

3.1 Gaussian random fields

Geostatistical models for large datasets are increasingly common, particularly with the use of automatic collection methods such as remote sensing, and composite likelihood methods for approximate inference are very appealing. A typical model in geostatistics applications is a Gaussian random field $Y = \{Y(s) : s \in \mathcal{S} \subset \mathbb{R}^2\}$ with mean function $\mu(s)$ and covariance matrix $\Sigma(\theta)$ whose entries reflect spatial correlation; Cressie (1993) gives several examples of parametric spatial correlation functions. Classical geostatistics estimation of θ is based on various methods of curve fitting to the sample variogram (Cressie, 1993). These methods have been strongly criticised, as there is considerable arbitrariness in tuning the fitting algorithms, and the resulting estimators are often inefficient (Diggle and Ribeiro, 2007, §6.3). Maximum likelihood estimation would be more efficient, but requires the inversion of the covariance matrix $\Sigma(\theta)$, usually with a computational cost of order $\mathcal{O}(n^3)$. This cost is prohibitive with many modern spatial, or spatio-temporal, data sets.

Stemming from the work by Besag (1974), Vecchia (1988) proposed approximating the full likelihood with the composite conditional likelihood

$$\mathcal{L}_{\mathrm{CC}}(\theta; y) = f(y_1; \theta) \prod_{i=2}^{n} f(y_i | \mathcal{B}_i; \theta)$$

where $\mathcal{B}_i$ is a subset of $\{y_{i-1}, \ldots, y_1\}$ chosen so as to make feasible the computation of $\mathcal{L}_{\mathrm{C}}$. Vecchia (1988) suggested restricting $\mathcal{B}(y_i)$ to a number of neighbours of y_i. The use of this composite conditional likelihood is illustrated in Vecchia (1988) by the spatial analysis of water levels in 93 observation wells from an aquifer in the Saratoga Valley in Wyoming.

Stein et al. (2004) further developed Vecchia's proposal, and used it to approximate the restricted likelihood function. The authors show that statistical efficiency can be improved using blocks of observations in place of single observations,

$$\mathcal{L}_{\mathrm{CC}}(\theta; y) = f(z_1; \theta) \prod_{i=2}^{b} f(z_i | \mathcal{B}_i'; \theta)$$

where $z_1, \ldots, z_b$ are b blocks of data and $\mathcal{B}_i'$ is a subset of $\{z_{i-1}, \ldots, z_1\}$. This approximate restricted likelihood method is used in Stein et al. (2004) to analyse a data set of over $13,000$ measurements of levels of chlorophyll in Lake Michigan. The measurements were made in a highly irregular pattern, which creates some challenges in the choice of conditioning sets. It was found that including some distant observations in the conditioning sets leads to a remarkable improvement in the efficiency of the composite likelihood parameter estimators.

Difficulties with the composite likelihoods of Stein et al. (2004) and Vecchia (1988) arise with the selection of the observation order and of the conditioning sets $\mathcal{B}_i$ and $\mathcal{B}_i'$. To overcome such complications, in the pair of papers Caragea and Smith (2006; 2007) three different likelihood approximations all based on splitting the data into blocks are proposed. The first method, the "big blocks likelihood", consists in estimating θ from the joint density of the block means. The second method is denoted "small blocks" and it is the composite marginal likelihood formed by the product of densities for all the observations in each block,

$$\mathcal{L}_{\mathrm{MC}}(\theta; y) = \prod_{i=1}^{b} f(z_i; \theta),$$

where $z_1, \ldots, z_b$ are b blocks of data. Hence, while the big blocks likelihood captures large-sample properties of the process but ignores the within blocks dependence, the small blocks method does the opposite. A proposed compromise between the two, called a hybrid method, is to use the big blocks likelihood multiplied by the composite conditional likelihood formed by the product of conditional densities of the observations within blocks, conditioned on the block mean. Efficiency studies indicate poor performance of the big blocks method, while the small blocks and the hybrid methods work similarly with high efficiency. Caragea and Smith (2006) illustrated the good behaviour of the last two methods

for spatial estimation of rainfall trends across the south-central U.S.A.

A major reason for concern with maximum likelihood estimation is the difficulty in checking the assumption of multivariate normality. This difficulty is also shared by these blockwise strategies. In contrast, the pairwise likelihood (2.2) and the composite likelihood of differences (2.3) just require bivariate normality for pairs of observations, which is much simpler to validate. Pairwise likelihood was suggested for inference in geostatistical models first in Hjort and Omre (1994) and then further developed for image models by Nott and Rydén (1999). The composite likelihood based on differences (2.3) was proposed by Curriero and Lele (1999) and applied to temperature data in three dimensional geothermal fields in Mateu et al. (2007).

3.2 Spatial extremes

The rise in hazardous environmental events leads to much interest in statistical modelling of spatial extremes. A flexible approach to this problem is provided by max-stable models obtained from underlying Gaussian random fields constructed by building on unpublished work of Smith (1990). Despite the attractive properties of these models, both classical and Bayesian inference are impractical because of the curse of dimensionality with the likelihood computation, see Davison and Gholamrezaee (2009). At the present time, only expressions for bivariate marginal densities have been derived. Thus, pairwise likelihood inference is naturally considered as a surrogate for impossible ordinary likelihood analysis in Davison and Gholamrezaee (2009) and Padoan et al. (2010) with applications to maximum temperatures in Switzerland and maximum precipitations in the U.S.A., respectively. In both of those papers, computations are carried out with the `R` (R Development Core Team, 2009) package `SpatialExtremes` by Ribatet (2009), which seems to be the first publicly available software implementing composite likelihood methods.

A related approach is followed by Smith and Stephenson (2009) where the pairwise likelihood is used in place of the unfeasible ordinary likelihood for Bayesian inference in max-stable spatial processes. The approach is illustrated through analysis of annual maximum rainfall data in South-West England.

3.3 Serially correlated random effects

In longitudinal and panel studies, random effects models are popular choices

for modelling unobserved heterogeneity. In these models the outcomes are modelled as independent variables conditionally upon a subject-specific random effect, usually assumed to be constant for all the measurements. The latter assumption may be unrealistic in most cases: better models should take into account also the possible serial dependence within subject-specific measurements.

Consider longitudinal counts Y_{ij} observed at occasion $j = 1, \ldots, m_i$ for subject $i = 1, \ldots, n$. This type of data may be naturally modelled as conditionally independent Poisson variables

$$Y_{ij}|U_i \sim \text{Po}\{U_i \exp(x_{ij}^{\text{T}}\beta)\},$$

where U_i is a random effect, x_{ij} is a covariate vector and β are unknown regression coefficients. A common assumption is that $U_1, \ldots, U_n$ are independendent Gamma variables with unit mean. Accordingly, the marginal distribution of Y_{ij} is negative binomial. To account for serial dependence Henderson and Shimakura (2003) suggested to extend the above model by assuming instead different Gamma-distributed random effects U_{ij} for each measurement,

$$Y_{ij}|U_{ij} \sim \text{Po}\{U_{ij} \exp(x_{ij}^{\text{T}}\beta)\},$$

and to specify the joint distribution of U_{ij} to describe the serial dependence. For example, Henderson and Shimakura (2003) proposed an autoregressive dependence of type

$$\text{cor}(U_{ij}, U_{i'k}) = \begin{cases} \rho^{|j-k|} & \text{if } i = i' \\ 0 & \text{if } i \neq i'. \end{cases}$$

Unfortunately, the higher model flexibility of the above formulation is paid for in terms of computational complexity. The likelihood function involves a number of terms exponentially increasing with the series length m_i. Likelihood computation is impractical except in low dimensions. Hence, Henderson and Shimakura (2003) proposed that inference be based on the pairwise likelihood

$$\mathcal{L}_{\text{pair}}(\theta; y) = \prod_{i=1}^{n} \frac{1}{m_i - 1} \prod_{j=1}^{m_i - 1} \prod_{k=j+1}^{m_i} f(y_{ij}, y_{ik}; \theta).$$

The weights $1/(m_i - 1)$ are used to match the ordinary likelihood in the case of independence, as suggested in LeCessie and van Houwelingen (1994). Henderson and Shimakura (2003) illustrated pairwise likelihood inference for the above

model by the analysis of a clinical trial on the number of analgesic doses taken by hospital patients in successive time intervals following abdominal surgery.

A further development of the Henderson and Shimakura work is provided by Fiocco et al. (2009) who modified the autoregressive Gamma process U_{ij} to enhance numerical stability when large counts are involved. Furthermore, Fiocco et al. (2009) suggest a two-step composite likelihood analysis where first regression and oversdispersion parameters are estimated from the independence likelihood, and then dependence parameters are obtained separately from the pairwise likelihood. In their simulation studies, Fiocco et al. (2009) found that this two-step approach loses little in efficiency with respect to joint estimation of all the parameters from the pairwise likelihood and applied this approach to a meta-analysis study for survival curves.

A motivation similar to that of Henderson and Shimakura (2003) and Fiocco et al. (2009) underlies the work by Varin and Czado (2010) who suggested an autoregressive mixed probit model for ordinal and binary longitudinal outcomes. The response Y_{ij} is viewed as a censored version of a continuous unobserved variable Y_{ij}^*,

$$Y_{ij} = y_{ij} \quad \leftrightarrow \quad \alpha_{y_{ij}-1} < Y_{ij}^* \leq \alpha_{y_{ij}}, \quad y_{ij} \in \{1, \ldots, h\},$$

where $-\infty \equiv \alpha_0 < \alpha_1 < \ldots < \alpha_{h-1} < \alpha_h \equiv \infty$ are suitable threshold parameters. The unobserved Y_{ij}^* is modelled with a normal linear mixed model

$$Y_{ij}^* = x_{ij}^{\mathrm{T}}\beta + U_i + \epsilon_{ij},$$

where $U_1, \ldots, U_n$ are n independent normal distributed random effects with zero mean and variance σ^2. To account for serial dependence, the errors ϵ_{ij} are assumed to be generated from an autoregressive process of order one,

$$\epsilon_{ij} = \rho\epsilon_{i,j-1} + (1-\rho^2)^{1/2}\,\eta_{ij}$$

where η_{ij} are independent standard normal innovations. Accordingly, the likelihood function is a product of n rectangular normal probabilities of dimensions $m_1, \ldots, m_n$. With the exception of longitudinal studies with a small number of measurements m_i, the evaluation of the likelihood requires time-consuming Monte Carlo methods with possible numerical instabilities. Hence, Varin and

Czado (2010) proposed the use of pairwise likelihood inference based on pairs of observations less than q units apart,

$$\mathcal{L}^{(q)}_{\text{pair}}(\theta; y) = \prod_{i=1}^{n} \prod_{j=q}^{m_i-1} \prod_{d=1}^{q} f(y_{ij}, y_{ij-d}; \theta).$$

The bivariate probabilities $f(y_{ij}, y_{ij-d}; \theta)$ are easily computed with very precise deterministic quadrature methods available in standard statistical software, thus avoiding the need for simulations. This work is motivated by the analysis of a longitudinal study on the determinants of headache severity. The longitudinal data consist of pain severity diaries compiled by the patients four times a day for periods of different lengths, from four days to almost one year of consecutive measurements. The outcome is the severity of headache measured on an ordinal scale with six levels. Covariate data included personal and clinical information, as well as weather conditions.

3.4 Spatially correlated random effects

The numerical difficulties described in case of serially correlated random effects further increase with spatially correlated random effects. Consider a generalized linear model with linear predictor

$$g\{\mathrm{E}(Y(s)\} = x(s)^{\mathrm{T}}\beta + U(s), \quad s \in \mathcal{S} \subset \mathbb{R}^2,$$

where g is a suitable link function and $\{U(s) : s \in \mathcal{S}\}$ is a zero-mean stationary Gaussian random field. Models of this type are termed generalized linear geostatistical models in Diggle and Ribeiro (2007). Given n observed locations $s_1, \ldots, s_n$, the likelihood function is expressed in terms of a single n dimensional integral,

$$\mathcal{L}(\theta; y) = \int_{\mathbb{R}^n} \prod_{i=1}^{n} f\{y(s_i)|u(s_i); \theta\} f\{u(s_1), \ldots, u(s_n); \theta\} \mathrm{d}u(s_1) \ldots \mathrm{d}u(s_n),$$

whose approximation may be difficult even for moderate n. Typical solutions are based on simulation algorithms such as Monte Carlo expectation-maximization or Markov chain Monte Carlo algorithms, see Diggle and Ribeiro (2007) for references. For large data sets, simulation methods become very demanding and thus pairwise likelihood represents an effective alternative. This was first studied

by Heagerty and Lele (1998) for binary data with probit link. They proposed a weighted pairwise likelihood

$$\mathcal{L}^{(q)}_{\text{pair}}(\theta; y) = \prod_{i=1}^{n-1} \prod_{j=i+1}^{n} f\{y(s_i), y(s_j); \theta\}^{w(\|s_i - s_j\|_2; q)},$$

with dummy weights, $w(d; q) = 1$ if the distance d is less than q and $w(d; q) = 0$ otherwise, used to exclude pairs of observations more than q units apart. Heagerty and Lele (1998) used these ideas for spatial modelling of defoliation by gypsy moths in Massachusetts.

Varin et al. (2005) investigate pairwise likelihood for generalized linear models and suggest that excluding pairs formed by observations too distant may be not only numerically efficient but also statistically efficient. Apanasovich et al. (2008) consider pairwise likelihood inference for logistic regression with a linear predictor expressed by the sum of a nonparametric term and a spatially correlated random effect accounting for short range dependence. This last work is motivated by spatial modelling of aberrant crypt foci in colon carcinogenesis.

3.5 Joint mixed models

Correlated random effects are also used for joint modelling of multivariate longitudinal profiles. Let $(Y_{ij}^{(1)}, \ldots, Y_{ij}^{(D)})^{\mathrm{T}}$ be a vector of D outcomes for subject $i = 1, \ldots, n$ at occasion $j = 1, \ldots, m_i$. A possible modelling strategy for data of this type consists in assuming a mixed model for each single outcome and then modelling the association among the outcomes with a suitable covariance matrix for the random effects. Suppose for ease of exposition a random intercept generalized linear model for each outcome,

$$g\{\mathrm{E}(Y_{ij}^{(d)})\} = x_{ij}^{\mathrm{T}}\beta + U_i^{(d)}, \quad d = 1, \ldots, D,$$

where $U_i^{(d)}$ is a random effect specific for outcome d and subject i $(i = 1, \ldots, n)$. The various univariate mixed models can be combined by assuming a D-dimensional multivariate normal distribution for all the random effects, $U_i^{(1)}, \ldots, U_i^{(D)}$, of a single subject $(i = 1, \ldots, n)$.

With the assumption of independence among different subjects, the likelihood is

$$\mathcal{L}(\theta; y) = \prod_{i=1}^{n} \mathcal{L}_i(\theta; y_i^{(1)}, \ldots, y_i^{(D)}),$$

with $y_i^{(d)} = (y_{i1}^{(d)}, \ldots, y_{im_i}^{(d)})^{\mathrm{T}}$ indicating the vector of all observations of outcome d for subject i. When the dimension D of the outcomes increases, the number of random effects parameters, $\binom{D}{2} + D$, grows quadratically, making the maximization of the likelihood quickly out of reach even in the case of normal linear mixed models where the likelihood has an analytic form.

Molenberghs and Verbeke (2005, §25) proposed to alleviate these computational difficulties by the method of "pairwise fitting". Consider the composite marginal likelihood constructed from all pairs of outcomes,

$$\mathcal{L}_{\mathrm{MC}}(\theta_{1,2}, \ldots, \theta_{D-1,D}; y) = \prod_{r=1}^{D-1} \prod_{s=r+1}^{D} \mathcal{L}(\theta_{r,s}; y^{(r)}, y^{(s)}), \tag{3.1}$$

where $\mathcal{L}(\theta_{r,s}; y^{(r)}, y^{(s)})$ is the likelihood based on outcomes r and s only. In contrast to previously discussed composite likelihoods, here different pair-specific parameters are assumed, *i.e.* $\theta_{r,s}$ is the subset of θ pertaining to the assumed distribution of $(Y^{(r)}, Y^{(s)})$. This separate parameterization is necessary to allow distinct maximization of each term $\mathcal{L}(\theta_{r,s}; y^{(r)}, y^{(s)})$ forming the composite likelihood (3.1), and thus resolve the computational difficulties associated with joint maximization.

Let $\omega = (\theta_{1,2}, \ldots, \theta_{D-1,D})^{\mathrm{T}}$ be the vector containing all the $\binom{D}{2}$ pair-specific parameters. Then, $\hat{\omega} = (\hat{\theta}_{1,2}, \ldots, \hat{\theta}_{D-1,D})^{\mathrm{T}}$ locates the maximum of the composite likelihood (3.1). Accordingly, we have

$$\sqrt{n}(\hat{\omega} - \omega) \xrightarrow{d} \mathcal{N}\{0, \mathrm{G}^{-1}(\omega)\}.$$

Obviously, there is a one-to-many correspondence between ω and the original parameter θ, for example, $\theta_{r,s}$ and $\theta_{r,t}$ have some components of θ in common. A single estimate of θ may then be obtained by averaging all the corresponding pair-specific estimates in $\hat{\omega}$. If we denote by A the weight matrix such that $\hat{\theta} = \mathrm{A}\hat{\omega}$, then inference will be based on the asymptotic distribution

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}\{0, \mathrm{A}^{\mathrm{T}}\mathrm{G}^{-1}(\omega)\mathrm{A}\}.$$

Further details of the pairwise fitting method can be found in a series of papers by S. Fieuws and his colleagues with applications to multivariate longitudinal profiles of hearing thresholds (Fieuws and Verbeke, 2006; Fieuws, Verbeke and

Molenberghs, 2007), batteries of binary questionnaires on psychocognitive functioning (Fieuws et al., 2006; Fieuws, Verbeke and Molenberghs, 2007) and joint analysis of several biochemical and physiological markers for failure of renal graft (Fieuws, Verbeke, Maes and Vanrenterghem, 2007). Barry and Bowman (2008) applied the pairwise fitting method to a longitudinal study designed to compare the facial shapes of a group of 49 infants with unilateral cleft lip and palate with those of a group of 100 age-matched controls.

3.6 Time-varying correlation matrices

Engle et al. (2009) proposed composite likelihood methods for risk management with high dimensional portfolios. Consider a K dimensional vector of log-returns r_t observed at times $t = 1, \ldots, T$. Risk management models assume that r_t is the martingale difference sequence

$$\mathrm{E}(r_t|\mathcal{F}_{t-1}) = 0, \quad \mathrm{Cov}(r_t|\mathcal{F}_{t-1}) = \mathrm{H}_t,$$

where $\mathcal{F}_{t-1}$ is the information up to time $t-1$ and H_t is a time-varying covariance matrix. Models proposed for H_t are parametrized in terms of dynamics parameters of interest θ and of nuisance parameters λ. Standard inference is based on a two-step approach. First, nuisance parameters are estimated using a method of moments. Then, parameters of interest are obtained by maximizing a misspecified likelihood constructed under working assumptions of multinormality with the nuisance parameters kept fixed at their moment-based estimates.

There are two sources of difficulty with the above fitting method. First, the method needs the inversion of T correlation matrices H_t, each requiring $\mathcal{O}(K^3)$ operations. Secondly, even if these inversions were possible, the precision of the resulting estimators for θ would quickly fail because the dimension of nuisance parameters grows as the number of assets K increases.

In order to overcome these difficulties, Engle et al. (2009) investigate the use of composite marginal likelihoods formed by summing up (misspecified) log-likelihoods of subsets of assets. This approach resolves the numerical difficulties connected with the inversion of high dimensional matrices. The problem of increasing numbers of nuisance parameters is addressed by using moment estimators for the nuisance parameters specific to each asset, and assuming a common set of parameters across assets; these common parameters are estimated by composite likelihood. This methodology is developed further for composite likelihood

analysis of a panel of GARCH models in Pakel et al. (2009), and simulation studies indicate that composite likelihood methods work very well when there are a large number of short series.

3.7 Marginal regression models with missing data

Statistical analysis of longitudinal data is complicated by the likely occurrence of missing responses. The popular method of generalized estimating equations (GEEs) devised by Liang and Zeger (1986) provides valid inference under the assumption of ignorable missing data (missing completely at random). Complications arise when such an assumption cannot be trusted. If the weaker assumption of missing-at-random is valid, then GEEs can be saved by the use of inverse probability weights as in Robins (1995). A difficulty with this strategy is that it requires correct specification of the missing data process, something that can be awkward in practice. Alternatively, one may base inference on the observed likelihood. However, this strategy suffers from lack of robustness because it relies on correct specification of the joint distribution of all observed responses.

In this context, composite likelihood methods are attractive as a compromise between advantages from likelihood-type analysis and robustness to model specification of GEEs. In the following lines, we summarize composite likelihood inference for marginal regression in presence of noningnorable missing data.

If only parameters in the univariate margins are of interest, Troxel et al. (2003) suggested basing inference under missing at random assumptions on the following independence likelihood:

$$\begin{aligned}\mathcal{L}_{\text{ind}}(\beta,\gamma;y,r) &= \prod_{i=1}^{n}\prod_{j=1}^{m_i}\{f(y_{ij},r_{ij};\beta,\gamma)\}^{r_{ij}}\left\{\int_{y_{ij}} f(y_{ij},r_{ij};\beta,\gamma)\mathrm{d}y_{ij}\right\}^{1-r_{ij}}\\ &= \prod_{i=1}^{n}\prod_{j=1}^{m_i}\{f(y_{ij};\beta)\pi_{ij}(\gamma)\}^{r_{ij}}\left[\int_{y_{ij}} f(y_{ij};\beta)\{1-\pi_{ij}(\gamma)\}\mathrm{d}y_{ij}\right]^{1-r_{ij}}\end{aligned}$$

where β are marginal regression parameters, r_{ij} indicates whether observation j on subject i has been observed or not and $\pi_{ij}(\gamma)$ is the probability of having observed it modelled as a function of parameter γ. This independence likelihood thus requires only the correct specification of univariate margins $f(y_{ij};\beta)$ and of observation probabilities $\pi_{ij}(\gamma)$. This approach is applied in Troxel et al. (2003) to evaluation of adjuvant chemotherapy following surgery in a longitudinal study

of 430 breast cancer patients with up to 37% missing responses. See also Parzen et al. (2006) for another illustration using data from the well-known Six Cities longitudinal study on the health effects of air pollution.

In situations where the association between responses is substantial, this independence likelihood may lead to sensible, but inefficient, inferences on regressors β. For such situations, Parzen et al. (2007) suggest to incorporate information about dependence by moving to the pairwise likelihood

$$\mathcal{L}_{\text{pair}}(\beta, \alpha, \gamma; y, r) = \mathcal{L}_1 \times \mathcal{L}_2 \times \mathcal{L}_3 \times \mathcal{L}_4,$$

with

$$\mathcal{L}_1 = \prod_{i=1}^{n} \prod_{j=1}^{m_i-1} \prod_{k=j+1}^{m_i} \{f(y_{ij}, y_{ik}, r_{ij}, r_{ik}; \beta, \alpha, \gamma)\}^{r_{ij}r_{ik}},$$

$$\mathcal{L}_2 = \prod_{i=1}^{n} \prod_{j=1}^{m_i-1} \prod_{k=j+1}^{m_i} \left\{ \int_{y_{ij}} f(y_{ij}, y_{ik}, r_{ij}, r_{ik}; \beta, \alpha, \gamma) \mathrm{d}y_{ij} \right\}^{(1-r_{ij})r_{ik}},$$

$$\mathcal{L}_3 = \prod_{i=1}^{n} \prod_{j=1}^{m_i-1} \prod_{k=j+1}^{m_i} \left\{ \int_{y_{ik}} f(y_{ij}, y_{ik}, r_{ij}, r_{ik}; \beta, \alpha, \gamma) \mathrm{d}y_{ij} \right\}^{r_{ij}(1-r_{ik})},$$

$$\mathcal{L}_4 = \prod_{i=1}^{n} \prod_{j=1}^{m_i-1} \prod_{k=j+1}^{m_i} \left\{ \int_{y_{ij}} \int_{y_{ik}} f(y_{ij}, y_{ik}, r_{ij}, r_{ik}; \beta, \alpha, \gamma) \mathrm{d}y_{ij} \mathrm{d}y_{ik} \right\}^{(1-r_{ij})(1-r_{ik})},$$

where α is a vector of association parameters involved in the joint distribution of a pair of responses. This pairwise likelihood is contrasted inParzen et al. (2007) with the previously described independence likelihood of Troxel et al. (2003), again with analysis of data from the Six Cities study. The results show advantages from modelling also the dependence between responses.

Although the above pairwise likelihood may improve estimatmion efficiency compared to the independence likelihood, this comes at the cost of requiring correct specification of bivariate margins both of responses and of missingness indicators. In particular, the specification of the missing data mechanism even only for pairs is a critical aspect. Yi et al. (2009) showed how to overcome this. They assume that given any pair of responses (y_{ij}, y_{ik}) and covariates x_i, the missing data process does not comprise information on parameter β and α. With this assumption, inference can rely on the pairwise likelihood constructed from

the observed pairs of responses only,

$$\mathcal{L}_{\text{pair}}(\beta, \alpha; y) = \prod_{i=1}^{n} \prod_{j=1}^{m_i - 1} \prod_{k=j+1}^{m_i} f(y_{ij}, y_{ik}; \beta, \alpha)^{r_{ij} r_{ik}},$$

without requiring specification of the missing process distribution. Note that inferences from the observed pairwise likelihood are valid without assuming the missing at random mechanism.

4. Properties

4.1. Introduction

The motivation for the use of any version of composite likelihood is usually computational: to avoid computing, or in some cases modelling, the joint distribution of a possibly high-dimensional response vector. This is particularly true in applications of composite likelihood to mixed and random effects models, where the likelihood requires integration over the distribution of the random effects, as described in Section 3. Within this context, where composite likelihood is essentially a mis-specified model, interest has often focused on the relative efficiency of estimation from composite likelihood relative to the full likelihood. In §4.1 below we summarize the main results on efficiency of composite likelihood estimation.

Another motivation for the use of composite likelihood is a notion of robustness: in this case robustness under possible misspecification of the higher order dimensional distributions. For example, if pairwise likelihood is used for dependent binary data, it is not necessary to choose a model for joint probabilities of triples and quadruples, and to the extent that a number of possibilities for these could be consistent with the modelled joint probabilities of pairs, composite likelihood is by construction robust against these alternatives. This is a different notion of robustness than that in robust point estimation, and closer in spirit to the type of robustness achieved by generalized estimating equations. However, for many high-dimensional models, it is not clear what types of higher-order joint densities are indeed compatible with the modelled lower order marginal densities, so it is difficult to study the robustness issue in any generality. In §4.2 below we summarize what seems to be known about robustness in the literature.

Composite likelihood has also been used to construct joint distributions, in settings where there are not obvious high dimensional distributions at hand, for example in the use of frailty models in survival data (Fiocco et al., 2009).

Another feature of composite likelihood, noted for example in Liang and Yu (2003), is that the likelihood surface can be much smoother than the full joint likelihood, and thus easier to maximize. This is related to, but slightly different than, the ease of computation of composite likelihood, and closer in spirit to robustness of composite likelihood: by not specifying very high dimensional characteristics of the model, we are perhaps allowing a less complex structure on the parameter space as well. Renard et al. (2004) use the term computational robustness; in simulations they found that pairwise likelihood is more robust to convergence than their comparison method based on penalized quasi-likelihood. Computational aspects are considered in more detail in §5.

4.2 Relative efficiency

The seemingly high efficiency of composite likelihood methods in many applications has contributed to the increased interest in, and literature on, these methods. Three possible types of efficiency comparisons are: (i) asymptotic efficiency computed by an analytical calculation of $\mathrm{G}(\theta)$ and comparison with the Fisher information $\mathrm{I}(\theta)$, (ii) estimated asymptotic efficiency using simulation based estimates of $\mathrm{G}(\theta)$ and $\mathrm{I}(\theta)$, and (iii) empirical efficiency using simulation based estimates of $\mathrm{var}(\hat{\theta}_{\mathrm{CL}})$ and $\mathrm{var}(\hat{\theta})$. The first gives the clearest interpretation, although under the model assumption of the 'asymptotic ideal', whereas the third is closer to what may be achieved with finite sample sizes. A drawback of simulation based studies is that many aspects of the model must be specified in advance, so the relevance of the results to other, slightly different, models is not always clear. When θ is a vector an overall summary of the comparison of $\mathrm{G}(\theta)$ with $\mathrm{I}(\theta)$ can be computed using the ratio of the determinants, but more usually the diagonal components corresponding to particular parameters of interest are compared.

In exceptional cases pairwise likelihood estimators are fully efficient, and even identical to the maximum likelihood estimators. Mardia et al. (2007) show that composite conditional estimators are identical to maximum likelihood estimators in the multivariate normal distribution with arbitrary means and covariances, and Zi (2009) gives the same result for pairwise likelihood. Mardia et al. (2009) provide an explanation for this, by showing that composite conditional estimators are fully efficient in exponential families that have a certain closure property

under subsetting. Under further restrictions, composite marginal estimators are also fully efficient. An interesting special case is the equi-correlated multivariate normal distribution: a single observation vector has mean μ and covariance matrix $\sigma^2\{(1-\rho)I + \rho 11^{\mathrm{T}}\}$ where I is the identity matrix of dimension m and 1 is an m-vector of 1's. With μ and σ unknown, both pairwise maximum likelihood estimators and composite conditional maximum likelihood estimators are identical to the maximum likelihood estimator. If μ is fixed the same result holds, but if σ^2 is fixed, then $\hat{\rho}_{CL}$ is not fully efficient. Although this model is not covered by the closure result of Mardia et al. (2009), they adapt their discussion to show why the result continues to be true. It is possible that the method of Mardia et al. (2009) may also be adapted to explain the relatively high efficiency of composite likelihood methods in more complex applications, at least in some special cases. For example, while the bivariate von Mises distribution is not in the class of exponential families treated in Mardia et al. (2009), that paper shows that it is close to that class for most parameter values: this clarifies some results presented in Mardia et al. (2008) on this model.

The quadratic exponential distribution was proposed as a model for multivariate binary data in Cox (1972), and inference for this model was developed in Zhao and Prenctice (1990). As noted in Cox and Reid (2004) its likelihood function is equal to the pairwise likelihood function for binary data generated by a probit link. This provides a simple example where pairwise likelihood has full efficiency. Two-way contingency tables also have pairwise likelihood estimators equal to maximum likelihood estimators (Mardia et al., 2009).

Hjort and Varin (2008) also study in detail properties of composite conditional and composite marginal likelihoods in a simplified class of models. In their case they restrict attention to Markov chain models, and both theoretical analysis and detailed calculations provide strong evidence that composite marginal likelihood inference is both efficient and robust, and preferable to composite conditional likelihood inference. In their case the full likelihood is given by

$$\ell(\theta; y) = \sum_{a,b} y_{a,b} \log p_{a,b}(\theta),$$

where $y_{a,b}$ is the number of transitions from a to b, $p_{a,b}(\theta)$ is the stationary transition probability function, and a, b range over the number of states in the

Markov chain. This is a curved exponential family model, so the theory of Mardia et al. (2009) does not apply. The pairwise log-likelihood function is

$$c\ell(\theta; y) = \sum_{a,b} y_{a,b} \log p_{a,b}(\theta) + \sum_{a} y_{a+} \log p_a(\theta), \tag{4.1}$$

where $y_{a+} = \sum_b y_{a,b}$ and $p_a(\theta)$ is the equilibrium probability that the chain is in state a. Equation (4.1) is interpreted in Hjort and Varin (2008) as a penalized log-likelihood, with a penalty function that is targetted on matching the equilibrium distribution. This provides a different explanation of the efficiency and robustness of pairwise likelihood inference.

The papers by Mardia et al. (2009) and Hjort and Varin (2008) seek to establish some general results about composite likelihood, albeit in relatively specialized settings. The rest of the literature that we have reviewed on composite likelihood is typically concerned with comparisons in particular models motivated by applications. In the paragraphs below we highlight recent work on efficiency that seems to us to be particularly useful.

In models for clustered data, where observations $y_{ij}, j = 1, \ldots, n_i$ within the ith cluster are correlated, asymptotic relative efficiency can often be assessed by obtaining analytical expressions for $\mathrm{G}(\theta)$ and $\mathrm{J}(\theta)$. Within this context, extensive studies of asymptotic relative efficiency are available, and there is also a literature on the choice of weights, usually related to cluster size, for achieving optimal efficiency. For pairwise likelihood, Joe and Lee (2009) investigate the choice of weights for clustered data in detail, and show that the best choice of weights depends on the strength of the dependence within clusters. The models investigated analytically are the multivariate normal, where direct comparisons to the maximum likelihood estimator can be made, and the multivariate binary, created by dichotomizing multivariate normal observations. The weights $1/(n_i - 1)$, recommended in Kuk and Nott (2000), LeCessie and van Houwelingen (1994) and Zhao and Joe (2005), are suitable for the limiting case of independence within clusters, but the weights $1/\{n_i(n_i - 1)\}$ are optimal for very strong dependence. A compromise suggested in Joe and Lee (2009) is $1/[(n_i - 1)\{1 + 0.5(n_i - 1)\}]$, which works well for a range of parameter values and models. Most applications to date however have used the simpler weights $1/(n_i - 1)$. Joe and Lee (2009) also show that the best choice of weights depends on which parameter is to be

estimated, providing further detail on the earlier results of Kuk and Nott (2000) and others that unweighted pairwise likelihood can be preferable for inference about the association parameters, whereas weighting improves on the estimation of the parameters in the mean.

When modelling clustered data the parameters in the one-dimensional margins are usually regression coefficients and variances, and the association parameters only appear in the two-dimensional margins. This suggests using separate approaches for inference on these two sets of parameters, and several suggestions along these lines have appeared in various contexts. Zhao and Joe (2005) explores using the independence likelihood for the marginal parameters and pairwise likelihood for the association parameters, although in most cases the full pairwise likelihood method turns out to be more efficient. Kuk (2007) suggests a quite promising hybrid method that uses optimal score functions for the marginal parameters, and pairwise likelihood for the association parameters, regarded as nuisance parameters. This hybrid method is shown to be related to, but better than, alternating logistic regression (Carey et al., 2003), and is illustrated on ordinal count data, as well as negative binomial count data; the latter application is also treated in Henderson and Shimakura (2003). For further discussion of this approach see Varin (2008).

The same data structure $y_{ij}, j = 1, \ldots, n_i$ may arise as longitudinal data, in which case serial dependence of the observations is usually part of the model. In this case the inferential problem is more similar to time series analysis, with the difference that longitudinal data is typically n independent short time series, rather than a single long time series. Asymptotic efficiency for longitudinal data is studied analytically in Joe and Lee (2009). The weighting schemes typically proposed in time series analysis downweight observations that are far apart in time, and Joe and Lee (2009) find that choosing weights so that the pairwise likelihood is constructed only from adjacent pairs is preferable to the full pairwise likelihood involving all possible pairs in the sequence.

For time series models both marginal and conditional composite likelihoods have been proposed, with a possible weighting scheme chosen to downweight observations that are far apart in time. Explicit comparison of the simulation variance for composite marginal likelihood of different orders is illustrated in

Varin and Vidoni (2006), where again it is shown that including too distant observations in the composite marginal formulation can lead to a loss of efficiency. Simulations of non-stationary time series are presented in a particular model for ecology in Lele (2006), where the pairwise likelihood is shown to be more efficient than the independence likelihood.

There are a number of investigations of asymptotic relative efficiency for clustered and longitudinal data that rely on simulations, rather than analytical calculations of the asymptotic variances. Such studies can consider more complex models for marginal and association parameters, but it is difficult to gain an overall picture of the results. Examples of simulation studies that show high relative efficiency for pairwise likelihood in binary data include Renard et al. (2002), Renard et al. (2004), Fieuws and Verbeke (2006) and Feddag and Bacci (2009). The last paper considers a multidimensional Rasch model proposed for longitudinal studies in item response theory. In all these papers pairwise likelihood has good simulation-based efficiency relative to inference based on the full likelihood function, or in some cases approximations to it, but there is likely to be a statistical 'file-drawer' problem, in that situations for which composite likelihood performs poorly are perhaps unlikely to be published, at least until a method can be developed that seems to work well.

Sparse clustered binary data may arise in finely stratified studies, and two versions of composite likelihood are suggested in Hanfelt (2004) and in Wang and Williamson (2005), using Liang (1987)'s composite conditional likelihood (2.1). Simulations in Wang and Williamson (2005) compare composite likelihood estimators of marginal and association parameters to estimators derived from an estimating equations approach. The two methods have comparable efficiency; the authors note that the composite likelihood equations for the association parameters very often have multiple roots, which makes numerical work based on composite likelihood rather difficult in this setting. It would be useful to have an explanation for this, as most authors who comment on numerical aspects of composite likelihood estimation report that composite likelihood functions are well behaved and relatively easy to maximize.

In the approach of Hanfelt (2004) to sparse binary data there are an increasing number of nuisance parameters, and an adaptation of the estimating equation

for the association parameter derived from the composite conditional likelihood is needed. Hjort and Varin (2008) note that in the Neyman-Scott model of several normal groups with common mean but separate variances the pairwise likelihood based on differences gives consistent inference for the common variance, even as the number of groups increases. Composite likelihood inference with very large numbers of nuisance parameters is also considered in Engle et al. (2009) and Pakel et al. (2009).

Heagerty and Lele (1998) proposed the use of pairwise likelihood for spatial binary data generated through a multivariate probit model. Limited simulations there suggest that the pairwise likelihood estimator is efficient for estimating parameters in the mean, but somewhat less efficient in estimation of variance parameters. See Bhat, Sener and Eluru (2010) for an extension to regression analysis of spatially correlated ordinal responses. A general approach to spatial generalized linear mixed models is presented in Varin et al. (2005), and simulations are presented showing that pairwise likelihood inference for both mean and variance parameters in a Poisson random effects model does better than inference based on an high-dimensional Laplace approximation of the full likelihood. Several computational issues arise in fitting pairwise and full log-likelihoods in spatial generalized linear mixed models and the authors describe an EM-type algorithm; see Section 5.

Caragea and Smith (2007) used analytical calculations of asymptotic efficiency, as well as simulations, to choose among three possible composite likelihood approaches for Gaussian random fields, as described in §3.1 above. Their conclusions were broadly that a method that uses groups of nearby observations ("small blocks") is more efficient than a version closer to independence likelihood, and that for estimating regression parameters a hybrid method was slightly better. Simulations of spatial point processes presented in Guan (2006) show that adaptive estimation the weights assigned to the likelihood of each pair can be effective.

While most simulation studies show that some version of composite likelihood has high efficiency, a warning is presented in Oliveira (2004), where a new spatial model for rainfall is proposed. This model is based on a mixture of discrete and continuous spatial processes, to represent both the occurrence and amount

of rainfall, and it is noted that simulations indicate very poor performance of pairwise likelihood for estimating parameters in the spatial correlation functions.

Simulation efficiency of pairwise likelihood in general state space models is considered in Varin and Vidoni (2009) and Joe and Lee (2009), and Andrieu et al. (2005) develop a version of composite likelihood adapted to sequential Monte Carlo inference.

4.3 Robustness

Many authors refer to composite likelihood inference as robust, because composite likelihood requires only model assumptions on the lower dimensional conditional or marginal densities, and not detailed specification of the full joint distribution. Thus if there are several joint distributions with the same lower dimensional marginal or conditional distributions, the inference will be the same for all members of that family.

A small number of papers investigate robustness in more detail, usually through simulations from a mis-specified model. For example, Lele and Taper (2002) investigated the behaviour of $\hat{\theta}_{\mathrm{CL}}$ from the likelihood based on pairwise differences, (2.3), in their case a one-way random effects model, first assuming normality for the distribution of the random effects, and then simulating the random effects under non-normal distributions. They concluded that composite likelihood estimators and restricted maximum likelihood (REML) estimators of variance components behaved similarly under model misspecification. The REML likelihood is the likelihood function for the marginal distribution of the residuals, which for normal theory models is the same as the likelihood based on pairwise differences, so may be very close to (2.3) in the models that Lele and Taper (2002) studied. Wang and Williamson (2005) present simulations of sparse clustered binary data, under a model for which the correlation structure is misspecified, and their results also indicate that composite likelihood methods continue to have high efficiency.

In longitudinal data analysis it is not unusual to have missing observations, and modelling this can be important for valid inferences. This is considered in detail in Parzen et al. (2007), and again in Yi et al. (2009), as discussed in Section 3.4 above. The fact that some versions of composite likelihood are indeed robust to the specification of the missing data mechanism is another very attractive

feature of composite likelihood.

The inverse of the Godambe information, $\mathrm{G}(\theta)$, is often called the robust variance estimate, as it is computed under the assumption that the model is misspecified, and composite likelihood models are by definition misspecified. However the use of $\mathrm{G}^{-1}(\theta)$ as a variance estimator does not guarantee, for example, that the composite likelihood estimator will have high efficiency under a range of models consistent with the composite likelihood; these need to be investigated on their own merits.

Liang and Qin (2000) use a specialized version of composite conditional likelihood for a number of non-standard regression models, where modelling of the distribution of the explanatory variables may be needed. Their simulations address robustness to misspecification of this aspect of the modelling, noting that the composite maximum likelihood estimator continues to have small bias, but somewhat larger variance, under this misspecification.

Finally, Kent (1982) called the log-likelihood ratio statistic W robust if its asymptotic distribution was χ^2_p, rather than the more complex form given after (2.5), and discussed a special class of exponential family models that guaranteed this result by showing that the score equations were information unbiased. This line of argument is developed further in Mardia et al. (2009).

4.4 Identifiability

It is not clear whether or not composite likelihood methods give meaningful results if there is no joint distribution compatible with the component densities used to construct the composite likelihood. In the case that the composite likelihood is constructed from conditional distributions, the Hammersley-Clifford theorem specifies when there is a genuine joint distribution consistent with these conditional distributions, and this was used in Besag (1974) in his development of pseudo-likelihood for spatial data. This issue is pursued in Wang and Ip (2008), where the key notion of interactions is defined, and their role in ensuring the compatibility of conditional and joint distributions is emphasized; see also Arnold et al. (2001).

There is not an analogous result for composite marginal likelihood, although there is likely to be a connection to the theory of construction of joint distributions using copulas. Several papers on the use of composite marginal likelihood

use a copula construction (Bhat, Sener and Eluru, 2010; Tibaldi et al., 2004; Andersen, 2004) but many applications of composite marginal likelihood do not. For example, the development of composite likelihood for spatial extremes described in §3.2 uses pairwise marginals as a proxy for a genuine joint distribution.

However we may consider the composite Kullback-Leibler divergence,

$$\mathrm{CKL}(g, f; \theta) = \sum_{k=1}^{K} w_k \mathrm{E}_g \left\{ \log g(y \in \mathcal{A}_k) - \log f(y \in \mathcal{A}_k; \theta) \right\},$$

consisting of the linear combination of the Kullback-Leibler divergences for each term of the composite likelihood. Under some regularity conditions the maximum composite likelihood estimator $\hat{\theta}_{\mathrm{CL}}$ will be consistent for the parameter value minimizing CKL, and inference about this pseudo-parameter may be of useful for particular applications. We could also view the estimating equation from the composite likelihood as a reasonable specification of knowledge about parameters of lower dimensional marginal distributions, in the spirit of generalized estimating equations; see Varin (2008). This might be especially true for estimating parameters in the mean function.

Joe and Lee (2009) notes in passing that unless the component likelihoods in a composite likelihood construction are "rich enough to identify the parameter", the composite likelihood estimator will not be consistent. Presumably if a full joint distribution exists in which the parameters of the components are (subvectors of the) parameters of the full joint distribution this guarantees identifiability. However it seems possible that the parameters of the component densities could be identifiable under weaker conditions.

In the approach outlined in §3.3.3, each component marginal density has its own parameter θ_{rs}, say, and the estimator used for the notional parameter θ of interest is a linear combination of the pairwise estimators $\hat{\theta}_{rs}$. The connection of this to identifiability of joint densities is not clear.

5. Computational aspects

5.1. Standard errors

Standard errors and confidence interval computation require the estimation of the Godambe matrix and its components. Again, it is useful to distinguish between the case of n large with m fixed, and vice-versa. The first case is simpler with easily computed sample estimates of the sensitivity and variability matrices.

The sample estimate of the sensitivity matrix is given by

$$\hat{\mathrm{H}}(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \nabla u(\hat{\theta}_{\mathrm{CL}}; y_i)$$

where $u(\theta; y_i) = \nabla c\ell(\theta; y_i)$. Computation of Hessians can be avoided by exploiting the second Bartlett identity, which remains valid for each individual likelihood term forming the composite likelihood. This yields the alternative estimate

$$\hat{\mathrm{H}}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} u(\hat{\theta}_{\mathrm{CL}}; y_{ij}) u(\hat{\theta}_{\mathrm{CL}}; y_{ij})^{\mathrm{T}}.$$

The sample estimate of the variability matrix is expressed by the outer product of the composite scores computed at $\hat{\theta}_{\mathrm{CL}}$,

$$\hat{\mathrm{J}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} u(\hat{\theta}_{\mathrm{CL}}; y_i) u(\hat{\theta}_{\mathrm{CL}}; y_i)^{\mathrm{T}}.$$

The above empirical estimates of H and J may be imprecise when n is not sufficiently large compared to the dimension of θ. This is well known in the longitudinal literature where resampling methods, such as jackknife or bootstrap, are often used to obtain more robust estimates of the covariance matrix of $\hat{\theta}_{\mathrm{CL}}$; see for example Lipsitz et al. (1994). The jackknife covariance matrix is given by

$$\mathrm{var}_{\mathrm{jack}}(\hat{\theta}_{\mathrm{CL}}) = \frac{n-1}{n}\sum_{i=1}^{n} (\hat{\theta}_{\mathrm{CL}}^{(-i)} - \hat{\theta}_{\mathrm{CL}})(\hat{\theta}_{\mathrm{CL}}^{(-i)} - \hat{\theta}_{\mathrm{CL}})^{\mathrm{T}},$$

where $\hat{\theta}_{\mathrm{CL}}^{(-i)}$ is the composite likelihood estimator of θ with y_i deleted. Zhao and Joe (2005) use $\mathrm{var}_{\mathrm{jack}}$ for estimation of the standard errors of maximum pairwise likelihood estimators with clustered data. A further possible advantage of the jackknife method is the possibility to obtain an approximate bias correction of $\hat{\theta}_{\mathrm{CL}}$. In certain applications the computation of the set of $\hat{\theta}_{\mathrm{CL}}^{(-i)}$ can be excessively expensive, and then it may be convenient to consider a first order approximation where $\hat{\theta}_{\mathrm{CL}}^{(-i)}$ is approximated with a single step of the Newton-Raphson algorithm.

More difficult is the case of m large when n is fixed, with the extreme situation of $n = 1$ when a single time-series or spatial process is observed. While the sample estimate of the sensitivity matrix H has the usual form, difficulties arise for the variability matrix J. A sample estimate of the latter is possible only

if the data can be grouped into pseudo-independent replicates. Considering a temporal or spatial process with good mixing properties, a sample estimate of J can be obtained by splitting the region under study into subregions treated as approximately independent:

$$\hat{\mathrm{J}}_{\mathrm{ws}}(\theta) = \frac{1}{k}\sum_{i=1}^{k} |\mathcal{S}_i|\, u(\hat{\theta}_{\mathrm{CL}}; y \in \mathcal{S}_i) u(\hat{\theta}_{\mathrm{CL}}; y \in \mathcal{S}_i)^{\mathrm{T}},$$

where $\mathcal{S}_1, \ldots, \mathcal{S}_k$ are k possibly overlapping subregions and $|\mathcal{A}|$ denotes the dimension of set $\mathcal{A}$. Heagerty and Lele (1998) term this method window subsampling and use it for pairwise likelihood inference with spatial binary data. For more details and guidance on the choice of the subregions, we refer to Lumley and Heagerty (1999).

When conditions for ensuring the validity of window subsamping or other empirical estimates are not satisfied, estimation of J must be done under model assumptions. In certain contexts, it may be possible to compute J explicitly. For example, in the case of the pairwise likelihood, model-based estimation of J typically requires computation of four-dimensional expectations. When it is easy to simulate data from the complete model Monte Carlo simulations can be used either for estimating the J matrix with

$$\hat{\mathrm{J}}_{\mathrm{mc}}(\theta) = \frac{1}{B}\sum_{i=1}^{B} u(\hat{\theta}_{\mathrm{CL}}; y^{(b)}) u(\hat{\theta}_{\mathrm{CL}}; y^{(b)})^{\mathrm{T}},$$

where $y^{(1)}, \ldots, y^{(B)}$ are independent draws from the fitted model, or for direct estimation of the covariance matrix of $\hat{\theta}_{\mathrm{CL}}$ from repeated fitting of simulated data.

5.2 Composite likelihood Expectation-Maximization algorithm

The expectation-maximization (Dempster et al., 1977; EM) algorithm and its variants are popular methods to obtain maximum likelihood estimates in a number of situations. Examples include missing data, censored data, latent variables, finite mixture models, and hidden Markov models. See McLachlan and Krishnan (2008) for a book length exposition.

The EM algorithm can be straightforwardly extended to maximization of composite likelihoods. This can be useful for models where the expectation step involves high-dimensional integration, thus making impractical the use of a stan-

dard EM algorithm. The first example of the use of a composite EM algorithm seems to be the pairwise EM algorithm proposed by Liang and Yu (2003) in network tomography, see also Castro et al. (2004). Varin et al. (2005) consider an approximate version of the same algorithm for inference in spatial generalized linear mixed models discussed in §3.4. Gao and Song (2010) discusses properties of a general composite marginal likelihood EM algorithm and gives an illustration of the pairwise version for multivariate hidden Markov models applied to time-course microarray data.

Here we briefly summarize only the pairwise EM algorithm. Let $x_1, \ldots, x_m$ be the complete data and $y_1, \ldots, y_m$ the observed data. Denote by $\theta^{(0)}$ a starting value for θ. Given $\theta^{(k)}$, the pairwise EM algorithm iterate at step k, the next iterate $\theta^{(k+1)}$ is the value such that

$$Q(\theta^{(k+1)}|\theta^{(k)}) \geq Q(\theta|\theta^{(k)}), \quad \text{for any } \theta \in \Theta,$$

where $Q(\theta|\theta^{(k)})$ is the sum of bivariate conditional probabilities

$$Q(\theta|\theta^{(k)}) = \sum_{r=1}^{m-1} \sum_{s=r+1}^{m} \mathrm{E}\{\log f(x_r, x_s; \theta)|y_r, y_s; \theta^{(k-1)}\}.$$

As shown in detail by Gao and Song (2010), it is easy to prove that this algorithm shares the three key properties of standard EM algorithms, namely (i) the ascent property

$$\mathcal{L}_{\mathrm{pair}}(\theta^{(k+1)}; y) \geq \mathcal{L}_{\mathrm{pair}}(\theta^{(k)}; y), \; k = 1, 2, \ldots$$

(ii) convergence to a stationary point of the objective function and (iii) convergence rate depending on the curvature of the objective function.

5.3 Low-dimensional integration versus high-dimensional integration

In many applications, the motivation for composite likelihood inference is to substitute awkward high-dimensional integration involved in full likelihoods with low-dimensional integrals. The latter can often be computed by using accurate deterministic quadrature rules. For example, Bellio and Varin (2005) approximate integrals involved in logistic regression models with random effects using normal scale mixtures and bivariate quadrature rules.

In contrast, high-dimensional integrals typically require Monte Carlo simulation methods with various potential difficulties. First, the computational time

may be too large for practical purposes. Secondly, the simulation error may be substantial and difficult to evaluate, making the optimization of the approximated likelihood troublesome. A third reason for concern regards reproducibility of results, especially for a non-technical audience.

A possible advantage of simulated maximum likelihood versus composite likelihood methods is the possibility to base inference on the standard asymptotic results, without the need to compute the more difficult Godambe information or to modify the chi-squared distribution of the likelihood ratio test. However, some authors suggest the use of the Godambe information also for simulated maximum likelihood to take into account the simulation error due to the use of a finite number of draws; see for example McFadden and Train (2000). Thus, the potential simplicity of maximum likelihood inference is lost when using simulations to approximate the likelihood. For a comparison between simulated maximum likelihood based on quasi-Monte Carlo rules and pairwise likelihood for ordinal probit models see Bhat, Varin and Ferdous (2010).

5.4 Combinatorial difficulties

As observed by one referee, another computational motivation for preferring the composite likelihood method is the combinatorial difficulty associated with some likelihood-type analyses based on the complete data. Examples of this include computation of the partial likelihood (Cox, 1975) for the proportional hazards model when the number of events is large, and computation of the conditional likelihood for case-control studies with a large number of cases. Other combinatorial difficulties arise when the computation of the joint distribution of the data requires conditioning on the order statistics, thus involving $n!$ permutations, where n is the sample size (Kalbfleisch, 1978). While the difficulty of computing high-dimensional integrals leads naturally to composite marginal likelihoods, avoiding these combinatorial difficulties leads to the use of composite conditional likelihoods, as in Liang (1987) and Liang and Qin (2000).

6. Conclusions

6.1 Relations with other approaches

In many applications of marginal or conditional composite likelihood, the approach of generalized estimating equations originated in Liang and Zeger (1986) is a natural alternative. This approach defines an estimating equation through

a model for the mean, and accommodates correlation among observations, and non-homogeneous variances, by weighting the estimating equation appropriately. Liang and Zeger (1986) showed that as long as the estimating equation for the mean is correctly specified, the resulting estimator will be consistent, and suggested using a working covariance matrix to this end. Many refinements have since been suggested, and the method is very convenient for semi-parametric modelling of complex data. A possible drawback of the method is that there is no objective function, which can be useful for comparing multiple roots of the estimating equation. For clustered binary data Molenberghs and Verbeke (2005, Ch. 9) give a detailed comparison of the estimating equations from pairwise likelihood, with weight $1/(n_i-1)$ for clusters of size n_i, to two versions of generalized estimating equations, GEE1 and GEE2, where the latter requires modelling of the first four moments of the data; they argue that pairwise likelihood is a compromise between the two, with computational complexity similar to GEE1, but efficiency closer to GEE2.

Many of the more complex applications of composite likelihood, particularly in longitudinal or clustered data, provide comparisons using simulation studies to some type of estimating equation, usually a generalized estimating equation; see for example Geys et al. (2001), Hanfelt (2004), and Zhao and Ma (2009). Hybrid methods that combine features of composite likelihood with generalized estimating equations, as in Kuk (2007), seem quite promising. In the other direction, Oman et al. (2007) used a generalized estimating equation approach to simplify the computation of pairwise likelihood.

In its most general form composite likelihood encompasses many types of likelihood-like functions suggested in the statistical literature, including partial likelihood for censored survival data and its many extensions, as well as nonparametric likelihoods for counting processes. For example, Wellner and Zhang (2007) and Wellner and Zhang (2000) propose non-parametric and semi-parametric estimators for panel count data using an independence likelihood, and Andersen (2004) uses pairwise likelihood in the proportional hazards model. Other extensions to likelihood composition include the weighted likelihood of Zidek and Hu (1997) and the partitioning of likelihood for maximization by parts in Kalbfleisch et al. (2005).

6.2 Some challenges

Using the most general definition of composite likelihood, it may be difficult to derive very many specific properties beyond perhaps consistency of the point estimator, as the range of models is simply too broad (Lindsay et al., 2010). In this subsection we consider some ideas that seem promising for further research in the theory and application of composite marginal likelihood and composite conditional likelihood.

Some theoretical issues are mentioned in passing in Section 4, and others are addressed in papers in this special issue. One important question is the relation between the lower dimensional marginal or conditional distributions used to construct composite likelihood and the underlying joint distribution. For conditional densities, the Hammersley-Clifford theorem (Besag, 1974) provides some guarantees about the existence of a full joint distribution, even if this distribution is not computable. There is not a similar result for the compounding of low-dimensional joint densities, beyond univariate marginal densities, where the independence likelihood corresponds trivially to a full joint distribution. While some authors state that the existence of such a full joint density is needed for sensible inference from composite likelihood, others argue that the parameters in the low-dimensional margins may be interpretable anyway. Varin and Vidoni (2005) argue this from the point of view of Kullback-Leibler divergence, whereas Faes et al. (2008) devise a method of relating individual parameters in component densities to a common parameter of interest. The theory of construction of multivariate distributions using copulas (Joe, 1997) may be useful for exploring these ideas further.

In work as yet unpublished, presented at a session on composite likelihood at the Joint Statistical Meetings in Washington, 2009, G. Y.Yi gave several illustrative examples that raise questions about modelling based on composite likelihood, and questions about the comparison of composite likelihood to full likelihood methods. A modelling example assumes the following pairwise distribution for binary data:

$$f(y_j, y_k; \beta) \propto \frac{\exp(\beta y_j + \beta y_k + \beta_{jk} y_j y_k)}{1 + \exp(\beta y_j + \beta y_k + \beta_{jk} y_j y_k)}.$$

If the binary vector has, for example, length 3, then pairwise likelihood will lead to

different estimates of β_{12}, β_{13} and β_{23}, yet the strong form of the marginal model assumed constrains the full joint density to have $\beta_{12} = \beta_{13} = \beta_{23}$. Similarly, it is not difficult to construct examples where pairwise conditional densities are not compatible with any joint density. A somewhat different point arises in the symmetric normal example of pairwise likelihood, where Y follows a normal distribution with covariance matrix $\Sigma = (1-\rho)I + \rho 11^{\mathrm{T}}$, where I is the identity and 1 is an m-vector of 1's. In the full joint distribution for a vector Y of length m, Σ is only positive definite for $-1/(m-1) < \rho < 1$, whereas the pairwise composite likelihood requires for only $-1/2 < \rho < 1$; thus pairwise likelihood would be expected to be unsuitable for $\rho < 0$, a point confirmed in Mardia et al. (2007).

Several recent papers, including some in this special issue, (Okabayashi et al., 2010; Davis and Yau, 2010) address the question of how many terms to include in a composite likelihood, following along the lines of Hjort and Varin (2008). The answer is likely to depend fairly strongly on the application, but from a theoretical point of view it is possible to study in a more general way the information accumulation provided, or not provided, by adding additional terms. Lindsay et al. (2010) discusses in detail a number of aspects of the design of composite likelihoods, with particular emphasis on the components of the composite likelihood score function. This design issue seems especially relevant for the context of spatial and time series applications, where replication is obtained from observations sufficiently distant in time or space. A special aspect of this is the weighting of subsets of observations, which has been studied in some detail particularly for clustered and/or longitudinal data. A more general understanding of the asymptotic efficiency of composite marginal likelihood would be welcome. Some progress is made in Mardia et al. (2009), but the conditions there seem rather strong, and apply more easily to composite conditional likelihood.

One motivation for the use of composite marginal likelihood is that it is easier to model the univariate, bivariate or trivariate dependence than the full joint dependence. The claim is often made that these models are more robust than the full joint models, but it seems difficult to make this precise. This is partly because it is not always clear how much the specification of, for example, the bivariate marginals constrains the full joint distribution. A related question,

suggested by Bruce Lindsay in personal communication, is whether there may be a higher dimensional model in which J = H so that asymptotic efficiency is achieved at this particular model.

Most of the study of inference from composite likelihood has focussed on point estimation and estimation of the asymptotic variance of the composite maximum likelihood estimator. This follows fairly directly from the theory of estimating equations, at least if the dimension of the vector is fixed while the sample size increases. In ordinary likelihood theory inference based on the log-likelihood ratio statistic is often preferred, but its asymptotic distribution for the composite likelihood analogue is difficult to work with. Promising alternatives include Satterthwaite (1946) type corrections mentioned in Section 2.3 and the adjustment developed in Pace et al. (2010). Another aspect of the asymptotic theory that needs further study is the case of increasing dimension q with fixed or slowly increasing sample size: this is particularly important for genetics applications. There does not seem to be a rigorous proof that the composite maximum likelihood estimator is, or is not, consistent, under various conditions on q and n: some heuristics were sketched in Cox and Reid (2004).

One very important methodological development that is touched on in several papers in this issue, as well as in other literature, is the use of composite likelihood methods with missing data and with mis-measured data. This seems a particularly important applied issue, as composite likelihood methods are so often used with longitudinal data in medical applications, where missed visits may be nearly unavoidable, and lead to gaps in the series available for each subject in the study. In some models the mechanism for missingness does not need to be modelled in the composite likelihood approach (Yi et al., 2009; Yi and He, 2010). Molenberghs et al. (2010) use ideas of double robustness from the theory of estimating equations to adjust for missingness under the assumption of missing at random; and Gao and Song (2009) develop EM-type algorithms that however require missingness to be completely at random.

Another important methodological aspect is computation, particularly estimation of $\mathrm{J}(\theta)$ when there is not internal replication. Several strategies are suggested in Section 5.1, but a systematic comparison in a broad range of models could be worthwhile.

Composite likelihood is being developed, often independently by the researchers in different fields, for use in a wide variety of application areas well beyond the spatial or longitudinal data for which it was originally developed: computer experiments, network analysis, population genetics, and investment portfolios, to name only a few of the recent applications. It is a natural way to simplify the modelling of complex systems, and seems likely to become well established as an alternative approach to full likelihood inference.

Acknowledgments

We are grateful to Don Fraser and Grace Yun Yi for helpful discussions, and to a referee and associate editor for comments on an earlier draft. The research was partially supported by MIUR Italy, NSERC Canada and ESRC UK. The 2008 workshop at Warwick was supported by UK research councils ESRC, via the National Centre for Research Methods, and EPSRC, via the Centre for Research in Statistical Methodology.

References

Andersen, E. (2004), 'Composite likelihood and two-stage estimation in family studies', *Biostatistics* **5**(1), 15–30.

Andrieu, C., Doucet, A. and Tadic, V. (2005), On-line parameter estimation in general state-space models, *in* '44th Conference on Decision and Control', pp. 332–337.

Apanasovich, T., Ruppert, D., Lupton, J., Popovic, N. and Carroll, R. (2008), 'Aberrant crypt foci and semiparametric modeling of correlated binary data', *Biometrics* **64**(2), 490–500.

Arnold, B., Castillo, E. and Sarabia, J. (2001), 'Conditionally specified distributions: An introduction', *Statistical Science* **16**(3), 249–274.

Barry, S. and Bowman, A. (2008), 'Linear mixed models for longitudinal shape data with applications to facial modelling', *Biostatistics* **9**, 555–565.

Bellio, R. and Varin, C. (2005), 'A pairwise likelihood approach to generalized linear models with crossed random effects', *Statistical Modelling* **5**, 217–227.

Besag, J. (1974), 'Spatial interaction and the statistical analysis of lattice

systems', *Journal of the Royal Statistical Society. Series B (Methodological)* **36**(2), 192–236.

Bhat, C. R., Sener, P. N. and Eluru, N. (2010), 'A flexible spatially dependent discrete choice model: Formulation and application to teenagers' weekday recreational activity participation', *Transportation Research Part B* **44**(8–9), 903–921.

Bhat, C. R., Varin, C. and Ferdous, N. (2010), A comparison of the maximum simulated likelihood and composite marginal likelihood estimation approaches in the context of the multivariate ordered response model system. *Advances in Econometrics: Maximum Simulated Likelihood Methods and Applications*, Vol. 26, edited by W.H. Greene, Emerald Group Publishing Limited, to appear.

Caragea, P. and Smith, R. (2007), 'Asymptotic properties of computationally efficient alternative estimators for a class of multivariate normal models', *Journal of Multivariate Analysis* **98**, 1417–1440.

Caragea, P. and Smith, R. L. (2006), Approximate likelihoods for spatial processes. Preprint.

Carey, V., Zeger, S. and Diggle, P. (2003), 'Modelling multivariate binary data with alternating logistic regressions', *Biometrika* **80**(3), 517–526.

Castro, R., Coates, M., Liang, G., Nowak, R. and Yu, B. (2004), 'Network tomography: recent developments', *Statistical Science* **19**, 499–517.

Chandler, R. E. and Bate, S. (2007), 'Inference for clustered data using the independence log-likelihood', *Biometrika* **94**(1), 167–183.

Claeskens, G. and Hjort, N. (2008), *Model Selection and Model Averaging*, Cambridge University Press, Cambridge.

Cox, D. (1975), 'Partial likelihood', *Biometrika* **62**(2), 269–276.

Cox, D. R. (1972), 'The analysis of multivariate binary data', *Applied Statistics* **21**, 113–120.

Cox, D. and Reid, N. (2004), 'A note on pseudolikelihood constructed from marginal densities', *Biometrika* **91**(3), 729–737.

Cressie, N. (1993), *Statistics for Spatial Data*, Wiley, New York.

Curriero, F. and Lele, S. (1999), 'A composite likelihood approach to semivariogram estimation', *Journal of Agricultural, Biological, and Environmental Statistics* **4**(1), 9–28.

Davis, R. A. and Yau, C. Y. (2010), 'Comments on pairwise likelihood in time series models', *Statistica Sinica* .

Davison, A. and Gholamrezaee, M. (2009), Geostatistics of extremes, Technical report, EPFL. submitted.

Dempster, A., Laird, N. and Rubin, D. (1977), 'Maximum likelihood from incomplete data via the em algorithm', *Journal of the Royal Statistical Society. Series B (Methodological)* **39**, 1–22.

Diggle, P. and Ribeiro, P. (2007), *Model-based Geostatistics*, Springer, New York.

Engle, R. F., Shephard, N. and Sheppard, K. (2009), Fitting and testing vast dimensional time-varying covariance models. Preprint.

Faes, C., Aerts, M., Molenberghs, G., Geys, H., Teuns, G. and Bijnens, L. (2008), 'A high-dimensional joint model for longitudinal outcomes of different nature', *Statistics in medicine* **27**(22), 4408–4427.

Feddag, M.-L. and Bacci, S. (2009), 'Pairwise likelihood for the longitudinal mixed Rasch model', *Computational Statistics & Data Analysis* **53**(4), 1027–1037.

Fieuws, S. and Verbeke, G. (2006), 'Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles', *Biometrics* **62**(2), 424–431.

Fieuws, S., Verbeke, G., Boen, G. and Delecluse, C. (2006), 'High dimensional multivariate mixed models for binary questionnaire data', *Applied Statistics* **55**, 449–460.

Fieuws, S., Verbeke, G., Maes, B. and Vanrenterghem, Y. (2007), 'Predicting renal graft failure using multivariate longitudinal profiles', *Biostatistics* **9**(3), 419–431.

Fieuws, S., Verbeke, G. and Molenberghs, G. (2007), 'Random-effects models for multivariate repeated measures', *Statistical Methods in Medical Research* **16**(5), 387–397.

Fiocco, M., Putter, H. and van Houwelingen, J. C. (2009), 'A new serially correlated gamma-frailty process for longitudinal count data', *Biostatistics* **10**(2), 245–257.

Fujii, Y. and Yanagimoto, T. (2005), 'Pairwise conditional score functions: a generalization of the Mantel–Haenszel estimator', *Journal of Statistical Planning and Inference* **128**, 1–12.

Gao, X. and Song, P. X.-K. (2009), Composite likelihood Bayesian information criteria for model selection in high dimensional data. Preprint.

Gao, X. and Song, P. X.-K. (2010), 'Composite likelihood EM algorithm in incomplete high-dimensional correlated data analysis', *Statistica Sinica* .

Geys, H., Molenberghs, G. and Ryan, L. (1999), 'Pseudolikelihood modeling of multivariate outcomes in developmental toxicology', *Journal of the American Statistical Association* **94**(447), 734–745.

Geys, H., Regan, M., Catalano, P. and Molenberghs, G. (2001), 'Two latent variable risk assessment approaches for mixed continuous and discrete outcomes from developmental toxicity data', *Journal of Agricultural, Biological, and Environmental Statistics* **6**(3), 340–355.

Glasbey, C. (2001), 'Non-linear autoregressive time series with multivariate Gaussian mixtures as marginal distributions', *Applied Statistics* **50**(2), 143–154.

Godambe, V. (1960), 'An optimum property of regular maximum likelihood estimation', *The Annals of Mathematical Statistics* **31**(4), 1208–1211.

Guan, Y. (2006), 'A composite likelihood approach in fitting spatial point process models', *Journal of the American Statistical Association* **101**(476), 1502–1512.

Hanfelt, J. (2004), 'Composite conditional likelihood for sparse clustered data', *Journal of the Royal Statistical Society. Series B (Methodological)* **66**(1), 259–273.

Heagerty, P. and Lele, S. (1998), 'A composite likelihood approach to binary spatial data', *Journal of the American Statistical Association* **93**(443), 1099–1111.

Henderson, R. and Shimakura, S. (2003), 'A serially correlated gamma frailty model for longitudinal count data', *Biometrika* **90**(2), 335–366.

Hjort, N. and Omre, H. (1994), 'Topics in spatial statistics (with discussion, comments and rejoinder)', *Scandinavian Journal of Statistics* **21**, 289–357.

Hjort, N. and Varin, C. (2008), 'ML, PL, QL in Markov chain models', *Scandinavian Journal of Statistic* **35**(1), 64–82.

Joe, H. (1997), *Multivariate Models and Multivariate Dependence Concepts*, Chapman & Hall, London.

Joe, H. and Lee, Y. (2009), 'On weighting of bivariate margins in pairwise likelihood', *Journal of Multivariate Analysis* **100**(4), 670–685.

Kalbfleisch, J. (1978), 'Likelihood methods and nonparametric tests', *Journal of the American Statistical Association* **73**, 167–170.

Kalbfleisch, J. D., Song, P. X.-K. and Fan, Y. (2005), 'Maximization by parts in likelihood inference', *Journal of the American Statistical Association* **100**, 1145–1158.

Kent, J. (1982), 'Robust properties of likelihood ratio tests', *Biometrika* **69**(1), 19–27.

Kuk, A. (2007), 'A hybrid pairwise likelihood method', *Biometrika* **94**(4), 939–952.

Kuk, A. and Nott, D. (2000), 'A pairwise likelihood approach to analyzing correlated binary data', *Statistics and Probability Letters* **47**, 329–335.

Kuonen, D. (1999), 'Saddlepoint approximations for distributions of quadratic forms in normal variables', *Biometrika* **86**(4), 929–935.

Larribe, F. and Fearnhead, P. (2010), 'On composite likelihoods in statistical genetics', *Statistica Sinica* .

LeCessie, S. and van Houwelingen, J. C. (1994), 'Logistic regression for correlated binary data', *Applied Statistics* **43**(1), 95–108.

Lele, S. (2006), 'Sampling variability and estimates of density dependence: a composite-likelihood approach', *Ecology* **87**(1), 189–202.

Lele, S. and Taper, M. (2002), 'A composite likelihood approach to (co)variance components estimation', *Journal of Statistical Planning and Inference* **103**, 117–135.

Liang, G. and Yu, B. (2003), 'Maximum pseudo likelihood estimation in network tomography', *IEEE Transactions on Signal Processing* **51**, 2043–2053.

Liang, K.-Y. (1987), 'Extended Mantel-Haenszel estimating procedure for multivariate logistic regression models', *Biometrics* **43**(2), 289–299.

Liang, K.-Y. and Qin, J. (2000), 'Regression analysis under non-standard situations: a pairwise pseudolikelihood approach', *Journal of the Royal Statistical Society. Series B (Methodological)* **62**(4), 773–786.

Liang, K.-Y. and Zeger, S. (1986), 'Longitudinal data analysis using generalized linear models', *Biometrika* **73**(1), 13–22.

Lindsay, B. (1988), 'Composite likelihood methods', *Contemporary Mathematics* **80**, 220–239.

Lindsay, B. G. (1982), 'Conditional score functions: some optimality results', *Biometrika* **69**, 503–512.

Lindsay, B. G., Pilla, R. S. and Basak, P. (2000), 'Moment-based approximations of distributions using mixtures: theory and application', *Ann. Inst. Statist. Math.* **52**, 215–230.

Lindsay, B. G., Yi, G. Y. and Sun, J. (2010), 'Issues and strategies in the selection of composite likelihoods', *Statistica Sinica* .

Lipsitz, S., Dear, K. and Zhao, L. (1994), 'Jackknife estimators of variance for parameter estimates from estimating equations with applications to clustered survival data', *Biometrics* **50**, 842–846.

Lumley, T. and Heagerty, P. (1999), 'Weighted empirical adaptive variance estimators for correlated data regression', *Journal of the Royal Statistical Society. Series B (Methodological)* **61**(2), 459–477.

Mardia, K. V., Hughes, G. and Taylor, C. C. (2007), Efficiency of the pseudo-likelihood for multivariate normal and von mises distributions. Preprint.

Mardia, K. V., Hughes, G., Taylor, C. C. and Singh, H. (2008), 'A multivariate von Mises distribution with applications to bioinformatics', *Canadian Journal of Statistics* **36**(1), 99–109.

Mardia, K. V., Kent, J. T., Hughes, G. and Taylor, C. C. (2009), 'Maximum likelihood estimation using composite likelihoods for closed exponential families', *Biometrika* **96**, 975–982.

Mateu, J., Porcu, E., Christakos, G. and Bevilacqua, M. (2007), 'Fitting negative spatial covariances to geothermal field temperatures in Nea Kessani (Greece)', *Environmetrics* **18**(7), 759–773.

McCullagh, P. (1983), 'Quasi-likelihood functions', *Annals of Statistics* **11**, 59–67.

McFadden, D. and Train, K. (2000), 'Mixed MNL models for discrete responses', *Journal of Applied Econometrics* **15**, 447–470.

McLachlan, G. and Krishnan, T. (2008), *The EM Algorithm and Extensions. Second Edition*, Wiley, Hoboken, New Jersey.

Molenberghs, G., Kenward, M. G., Verbeke, G. and Birhanu, T. (2010), 'Pseudo-likelihood for incomplete data', *Statistica Sinica* .

Molenberghs, G. and Verbeke, G. (2005), *Models for Discrete Longitudinal Data*, Springer, New York.

Nott, D. and Rydén, T. (1999), 'Pairwise likelihood methods for inference in image models', *Biometrika* **86**(3), 661–676.

Okabayashi, S., Johnson, L. and Geyer, C. J. (2010), 'A composite likelihood extending pseudo-likelihood for Potts models', *Statistica Sinica* .

Oliveira, V. D. (2004), 'A simple model for spatial rainfall fields', *Stochastic Environmental Research and Risk Assessment* **18**, 131–140.

Oman, S., Landsman, V., Carmel, Y. and Kadmon, R. (2007), 'Analyzing spatially distributed binary data using independent-block estimating equations', *Biometrics* **63**(3), 892–890.

Pace, L., Salvan, A. and Sartori, N. (2010), 'Adjusting composite likelihood ratio statistics', *Statistica Sinica* .

Padoan, S., Ribatet, M. and Sisson, S. (2010), 'Likelihood-based inference for max-stable processes', *Journal of the American Statistical Association* **105**, 263–277.

Pakel, C., Shepard, N. and Sheppard, K. (2009), Nuisance parameters, composite likelihoods and a panel of garch models. manuscript.

Parzen, M., Lipsitz, S., Fitzmaurice, G. and Ibrahim, J. (2006), 'Pseudo-likelihood methods for longitudinal binary data with non-ignorable missing responses and covariates', *Statistics in Medicine* **25**, 2784–2796.

Parzen, M., Lipsitz, S., Fitzmaurice, G. and Ibrahim, J. (2007), 'Pseudo-likelihood methods for the analysis of longitudinal binary data subject to non-ignorable non-monotone missingness', *Journal of Data Science* **5**, 1–21.

R Development Core Team (2009), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL: *http://www.R-project.org*

Renard, D., Geys, H., Molenberghs, G., Burzykowski, T. and et al. (2002), 'Validation of surrogate endpoints in multiple randomized clinical trials with discrete outcomes', *Biometrical Journal* **8**, 921–935.

Renard, D., Molenberghs, G. and Geys, H. (2004), 'A pairwise likelihood approach to estimation in multilevel probit models', *Computational Statistics and Data Analysis* **44**, 649–667.

Ribatet, M. (2009), *A User's Guide to the SpatialExtremes Package*, EPFL, Lausanne, Switzerland.

Robins, J. (1995), 'Analysis of semiparametric regression models for repeated outcomes in the presence of missing data', *Journal of the American Statistical Association* **90**, 106–121.

Rotnitzky, A. and Jewell, N. (1990), 'Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data', *Biometrika* **77**(3), 485–497.

Rydén, T. (1994), 'Consistent and asymptotically normal parameter estimates for hidden Markov models', *The Annals of Statistics* **22**(4), 1884–1895.

Satterthwaite, F. E. (1946), 'An approximate distribution of estimates of variance components', *Biometrics Bulletin* **2**, 110–114.

Smith, E. and Stephenson, A. (2009), 'An extended Gaussian max-stable process model for spatial extremes', *Journal of Statistical Planning and Inference* **139**, 1266–1275.

Smith, R. (1990), Max-stable processes and spatial extremes. Unpublished.

Stein, M., Chi, Z. and Welty, L. (2004), 'Approximating likelihoods for large spatial data sets', *Journal of the Royal Statistical Society. Series B (Methodological)* **66**(2), 275–296.

Tibaldi, F., Molenberghs, G., Burzykowski, T. and Geys, H. (2004), 'Pseudo-likelihood estimation for a marginal multivariate survival model', *Statistics in Medicine* **23**, 924–963.

Troxel, A., Lipsitz, S. and Harrington, D. (2003), 'Marginal models for the analysis of longitudinal measurements with nonignorable non-monotone missing data', *Biometrika* **85**(3), 661–672.

Vandekerkhove, P. (2005), 'Consistent and asymptotically normal parameter estimates for hidden Markov mixtures of Markov models', *Bernoulli* **11**(11), 103–129.

Varin, C. (2008), 'On composite marginal likelihoods', *Advances in Statistical Analysis* **92**(1), 1–28.

Varin, C. and Czado, C. (2010), 'A mixed autoregressive probit model for ordinal longitudinal data', *Biostatistics* **11**, 127–138.

Varin, C., Høst, G. and Skare, Ø. (2005), 'Pairwise likelihood inference in spatial generalized linear mixed models', *Computational Statistics and Data Analysis* **49**, 1173–1191.

Varin, C. and Vidoni, P. (2005), 'A note on composite likelihood inference and model selection', *Biometrika* **92**(3), 519–528.

Varin, C. and Vidoni, P. (2006), 'Pairwise likelihood inference for ordinal categorical time series', *Computational Statistics and Data Analysis* **51**, 2365–2373.

Varin, C. and Vidoni, P. (2009), 'Pairwise likelihood inference for general state space models', *Econometric Reviews* **28**(1), 170–185.

Vecchia, A. V. (1988), 'Estimation and model identification for continuous spatial processes', *Journal of the Royal Statistical Society. Series B (Methodological)* **50**(2), 297–312.

Wang, M. and Williamson, J. M. (2005), 'Generalization of the Mantel-Haenszel estimating function for sparse clustered binary data', *Biometrics* **61**(4), 973–981.

Wang, Y. and Ip, E. (2008), 'Conditionally specified continuous distributions', *Biometrika* **95**(3), 735–746.

Wedderburn, R. (1974), 'Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method', *Biometrika* **61**, 439–447.

Wellner, J. A. and Zhang, Y. (2000), 'Two estimators of the mean of a counting process with panel count data', *Annals of Statistics* **28**, 779–814.

Wellner, J. A. and Zhang, Y. (2007), 'Two likelihood-based semiparametric estimation methdos for panel count data with covariates', *Annals of Statistics* **35**, 2106–2142.

White (1994), *Estimation, Inference and Specification Analysis*, Cambridge University Press, Cambridge.

Yi, G. Y. and He, W. (2010), 'A pairwise likelihood method for correlated binary data with/without missing observations under generalized partially linear single-index models', *Statistica Sinica* .

Yi, G. Y., Zeng, L. and Cook, R. J. (2009), A robust pairwise likelihood method for incomplete longitudinal binary data arising in clusters. *Canadian Journal of Statistics*, to appear.

Zhao, H. and Ma, W.-Q. (2009), 'A pairwise likelihood procedure for analyzing exchangeable binary data with random cluster sizes', *Communications in Statistics-Theory and Methods* **38**(5), 594–606.

Zhao, L. P. and Prenctice, R. L. (1990), 'Correlated binary regression using a quadratic exponential model', *Biometrika* **77**, 642–648.

Zhao, Y. and Joe, H. (2005), 'Composite likelihood estimation in multivariate data analysis', *Canadian Journal of Statistics* **33**, 335–356.

Zi, J. (2009), On some aspects of composite likelihood. PhD dissertation, University of Toronto.

Zidek, J. V. and Hu, F. (1997), 'The asymptotic properties of the maximum-relevance weighted likelihood estimators', *Canadian Journal of Statistics* **25**, 45–59.

Department of Statistics, Ca' Foscari University, 35121 Venice, Italy

E-mail: sammy@unive.it

Department of Statistics, University of Toronto, Toronto M5S 3GS, Canada

E-mail: reid@utstat.utoronto.ca

Department of Statistics, University of Warwick, Coventry CV4 7AL, UK

E-mail: d.firth@warwick.ac.uk