The Canadian Journal of Statistics
Vol. 40, No. 3, 2012, Pages 1–14
La revue canadienne de statistique

Likelihood inference in complex settings

N. REID[1]*

Department of Statistics, University of Toronto, 100 St George St., Toronto, ON, Canada M5S 3G3

Key words and phrases: Composite likelihood; Gaussian processes ; higher order asymptotics; pivotal functions.

MSC 2010: Primary 62F99; secondary 62P25

Abstract: Inference based on the likelihood function owes much to theory developed some decades ago. What is the current role of likelihood in developing strategies for the analysis of very large data sets, often with very high dimension, and complex dependencies? This paper considers some aspects of this question with emphasis on problems in stochastic modelling, estimating equations, and survey methodology. *The Canadian Journal of Statistics* 40: 1–14; 2012 © 2012 Statistical Society of Canada

Résumé[Q2]*:* Inference based on the likelihood function owes much to theory developed some decades ago. What is the current role of likelihood in developing strategies for the analysis of very large data sets, often with very high dimension, and complex dependencies? This paper considers some aspects of this question with emphasis on problems in stochastic modelling, estimating equations, and survey methodology. *La revue canadienne de statistique* 40: 1–14; 2012 © 2012 Société statistique du Canada

1. INTRODUCTION

It is a pleasure and an honour to contribute to the celebration of Professor Thompson's career in statistics. In this paper I will discuss some aspects of likelihood inference for complex problems, attempting to highlight some of the application areas that have benefitted from Mary's expertise, including stochastic processes and survey sampling.

By way of introduction, we may consider why the likelihood function plays such an important role in statistics. At a theoretical level, we can rely on the Lehmann–Scheffé results and say that the likelihood function captures all the information in the data, because it is a function of the sufficient statistics. This intuitive result has been mathematically generalized in various ways; for example, Fraser & Naderi (2007) prove a measure-theoretic result which they summarize as "the likelihood map is sufficient," building on Barndorff-Nielsen et al. (1996). From a more practical point of view, the likelihood function provides a set of summary statistics with known limiting distributions, and this leads to the construction of approximately pivotal functions that are easily used for inference based on the limiting normal distribution of these statistics. In some models, for example location models, the likelihood function gives exact inference, a result first derived in Fisher (1934). Developments of higher order approximations, based on refining the normal approximation to the standard summary statistics, show that combining the likelihood function with some extra information about its derivative on the sample space can give extremely accurate approximations to P values and confidence limits. Bayesian inference is based on the likelihood function combined with a prior, and in some cases the resulting inferences can be made very close to those from a non-Bayesian point of view by careful choice of the prior, although the resulting inference is not conventionally Bayesian.

* *Author to whom correspondence may be addressed.*
E-mail: reid@utstat.utoronto.ca

In the remainder of this Introduction I will set this out a little more precisely, and in Section 2 describe the extension of sufficiency embedded in higher order approximation. In Section 3 I look at a more complex model, and in Section 4 consider the role of composite likelihood in finding a compromise between computational tractability and the benefits of likelihood inference. I touch briefly on likelihood inference in sample surveys, and approximate Bayesian computation in Section 5.

1.1. Notation and Definitions

The likelihood inference that I discuss in this paper is essentially parametric, although there is an important literature on nonparametric likelihood inference and semi-parametric inference. We assume that we have a model $f(y;\theta)$ for an observation $y \in \mathbb{R}^n$, depending on a parameter $\theta \in \mathbb{R}^d$. The log-likelihood function is simply $\ell(\theta; y) = \log f(y;\theta)$, although more precisely $\ell(\theta; y) = \log f(y;\theta) + c(y)$ is an equivalence class of functions on the parameter space, up to additive terms that depend on the observations alone. The additive term is usually ignored, with the understanding that only differences in log-likelihoods are relevant. The maximum likelihood estimator $\hat\theta = \arg\sup \ell(\theta; y)$ is the point estimate associated with the likelihood function, and the log-likelihood ratio statistic $w(\theta_0) = 2\{\ell(\hat\theta; y) - \ell(\theta_0; y)\}$ provides a widely available test statistic for the hypothesis that $\theta = \theta_0$.

A more modern approach that downplays the traditional dichotomy of hypothesis testing and point estimation is to use asymptotic results to suggest distributional approximations, and to summarize this by describing a set of approximate pivotal quantities. For example, in regular models, under smoothness conditions on the family of models $\{f(y;\theta); \theta \in \mathbb{R}^d\}$, and (most easily) assuming the distribution of the components of y are independent, we have the asymptotic result

$$i^{1/2}(\theta)(\hat\theta - \theta) \xrightarrow{\mathcal{L}} N_d(0, I), \quad n \to \infty, \tag{1}$$

where $i(\theta) = \mathrm{E}\{-\partial^2\ell(\theta; y)/\partial\theta\partial\theta^T\}$ is the expected Fisher information in an observation Y from $f(y;\theta)$. Defining the observed Fisher information by

$$j(\hat\theta) = -\left.\frac{\partial^2\ell(\theta; y)}{\partial\theta\partial\theta^T}\right|_{\theta=\hat\theta}, \tag{2}$$

we can use (1) to derive the distribution approximation

$$q(\theta) = j^{1/2}(\hat\theta)(\hat\theta - \theta) \mathrel{\dot\sim} N(0, I). \tag{3}$$

This approximation, valid under the model $f(y;\theta)$, is usually called a first-order approximation as it is based on the limiting result (1). More precisely if the components of y are independent and identically distributed then $j(\hat\theta)$ is $O_p(n)$ and the relative error in using the normal density defined in (3) to the exact density is $O(n^{-1/2})$ in *moderate deviation regions*, $|\hat\theta - \theta| < \delta/\sqrt{n}$. In Equation (3), $j^{1/2}(\hat\theta)(\hat\theta - \theta)$ is an *approximate pivot*, a function of the data and the parameter with a known distribution. This can be used to construct P values or confidence limits for components of θ in the usual way, for example $\hat\theta_k \pm z_{\alpha/2}\{j^{-1}(\hat\theta)_{k,k}\}^{1/2}$ is an approximate $1 - \alpha$ confidence interval for θ_k, where $z_{\alpha/2}$ is the upper critical point of a standard normal, and $j^{-1}(\hat\theta)_{k,k}$ is the kth diagonal element of $j^{-1}(\hat\theta)$.

An approximate pivot based on the limiting distribution of the likelihood ratio statistic is

$$w(\theta) = 2\{\ell(\hat\theta) - \ell(\theta)\} \mathrel{\dot\sim} \chi^2_d \tag{4}$$

and in the special case that $d = 1$ we have

$$r(\theta) = \text{sign}(\hat{\theta} - \theta)[2\{\ell(\hat{\theta}) - \ell(\theta)\}]^{1/2} \dot{\sim} N(0, 1), \tag{5}$$

where $r(\theta)$ is often called the likelihood root.

Fisher's (1934) result giving the exact distribution of the maximum likelihood estimator in a location model is an exact pivot. Suppose $y = (y_1, \ldots, y_n)$ are independent and identically distributed with the distribution of each component given by $f_0(y_i - \theta)$, $\theta \in \mathbb{R}$. There is a one-to-one transformation of y to $(\hat{\theta}, a_1, \ldots, a_n)$, where $a_i = y_i - \hat{\theta}$, the vector a is constrained to $\mathbb{R}^{n-1}$ by the score equation defining $\hat{\theta}$, and a is *ancillary*; its distribution does not depend on θ. All the information about θ is contained in the conditional distribution of the maximum likelihood estimator, given a, and this has the exact density given by the renormalized likelihood function:

$$f(\hat{\theta} \mid a; \theta) = \frac{\exp\{\ell(\theta; y)\}}{\int \exp\{\ell(\theta; y)\}\, \mathrm{d}\theta} = \frac{\exp\left\{\sum \ell_0(\hat{\theta} + a_i - \theta)\right\}}{\int \exp\left\{\sum \ell_0(\hat{\theta} + a_i - \theta)\right\}} \tag{6}$$

where $\ell_0(\cdot) = \log f_0(\cdot)$ and the final expession makes explicit the dependence of the right hand side on $(\hat{\theta}, a)$. In other words, the likelihood function itself serves as a pivot (Hinkley, 1980).

2. HIGHER ORDER APPROXIMATION

The result in (6) holds, as an approximation, much more generally, as was outlined in a series of papers in *Biometrika* in 1980; the general result is often called Barndorff-Nielsen's p^* approximation, after Barndorff-Nielsen (1980, 1983), and is concisely expressed as

$$p^*(\hat{\theta} \mid a; \theta) = c(\theta, a)|j(\hat{\theta})|^{1/2} \exp\{\ell(\theta; \hat{\theta}, a) - \ell(\hat{\theta}; \hat{\theta}, a)\}, \tag{7}$$

where $c(\theta, a)$ is a normalizing constant for the density, often obtained numerically. This approximation to the conditional distribution of $\hat{\theta}$, given a, has relative error $O(n^{-3/2})$ in continuous models, for $\hat{\theta}$ in moderate deviations around θ; the derivation of this result is given in Skovgaard (1990). Note that to construct this approximation it is necessary to find a transformation from y to $(\hat{\theta}, a)$, where a is an ancillary statistic. This is quite easy in the location model, as at (6), and indeed in any transformation model, and this step is not needed in a full exponential family model, since the sufficient statistic has the same dimension as θ, but in general models it can be difficult to find such an ancillary statistic.

Since in any case we are often interested in the p^* approximation in order to compute tail area probabilities, or P values, an approximation to the integral of $p^*(\hat{\theta})$ with respect to $\hat{\theta}$ is of more interest than an approximation to the density function. This approximation is actually simpler, and can be obtained without finding an explicit form for an exact or approximately ancillary statistic. We avoid the specification of the transformation from $(y_1, \ldots, y_n)$ to $(\hat{\theta}, a)$ by defining a sample-space derivative of the log-likelihood function

$$\varphi(\theta) \equiv \ell_{;V}(\theta; y^0) = \frac{\partial}{\partial V(y)} \ell(\theta; y)\Big|_{y=y^0}. \tag{8}$$

This is a directional derivative at the observed data point, y^0 [illegible]rmined by the n vectors, each of length p, that make up the $n \times p$ matrix V; this directional derivative captures the aspect of the ancillary statistic that is needed to compute a P-value. The construction of the matrix V is described in detail in Brazzale, Davidson, & Reid (2007, Ch. 8), and illustrated there on a range of regression models. For completeness a brief description is given in the Appendix.

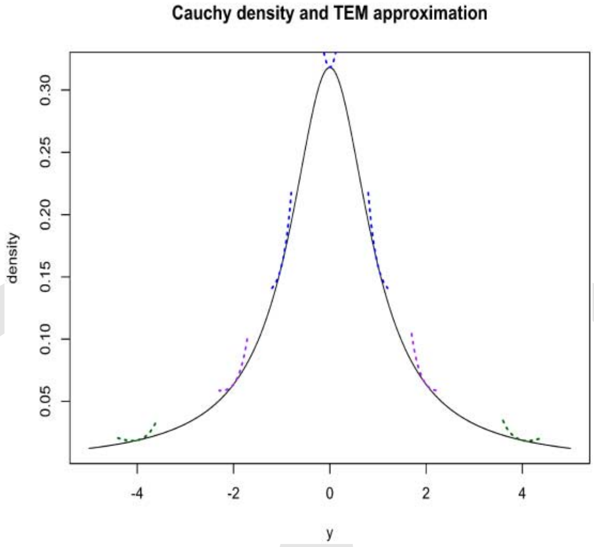

FIGURE 1: Comparison of the exact Cauchy density with a tangent exponential model approximation.

The reparametrization $\varphi(\theta)$ defined above can be used to define a *tangent exponential family* model on $\mathbb{R}^d$:

$$f_{\mathrm{TEM}}(s;\theta)\,\mathrm{d}s = \exp\{\varphi(\theta)'s + \ell^0(\theta)\}h(s)\,\mathrm{d}s, \tag{9}$$

where s is a score variable on $\mathbb{R}$, $\ell^0(\theta) = \ell(\theta; y^0)$ is the observed log-likelihood function, and $\varphi(\theta) = \varphi(\theta; y^0)$ is the directional derivative. The tangent exponential model is a p^*-type approximation to the conditional density, given an approximate ancillary statistic. Conditioning on the approximate ancillary statistic is implemented via the direction vectors that make up the matrix V used to define φ. Although (9) approximates the conditional model only to $O(n^{-1})$, when used to derive a P-value, that is, a distribution function at y^0, the resulting approximation is accurate to $O(n^{-3/2})$; see, for example, Fraser & Reid (1993) and Andrews, Fraser, & Wong (2005).

As an approximation to the density, $f(y;\theta_0)$, (9) is accurate only at the observed data point y^0; see Figure 1 for an illustration of the approximation to the Cauchy density. As an approximation to the log-likelihood function $\ell(\theta; y^0)$, it reproduces this function and its sample space derivative exactly. It is this property that enables use of the tangent exponential model to obtain approximations to P-values that have relative error $O(n^{-3/2})$ for continuous distributions. An illustration of the P-value approximation for the Cauchy is given in Brazzale, Davidson, & Reid (2007, Ch. 3.1).

In the case that θ is a scalar parameter, the P-value approximation derived from (9) is given by

$$p(\theta) = \Phi(r^*) \equiv \Phi\left\{r + \frac{1}{r}\log\left(\frac{Q}{r}\right)\right\}, \tag{10}$$

where $\Phi(\cdot)$ is the distribution function for a standard normal distribution, and r and Q are

$$r = \operatorname{sign}(\hat\theta - \theta)\left[2\left\{\ell^0(\hat\theta) - \ell^0(\theta)\right\}\right]^{1/2}, \tag{11}$$

$$Q = \operatorname{sign}(\hat\theta - \theta)|\varphi(\hat\theta) - \varphi(\theta)|\; j_{\varphi\varphi}^{-1/2}(\hat\varphi) \tag{12}$$

$$= \operatorname{sign}(\hat\theta - \theta)|\varphi(\hat\theta) - \varphi(\theta)|\; \varphi_\theta^{-1}(\hat\theta) j^{1/2}(\hat\theta); \tag{13}$$

note that r^* depends only on $\{\ell(\theta; y^0), \varphi(\theta; y^0)\}$ and their derivatives with respect to θ, evaluated at $\hat{\theta} = \hat{\theta}(y^0)$.

In (10) r^* is an approximate pivotal quantity, in the same way that r in (5) is an approximate pivotal quantity; the difference is that the normal approximation to the distribution of r^* is more accurate than the normal approximation to the distribution of r; the latter is accurate to $O(n^{-1/2})$, and the r^* approximation is accurate to $O(n^{-3/2})$ in continuous models.

Inference about a component of a vector parameter are more useful in practice, and versions of (3), (5) and r^* are similar in form to the scalar parameter versions, but require additional notation. We write $\theta = (\psi, \lambda)$, where ψ is the scalar parameter of interest and λ is the nuisance parameter, and $\hat{\theta}_\psi = (\psi, \hat{\lambda}_\psi)$ for the constrained maximum likelihood estimator of θ when ψ is fixed. Then (3) and (5) are replaced by

$$q(\psi) = \left\{j^{\psi\psi}(\hat{\theta})\right\}^{-1/2}(\hat{\psi} - \psi), \tag{14}$$
$$r(\psi) = \operatorname{sign}(\hat{\psi} - \psi)[2\{\ell(\hat{\theta}) - \ell(\hat{\theta}_\psi)\}]^{1/2},$$

where $j^{\psi\psi}(\hat{\theta})$ is the (ψ, ψ) component of the inverse of the observed Fisher information matrix. These pivotal quantities are simply the standardized maximum likelihood estimator, and the signed square root, respectively, treating the profile log-likelihood function $\ell_{\mathrm{p}}(\psi) = \ell(\hat{\theta}_\psi)$ as an ordinary log-likelihood function, because $\{j^{\psi\psi}(\hat{\theta})\}^{-1/2} = j_{\mathrm{p}}^{1/2}(\hat{\psi})$.

To construct the third-order pivotal quantity r^*, we use r as defined in (15) and extend the expression for Q in (13) to

$$Q = \operatorname{sign}(\hat{\psi} - \psi)\frac{|\varphi(\hat{\theta}) - \varphi(\hat{\theta}_\psi) \quad \varphi_\lambda(\hat{\theta}_\psi)|}{|\varphi_\theta(\hat{\theta})|}\left\{\frac{|j(\hat{\theta})|}{|j_{\lambda\lambda}(\hat{\theta}_\psi)|}\right\}^{1/2}, \tag{15}$$

where $j_{\lambda\lambda}(\theta)$ is the nuisance-parameter block of the observed Fisher information function. The two matrices in the second factor of (15) are $d \times d$: the numerator is shown partitioned into its first column, $\varphi(\hat{\theta}) - \varphi(\hat{\theta}_\psi)$, and the $(d-1) \times d$ sub-matrix $\varphi_\lambda(\hat{\theta}_\psi)$.

As an illustration, consider the model

$$y_i = \mu_i + \epsilon_i = \exp(\beta_0 + \beta_1 x_i) + \epsilon_i, \tag{16}$$

where we assume that ϵ_i follows a normal distribution with mean 0 and variance $\mu_i + \psi$. This is a simplified version of a model used in Hughes, Frick, & Hancock (2010) for analysing images developed by microscopic fluorescence. In their application the basic observation follows a Poisson distribution, but the counts tend to be large, and thus the normal approximation to the Poisson is used; ψ represents measurement error added to the Poisson fluctuations. In Hughes, Frick, & Hancock (2010) the model for μ_i was more complicated; it was derived from a spatial model for photon emission.

Figure 2 shows the distribution of r^* in simulations from (16), with ψ the parameter of interest, and β_0, β_1 treated as nuisance parameters. This is based on 1000 simulations of samples of size 30, with $x \sim U(0.5, 1)$ fixed for the simulations, and true values $(\beta_0, \beta_1, \psi) = (0, 8, 100)$. The details of the calculation of Q are given in a manuscript in preparation by Hoang and Reid.

3. MORE COMPLEX MODELS

Higher order approximation methods are easy to compute, and fun to use, on many regression type problems, but they are not well-suited to models with complex dependencies, as it can be difficult to determine the canonical parameter $\varphi(\theta)$ on which the approximation rests. The construction

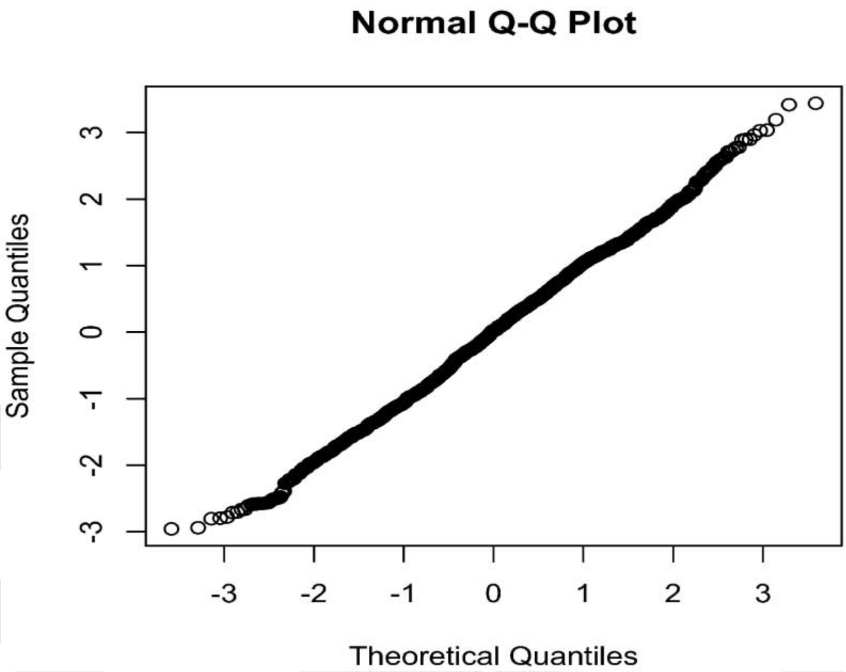

FIGURE 2: Simulations of the distribution of r^* for the Poisson–Normal model.

of this parameter implements conditioning on an approximately ancillary statistic, via a pivotal quantity z described in the Appendix, and when components of y are dependent the construction of the pivotal is more complicated, and may depend on the ordering of the components of y.

There are a great many application areas that require models with fairly complex structure, and Mary has made important contributions to many of these. Examples include analytic inference for survey data, stochastic processes in space or space-time (e.g., Serban, 2011; Bolin & Lindgren, 2011), models for extreme values in several dimensions, frailty models in survival data, various types of longitudinal data, family-based genetic data, estimation of recombination rates from SNP data, systems biology (Radde, Bar, & Achi[illegible]009), and many others.

A model widely used in the area of computer experiments and spatial data analysis is the Gaussian random field: suppose we have a scalar output y at a p-dimensional input $x = (x_1, \ldots, x_p)$, and that

$$y(x) = \phi(x)^T \beta + Z(x), \tag{17}$$

where $\phi(x)$ are a set of known basis functions, $Z(x)$ is a Gaussian process with

$$\mathrm{Cov}\{Z(x_1), Z(x_2)\} = \sigma^2 \prod_{i=1}^{p} R(|x_{1i} - x_{2i}|; \theta)$$

and the covariance matrix R is to be specified: for example

$$R(|x_{1i} - x_{2i}|) = \exp\left\{-\gamma_i |x_{1i} - x_{2i}|^{\alpha}\right\}.$$

In spatial data x is a two- or three-dimensional location vector, but in the context of computer experiments the p components of x represent different inputs in a simulator, for example, so anisotropic covariance matrices are more natural. The log-likelihood function for a sample of observations $y = (y_1, \ldots, y_m)$ at m locations $x_1, \ldots, x_m$, where each $x_i \in \mathbb{R}^p$ is easily written down,

$$\ell(\beta, \sigma, \theta) = -\frac{1}{2}\left\{m \log \sigma^2 + \log |R(\theta)| + \frac{1}{\sigma^2}(y - \Phi\beta)^{\mathrm{T}} R^{-1}(\theta)(y - \Phi\beta)\right\}; \tag{18}$$

here Φ is the design matrix of basis functions with m rows $\phi(x_i)^T$, each of length p. The computation of R^{-1} is $O(m^3)$, and for large m some simplification is needed. Two solutions proposed to simplify this computation are to enforce sparsity on the correlation matrix R, for example making in block diagonal, or to simplify the likelihood function, using ideas from composite likelihood.

A closely related model, sometimes appropriate in geostatistical applications, is the generalized linear model with mean function

$$E\{Y(x) \mid Z(x)\} = g\left\{\phi(x)^T \beta + Z(x)\right\}, \quad x \in \mathbb{R}^2 \text{ or } \mathbb{R}^3,$$

where now the random intercept $Z(x)$ is modelled as a stationary Gaussian process. The likelihood function involves integration over these random effects:

$$f(y; \theta) = \int_{\mathbb{R}^m} \prod_{i=1}^{n} f(y_i \mid z_i; \theta) f(\mathbf{z}; \theta)\, \mathrm{d}z_1 \ldots \mathrm{d}z_m,$$

where $z_i = z(x_i)$ is the random intercept associated with the ith location. This integral can be evaluated by simulation methods, but again composite likelihood provides a simplification.

4. COMPOSITE LIKELIHOOD

The strategy of constructing a function of the parameter that has similar properties to the likelihood function, but is easier to work with, goes back at least to the partial likelihood function for analysis of survival data (Cox, 1972) and the pseudo-likelihood for spatial data (Besag, 1974, 1975). One version of "likelihood-like" inference currently under active investigation is composite likelihood. Most applications of composite likelihood are targetted on models for multivariate observations, for which the evaluation of the joint distribution is very difficult, as, for example, the Gaussian process models described at the end of the previous section.

Suppose we have an m-dimensional variable Y with a model represented by a density $f(y; \theta)$, $\theta \in \mathbb{R}^d$. We define a set of marginal or conditional events $\{\mathcal{A}_1, \ldots, \mathcal{A}_K\}$ with associated "sub" log-likelihood

$$\ell_k(\theta; y) = \log f(y \in \mathcal{A}_k; \theta)$$

and define the composite log-likelihood by

$$\ell_C(\theta; y) = \sum_{k=1}^{K} \ell_k(\theta; y). \tag{19}$$

This is a "likelihood-like" inference function, obtained by pretending the sub-models are independent. Some general properties of estimating equations derived from composite likelihood were investigated in Lindsay (1988), who introduced the name composite likelihood. In many applications it makes sense to consider weighting the components, in which case we define

$$\ell_C(\theta; y) = \sum_{i=1}^{K} w_k \ell_k(\theta; y),$$

where $w_1, \ldots, w_k$ are a set of non-negative weights. Lindsay (1988) discussed optimal weighting based on the asymptotic variance of the resulting maximum composite likelihood estimator. Lindsay, Yi, & Sun (2011) provide a very general discussion of weighting, even allowing for the possibility of negative weights, with the goal of understanding how best to choose the set of component events $\{\mathcal{A}_k\}$.

Examples of composite likelihood include the independence log-likelihood $\ell_{\text{ind}}(\theta) = \sum_{r=1}^{m} \log f_1(y_r; \theta)$, which treats the components of y as independent, the pairwise log-likelihood

$$\ell_{\text{pair}}(\theta) = \sum_{r=1}^{m} \sum_{s>r} \log f_2(y_r, y_s; \theta), \tag{20}$$

where $f_2(y_r, y_s; \theta)$ is the marginal density of the pair (y_r, y_s), and Besag's (1975) pseudo-likelihood

$$\ell_{\text{pseudo}}(\theta) = \sum_{r=1}^{m} \log f(y_r \mid \{y_s : y_s \text{ neighbour of } y_r\}; \theta).$$

For the Gaussian random field (17), assuming $\sigma^2 = 1$, pairwise likelihood takes the form

$$\ell_{\text{pair}}(\theta) = -\frac{1}{2} \sum_{r=1}^{m-1} \sum_{s=r+1}^{m} \left\{ \log |R_{r,s}| + (y_{r,s} - \Phi_{r,s}\beta)^{\mathrm{T}} R_{r,s}^{-1} (y_{r,s} - \Phi_{r,s}\beta) \right\}, \tag{21}$$

where $y_{r,s} = (y_r, y_s)^T$, $\Phi_{r,s}$ is the $2 \times d$ sub-matrix of the design matrix Φ, and $R_{r,s}$ is the 2×2 correlation matrix for $y_{r,s}$. In $\ell_{\text{pair}}(\theta)$ the computational burden of computing the inverse of the $m \times m$ correlation matrix R is avoided.

Inference from composite likelihood proceeds by analogy with standard methods of likelihood inference; given a sample $y_1, \ldots, y_n$ of observations of y, we have

$$\ell_C(\theta; y) = \sum_{i=1}^{n} \ell_C(\theta; y_i) = \sum_{i=1}^{n} \sum_{k=1}^{K} \ell_k(\theta; y_i),$$

ignoring here the possibility of weighting. Because $\ell_C(\cdot)$ is built from component likelihoods, $U_C(\theta) = \ell'_C(\theta)$ is an unbiased estimating function, and under some regularity conditions on the model, a limiting normal distribution for U_C applies, leading to an approximation to the distribution of the maximum composite likelihood estimator $\hat{\theta}_C$, defined as the solution to $U_C(\hat{\theta}_C) = 0$:

$$\hat{\theta}_C \mathrel{\dot\sim} N\{\theta, G^{-1}(\theta)\},$$

where the approximate variance is given by the Godambe information

$$G(\theta) = \mathrm{E}\{-U'_C(\theta)\} \left[\mathrm{Var}\{U_C(\theta)\}\right]^{-1} \mathrm{E}\{-U'_C(\theta)\} = H(\theta) J^{-1}(\theta) H(\theta). \tag{22}$$

If $\mathrm{E}\{-U'_C(\theta)\} = \mathrm{Var}\{U_C(\theta)\}$ then the composite log-likelihood function is called information unbiased (Lindsay, 1988); this will not usually hold unless $\ell_C(\cdot)$ is a proper likelihood function.

The component sub-log-likelihoods $\ell_k(\cdot)$ are information unbiased, but not their sum. Godambe information arises as the asymptotic variance for estimating equations in the study of robustness (Kent, 1982). Godambe and Thompson established the optimality of the score equation among the family of unbiased estimating equations, first for the case of scalar parameter θ (Godambe, 1960), and later for the case of vector parameters (Godambe & Thompson, 1986); see also Godambe & Thompson (2009).

In the expression for pairwise likelihood for the Gaussian random field example in (21), $n = 1$, and the role of the sample size is taken by m. For the approximate inference outlined in the previous paragraph to be valid in this setting, it is necessary that the model provide internal replication, for example through the exponential decay of the spatial covariance matrix R. In general the asymptotic theory for composite likelihood when the length of the multivariate vector increases, but the sample size is fixed or increases slowly, needs to be considered on a case-by-case basis. Some discussion is provided in Cox & Reid (2004)[Q3], but rigorous asymptotic theory is lacking. Some work in the time series context is presented in Davis & Yau (2011).

A potential advantage of composite likelihood, beyond its use in providing estimating equations, is that it is itself an inference function: by analogy again we define the composite log-likelihood ratio statistic

$$w_C(\theta) = 2\left\{\ell_C\left(\hat{\theta}_C\right) - \ell_C(\theta)\right\}.$$

The asymptotic distribution of $w_C(\theta)$ is not χ^2_d, but rather a weighted sum of χ^2_1;

$$w_C(\theta) \dot{\sim} \sum_{i=1}^{d} \lambda_i \chi^2_{1i}, \tag{23}$$

where $\lambda_1, \ldots, \lambda_d$ are the eigenvalues of $H^{-1}(\theta)G(\theta) = J^{-1}(\theta)H(\theta)$, and $H(\theta)$, $J(\theta)$ and $G(\theta)$ are defined in (22). Pace, Sartori, & Salvan (2011) show how a rescaling of $w_C(\theta)$ can recover the more convenient χ^2_d approximation.

There are a great many applications of composite likelihood, several of which are surveyed in Varin, Reid, & Firth (2011), and many of these involve choosing some version of sub-likelihood event, $\{\mathcal{A}_k\}$, usually tailored to the application at hand, investigating both the computation and the quality of the inference (typically efficiency of $\hat{\theta}_C$), either by comparison with likelihood inference, if feasible, or by simulations. There is rather less work on strategies for the construction of composite likelihoods, which to date has proved difficult, partly because composite likelihood is so broadly defined. There are some surprises, though. In addressing the problem of optimal weights, for a given choice of sub-likelihoods, Lindsay, Yi, & Sun (2011) show that the optimal weights may be non-computable, or negative. Research in progress by Ximing Xu at the University of Toronto shows that surprises occur as well in the choice of the number and dimension of sub-likelihoods. In particular, examples can be constructed for which including additional sub-likelihood components leads to less efficient inferences. Similarly including higher-dimensional sub-likelihoods, for example going from independence likelihood to pairwise likelihood (20), can lead to less efficient estimation.

On the other hand, it seems plausible that inference based on a composite likelihood constructed from lower dimensional marginal distributions is robust against model misspecification of the full joint distribution, although even this seems difficult to make completely precise; some aspects are discussed in Xu & Reid (2011). At this stage of the development of the theory, there seem to be as many open problems as solved ones, and it is possible that theoretical results will be tied to particular classes of models. Kuk (2008) suggests an interesting hybrid strategy, using composite likelihood inference for aspects of the model of secondary interest, and more efficient techniques for estimation of the main parameters of interest. Yi, Zeng, & Cook (2011), and He & Ye (2011), discuss composite likelihood for longitudinal data with missing observations that has the very useful property of being independent of the mechanism that generates the missingness; this seems a very large advantage of composite likelihood, especially if it could be made more general.

5. CONNECTIONS TO SURVEY SAMPLING

In light of Mary's many contributions to theory and methods of survey sampling, and the complexity of using likelihood inference in survey sampling, it would seem of interest to investigate whether or not there is an opportunity for cross-fertilization of ideas. In survey sampling the parameters to be estimated are often properties of a well-defined population, from which a sample of more or less complexity may be available. A descriptive parameter for the population, $\theta_{\mathcal{P}}$, say, may be defined through an estimating equation

$$\sum_{i \in \mathcal{P}} U_i(\theta_{\mathcal{P}}) = 0,$$

where $\mathcal{P}$ is the population of interest. The estimating equation from the sample $\mathcal{S}$ is typically

$$\sum_{i \in \mathcal{S}} w_i U_i(\theta) = 0 \tag{24}$$

with the weights in the simplest case defined as $w_i = 1/\pi_i$, where π_i is the probability of selection of unit i. This leads to an estimate $\hat{\theta}_{\mathcal{P}}$ with variance given by the Godambe information function (22), but the version described here has no direct connection to likelihood inference, being essentially determined by the choice of estimating function $U(\cdot)$ and the sampling design.

However, it is usual in complex surveys that the estimating equation is motivated by a super-population model, that is, a probability density function for the distribution of values in the population, described by a model $f(y;\theta)$, $y \in \mathcal{P}$, $\theta \in \mathbb{R}^d$, say. In that setting $U_i(\cdot)$ would simply be the score function for the likelihood component $f(y_i;\theta)$, and the use of weights based on the sampling design in (24) is a form of model-assisted inference. As I understand it the goal is to obtain an estimator that is motivated by a plausible model for the data, but that has inferential properties valid under the sampling design, even if the super-population model is incorrect. Typically the weights used in (24) are more complicated than $1/\pi_i$, often written $1/\pi_i q_i$, where q_i represents various adjustments for non-response, post-stratification, and so on. It seems possible that there may be connections to be made to weighted composite likelihood, as in many applications of the latter the weights are designed to reflect properties of the sampling structure, such as cluster size, or observed cluster size in the case of missing data (Yi, Zeng, & Cook, 2011; He & Yi, 2011). Gelman (2007) and the many discussants to this article provide an interesting discussion of the art and science of designing survey weights; Little (2004) provides a helpful discussion of design-based and model-based inference in sample surveys.

Thompson (1997, Ch. 3) is one of the few places where one can find discussion of higher order asymptotic theory in sample surveys. While mathematically elegant, higher order asymptotic theory has not proved to be very useful for analysis of sample surveys; the use of bootstrap weights developed by Rao & Wu (1988) seems to have solved many inferential problems in a particularly elegant way. However, there are connections, as yet unexplored for survey data, between higher order asymptotics and the bootstrap!; see DiCiccio & Efron (1996).

Another likelihood approach for survey sampling is a nonparametric approach based on empirical likelihood, which builds on Hartley & Rao (1968) and Owen (1988). The usual empirical log-likelihood has the form for independent identically distributed sampling

$$\ell(F) = \sum_{i=1}^{n} \log p_i, \quad p_i > 0, \sum p_i = 1, \sum p_i y_i = \theta,$$

where p_i, the weights on observations y_i, are the unknown parameters in F, and $\theta = E_F(Y)$ is the parameter of interest. More generally θ can be defined by the solution of a particular estimating function. Chen & Sitter (1999) extended this to complex survey sampling by suggesting a weighted empirical likelihood of the form

$$\sum_{i \in \mathcal{S}} w_i \log p_i$$

with post-stratification constraints such as $\Sigma_{i \in \mathcal{S}} p_i x_i = \bar{X}_{\mathcal{P}}$ to incorporate known population information, and with weights inversely proportional to the probability of selection. This leads to estimates that are more efficient than the Horvitz-Thompson estimator. Wu & Rao (2006), and Rao & Wu (2010) suggested an alternative weighted empirical likelihood that enables a likelihood-ratio type χ^2 approximation, after a simple adjustment obtained from the design effect; the ratio of the variance of the sample mean under simple random sampling to the variance of the sample mean under the survey design.

This approach to likelihood-based inference is completely nonparametric, whereas composite likelihood is based on parametric families of distributions and their sub-families. There are however intriguing points of contact in the goals of making the inferences robust to model misspecification and in the use of rescaling to obtain a χ^2 limit for the likelihood-ratio type statistic. Further, composite likelihood methods are widely used in models for data with a complex structure due to sampling: often longitudinal, clustered, or hierarchical models, for example, and hierarchical structures feature naturally in aspects of survey sampling such as small area estimation and multiple frame surveys. Carillo, Chen, & Wu (2010) discuss a version of generalized estimating equations for longitudinal survey data; there are several parallels between generalized estimating equations and composite likelihood estimating equations but a detailed comparison is not yet available.

Another approach to computationally intractable likelihood functions is simulation: this was discussed in connection with applications to genetics in Geyer & Thompson (1996), but more recently a suite of methods under the name Approximate Bayesian Computation are being investigated; a review is provided in Marin et al. (2011). To date the method is mainly used to simulate from the posterior distribution, and to provide approximations to Bayes factors in problems of model choice. Central to the method is the choice of some summary statistics, and it is interesting to see how the notions of sufficiency and ancillarity come into the discussions: see, for example, Robert et al. (2011), Fearnhead & Prangle (2012), and Marin et al. (2010). As sufficiency is arguably the most important aspect of the likelihood function, and ancillarity is key to the development of pivotal quantities, it seems that the "classical" theory of likelihood has much to offer to modern uses of statistical methods.

APPENDIX

The construction of the $n \times p$ matrix V described in Section 2 is described in Brazzale, Davidson, & Reid (2007, Ch. 8). Broadly speaking, V is computed as the rate of change of the observation vector y, with respect to the parameter θ, where y and θ are linked through the model $f(y;\theta)$.

This is easiest to compute by the formula

$$V = -\left(\frac{\partial z}{\partial y}\right)^{-1}\left(\frac{\partial z}{\partial \theta}\right)\bigg|_{y=y^0,\theta=\hat{\theta}},$$

where $z = z(y, \theta)$ is a pivotal quantity with a fixed distribution. For example, if $y_i \sim N(\mu_i, \mu_i + \psi)$, as in Section 2, then we can write

$$z_i = (y_i - \mu_i)/\sqrt{\mu_i + \psi}$$

and the components z_i of z are independent and follow a standard normal distribution under the model. It is then easily verified for this model that the ith row of the matrix V is

$$\left(\frac{\hat{\mu}_i(2\hat{\psi} + y_i + \hat{\mu}_i)}{2(\hat{\mu}_i + \hat{\psi})}, \frac{\hat{\mu}_i(2\hat{\psi} + y_i + \hat{\mu}_i)x_i}{2(\hat{\mu}_i + \hat{\psi})}, \frac{y_i - \hat{\mu}_i}{2(\hat{\mu}_i + \hat{\psi})}\right)$$

and

$$\varphi(\theta) = \sum_{i=1}^{n} \frac{\partial \ell(\theta; y)}{\partial y_i}\bigg|_{y=y^0} V_i$$

and its derivatives are readily obtained.

ACKNOWLEDGEMENTS

This research was partially supported by the Natural Sciences and Engineering Research Council of Canada. I am grateful to Uyen Hoang, Ximing Xu and Bo Chen for helpful discussions of the material, and especially to Wei Lin for her extensive background research on likelihood methods in sampling. Two reviewers made very helpful comments and corrections to the first version.

BIBLIOGRAPHY

Andrews, D. A., Fraser, D. A. S., & Wong, A. C. M. (2005). Computation of distribution functions from likelihood information near observed data. *Journal of Statistical Planning and Inference*, 134, 180–193.

Barndorff-Nielsen, O. E. (1980). Conditionality resolutions. *Biometrika*, 67, 293–310.

Barndorff-Nielsen, O. E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, 70, 343–365.

Barndorff-Nielsen, O. E., Hoffman-Jørgensen, J., & Pederson, K. (1976). On the minimal sufficiency of the likelihood function. *Scandinavian Journal of Statistics*, 3, 37–38.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society Series B*, 36, 192–236.

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

Bolin, D. & Lindgren, F. (2011). Spatial models generated by nested stochastic partial differential equations, with an application to global ozone mapping. *Annals of Applied Statistics*, 5, 523–550.

Brazzale, A. R., Davidson, A. C., & Reid, N. (2007). *Applied Asymptotics*, Cambridge University Press, Cambridge.

Carillo, I. A., Chen, J., & Wu, C. (2010). The pseudo-GEE approach to the analysis of longitudinal surveys. *Canadian Journal of Statistics*, 38, 540–554.

Chen, J. & Sitter, R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica*, 9, 385–406.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society Series B*, 34, 187–220.

Davis, R. A. & Yau, C. Y. (2011). Comments on pairwise likelihood in time series models. *Statistica Sinica*, 21, 255–277.

DiCiccio, T. J. & Efron, B. (1996). Bootstrap confidence intervals. (with discussion). *Statistical Science*, 11, 189–228.

Fearnhead, P. & Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation (with discussion). *Journal of the Royal Statistical Society Series B*, 74, 419–474.

Fisher, R. A. (1934). Two new properties of mathematical likelihood. *Proceedings of the Royal Society, Series A*, 144, 285–304.

Fraser, D. A. S. & Naderi, A. (2007). Minimal sufficient statistics emerge from the observed likelihood function. *International Journal of Statistical Science*, 6, 55–61.

Fraser, D. A. S. & Reid, N. (1993). Third order asymptotic models: likelihood functions leading to accurate approximations for distribution functions. *Statistica Sinica*, 3, 67–82.

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, 22, 153–164.

Geyer, C. J. & Thompson, E. A. (1996). Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society Series B*, 54, 657–699.

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

Godambe, V. P. & Thompson, M. E. (1986). Parameters of superpopulation and survey population: Their relationships and estimation. *International Statistical Review*, 54, 127–139.

Godambe, V. P. & Thompson, M. E. (2009). Estimating functions and survey sampling. In *Handbook of Statistics. Sample Surveys: Inference and Analysis*, Vol. 29B, Rao, C. R. & Pfeffermann, D., editors[Q4]. pp. 669–687.

Hartley, H. O. & Rao, J. N. K. (1968). A new estimation theory for sample surveys. *Biometrika*, 55, 547–557.

He, W. & Yi, G. Y. (2011). A pairwise likelihood for correlated binary data with/without missing observations under generalized partially linear single-index models. *Statistica Sinica*, 21, 207–229.

Hinkley, D. V. (1980). Likelihood as approximate pivotal. *Biometrika*, 67, 287–292.

Hughes, J., Frick, J., & Hancock, W. (2010). Likelihood inference for particle location in fluorescence microscopy. *Annals of Applied Statistics*, 4, 830–848.

Kent, J. T. (1982). Robust properties of the likelihood ratio test. *Biometrika*, 69, 19–27.

Kuk, A. Y. C. (2008). A hybrid pairwise likelihood method. *Biometrika*, 94, 939–952.

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

Lindsay, B. G., Yi, G. Y., & Sun, J. (2011). Issues and strategies in the selection of composite likelihoods. *Statistica Sinica*, 21, 71–106.

Little, R. J. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, 99, 546–556.

Marin, J., Pillai, N. S., Robert, C. P., & Rousseau, J. (2010). Relevant statistics for Bayesian model choice. http://arxiv.org/abs/1110.4700

Marin, J.-M., Pudlo, P., Robert, C. P., & Ryder, R. J. (2011). Approximate Bayesian computational methods. *Statistics and Computing*, 21, 1–14.

Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75, 237–249.

Pace, L., Sartori, N., & Salvan, A. (2011). Adjusting composite likelihood ratio statistics. *Statistica Sinica*, 21, 129–148.

[illegible]e, N., Bar, N. S., & Achim, T. (2009). A comparison of likelihoods for dynamic stochastic models of biological networks. In *Proceedings of Workshop of Computational Biology (WCSB09), Aarhus, Denmark*, pp. 131–134.

Rao, J. N. K. & Wu, C. F. J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231–241.

Rao, J. N. K. & Wu, C. (2010). PseudoÐempirical likelihood inference for multiple frame surveys. *Journal of the American Statistical Association*, 105, 1494–1503.

Robert, C. P., Cornuet, J.-M., Marin, J.-M., & Pillai, N. S. (2011). Lack of confidence in approximate Bayesian computational (ABC) model choice. *Proceedings of the National Academy of Science*, 108, 15112–15117.

Serban, N. (2011). A space-time varying coefficient model: the equity of service accessibility. *Annals of Applied Statistics*, 5, 2024–2051.

Skovgaard, I. M. (1990). On the density of minimum contrast estimators. *Annals of Statistics*, 18, 779–789.

Thompson, M. E. (1997). *Theory of Sample Surveys.* Chapman & Hall, London.

Varin, C., Reid, N., & Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21, 5–42.

Wu, C. & Rao, J. N. K. (2006). Pseudo-empirical likelihood ratio confidence intervals for complex surveys. *Canadian Journal of Statistics*, 34, 359–375.

Xu, X. & Reid, N. (2011). On the robustness of maximum composite likelihood. *Journal of Statistical Planning and Inference*, 141, 3047–3054.

Yi, G. Y., Zeng, L., & Cook, R. J. (2011). A robust pairwise likelihood method for incomplete longitudinal binary data arising in clusters. *Canadian Journal of Statistics*, 39, 34–51.

Received 18 March 2012
Accepted 17 July 2012

Q1: Author: A running head short title was not supplied; please check if this one is suitable and, if not, please supply a short title of up to 45 characters that can be used instead.
Q2: Author: The journal will supply French conversion of English Abstract.
Q3: Author: Cox & Reid (2004) has not been cited in the text. Please indicate where it should be cited; or delete from the Reference List and renumber the References in the text and Reference List.
Q4: Author: Please provide the publisher's name and location.