Bernoulli **19**(4), 2013, 1404–1418
DOI: 10.3150/12-BEJSP03

Aspects of likelihood inference

NANCY REID

Department of Statistics, University of Toronto, 100 St. George St., Toronto, Canada M5S 3G3.
E-mail: reid@utstat.utoronto.ca; url: www.utstat.utoronto.ca/reid/

I review the classical theory of likelihood based inference and consider how it is being extended and developed for use in complex models and sampling schemes.

Keywords: approximate pivotal quantities; composite likelihood; Laplace approximation; nuisance parameter; parametric inference; r^* approximation

1. Introduction

Jakob Bernoulli's *Ars Conjectandi* established the field of probability theory, and founded a long and remarkable mathematical development of deducing patterns to be observed in sequences of random events. The theory of statistical inference works in the opposite direction, attempting to solve the inverse problem of deducing plausible models from a given set of observations. Laplace pioneered the study of this inverse problem, and indeed he referred to his method as that of inverse probability.

The likelihood function, introduced by Fisher (1922), puts this inversion front and centre, by writing the probability model as a function of unknown parameters in the model. This simple, almost trivial, change in point of view has profoundly influenced the development of statistical theory and methods. In the early days, computing data summaries based on the likelihood function could be computationally difficult, and various *ad hoc* simplifications were proposed and studied. By the late 1970s, however, the widespread availability of computing enabled a parallel development of widespread implementation of likelihood-based inference. The development of simulation and approximation methods that followed meant that both Bayesian and non-Bayesian inferences based on the likelihood function could be readily obtained.

As a result, construction of the likelihood function, and various summaries derived from it, is now a nearly ubiquitous starting point for a great many application areas. This has a unifying effect on the field of applied statistics, by providing a widely accepted standard as a starting point for inference.

With the explosion of data collection in recent decades, realistic probability models have continued to grow in complexity, and the calculation of the likelihood function can again be computationally very difficult. Several lines of research in active development concern methods to compute approximations to the likelihood function, or inference functions with some of the properties of likelihood functions, in these very complex settings.

In the following section, I will summarize the standard methods for inference based on the likelihood function, to establish notation, and then in Section 3 describe some aspects of more accurate inference, also based on the likelihood function. In Section 4, I describe some exten-

1350-7265

sions of the likelihood function that have been proposed for models with complex dependence structure, with particular emphasis on composite likelihood.

2. Inference based on the likelihood function

Suppose we have a probability model for an observable random vector $Y = (Y_1, \ldots, Y_n)$ of the form $f(y; \theta)$, where θ is a vector of unknown parameters in the model, and $f(y; \theta)$ is a density function with respect to a dominating measure, usually Lebesgue measure or counting measure, depending on whether our observations are discrete or continuous. Typical models used in applications assume that θ could potentially be any value in a set Θ; sometimes Θ is infinite-dimensional, but more usually $\Theta \subset \mathbb{R}^d$. The inverse problem mentioned in Section 1 is to construct inference about the value or values of $\theta \in \Theta$ that could plausibly have generated an observed value $y = y^0$. This is a considerable abstraction from realistic applied settings; in most scientific work such a problem will not be isolated from a series of investigations, but we can address at least some of the main issues in this setting.

The likelihood function is simply

$$L(\theta; y) \propto f(y; \theta); \tag{2.1}$$

i.e., there is an equivalence class of likelihood functions $L(\theta; y) = c(y) f(y; \theta)$, and only relative ratios $L(\theta_2; y)/L(\theta_1; y)$ are uniquely determined. From a mathematical point of view, (2.1) is a trivial re-expression of the model $f(y; \theta)$; the re-ordering of the arguments is simply to emphasize in the notation that we are more interested in the θ-section for fixed y than in the y-section for fixed θ. Used directly with a given observation y^0, $L(\theta; y^0)$ provides a ranking of relative plausibility of various values of θ, in light of the observed data.

A form of direct inference can be obtained by plotting the likelihood function, if the parameter space is one- or two-dimensional, and several writers, including Fisher, have suggested declaring values of θ in ranges determined by likelihood ratios as plausible, or implausible; for example, Fisher (1956) suggested that values of θ for which $L(\hat{\theta}; y)/L(\theta; y) > 15$, be declared 'implausible', where $\hat{\theta} = \hat{\theta}(y)$ is the maximum likelihood estimate of θ, i.e., the value for which the likelihood function is maximized, over θ, for a given y.

In general study of statistical theory and methods we are usually interested in properties of our statistical methods, in repeated sampling from the model $f(y; \theta_0)$, where θ_0 is the notional 'true' value of θ that generated the data. This requires considering the distribution of $L(\theta; Y)$, or relative ratios such as $L\{\hat{\theta}(Y); Y\}/L\{\theta(Y); Y\}$. To this end, some standard summary functions of $L(\theta; Y)$ are defined. Writing $\ell(\theta; Y) = \log L(\theta; Y)$ we define the *score function* $u(\theta; Y) = \partial \ell(\theta; Y)/\partial \theta$, and the observed and expected Fisher information functions:

$$j(\theta; Y) = -\frac{\partial^2 \ell(\theta; Y)}{\partial \theta \, \partial \theta^T}, \qquad i(\theta) = E\left\{-\frac{\partial^2 \ell(\theta; Y)}{\partial \theta \, \partial \theta^T}\right\}. \tag{2.2}$$

If the components of Y are independent, then $\ell(\theta; Y)$ is a sum of independent random variables, as is $u(\theta; Y)$, and under some conditions on the model the central limit theorem for $u(\cdot; Y)$

leads to the following asymptotic results, as $n \to \infty$:

$$s(\theta) = j^{-1/2}(\hat{\theta})u(\theta) \xrightarrow{\mathcal{L}} N(0, I), \tag{2.3}$$

$$q(\theta) = j^{1/2}(\hat{\theta})(\hat{\theta} - \theta) \xrightarrow{\mathcal{L}} N(0, I), \tag{2.4}$$

$$w(\theta) = 2\big\{\ell(\hat{\theta}) - \ell(\theta)\big\} \xrightarrow{\mathcal{L}} \chi^2_d, \tag{2.5}$$

where we suppress the dependence of each derived quantity on Y (and on n) for notational convenience. These results hold under the model $f(y; \theta)$; a more precise statement would use the true value θ_0 in $u(\theta)$, $(\hat{\theta} - \theta)$, and $\ell(\theta)$ above, and the model $f(y; \theta_0)$. However, the quantities $s(\theta)$, $q(\theta)$ and $w(\theta)$, considered as functions of both θ and Y, are approximate *pivotal quantities*, i.e., they have a known distribution, at least approximately. For $\theta \in \mathbb{R}$ we could plot, for example, $\Phi\{q(\theta)\}$ as a function of θ, where $\Phi(\cdot)$ is the standard normal distribution function, and obtain approximate p-values for testing any value of $\theta \in \mathbb{R}$ for fixed y. The approach to inference based on these pivotal quantities avoids the somewhat artificial distinction between point estimation and hypothesis testing. When $\theta \in \mathbb{R}$, an approximately standard normal pivotal quantity can be obtained from (2.5) as

$$r(\theta) = \text{sign}(\hat{\theta} - \theta)\big[2\big\{\ell(\hat{\theta}) - \ell(\theta)\big\}\big]^{1/2} \xrightarrow{\mathcal{L}} N(0, 1). \tag{2.6}$$

The likelihood function is also the starting point for Bayesian inference; if we model the unknown parameter as a random quantity with a postulated prior probability density function $\pi(\theta)$, then inference given an observed value $Y = y$ is based on the posterior distribution, with density

$$\pi(\theta \mid y) = \frac{\exp\{\ell(\theta; y)\}\pi(\theta)}{\int \exp\{\ell(\phi; y)\}\pi(\phi)\, d\phi}. \tag{2.7}$$

Bayesian inference is conceptually straightforward, given a prior density, and computational methods for estimating the integral in the denominator of (2.7), and associated integrals for marginal densities of components, or low-dimensional functions of θ, have enabled the application of Bayesian inference in models of considerable complexity. Two very useful methods include Laplace approximation of the relevant integrals, and Markov chain Monte Carlo simulation from the posterior. Difficulties with Bayesian inference include the specification of a prior density, and the meaning of probabilities for parameters of a mathematical model.

One way to assess the influence of the prior is to evaluate the properties of the resulting inference under the sampling model, and under regularity conditions similar to those needed to obtain (2.3), (2.4) and (2.5), a normal approximation to the posterior density can be derived:

$$\pi(\theta \mid y) \dot{\sim} N\big\{\hat{\theta}, j^{-1}(\hat{\theta})\big\}, \tag{2.8}$$

implying that inferences based on the posterior are asymptotically equivalent to those based on q. This simple result underlines the fact that Bayesian inference will in large samples give approximately correct inference under the model, and also that to distinguish between Bayesian and non-Bayesian approaches we need to consider the next order of approximation.

If $\theta \in \mathbb{R}^d$, then (2.3)–(2.5) can be used to construct confidence regions, or to test simple hypotheses of the form $\theta = \theta_0$, but in many settings θ can usefully be separated into a parameter of interest ψ, and a nuisance parameter λ, and analogous versions of the above limiting results in this context are

$$s(\psi) = j_{\mathrm{p}}^{-1/2}(\hat{\psi})\ell'_{\mathrm{p}}(\psi) \xrightarrow{\mathcal{L}} N(0, I), \tag{2.9}$$

$$q(\psi) = j_{\mathrm{p}}^{1/2}(\hat{\psi})(\hat{\psi} - \psi) \xrightarrow{\mathcal{L}} N(0, I), \tag{2.10}$$

$$w(\psi) = 2\big\{\ell_{\mathrm{p}}(\hat{\psi}) - \ell_{\mathrm{p}}(\psi)\big\} \xrightarrow{\mathcal{L}} \chi^2_{d_1}, \tag{2.11}$$

where $\ell_{\mathrm{p}}(\psi) = \ell(\psi, \hat{\lambda}_\psi)$ is the profile log-likelihood function, $\hat{\lambda}_\psi$ is the constrained maximum likelihood estimate of the nuisance parameter λ when ψ is fixed, d_1 is the dimension of ψ, and $j_{\mathrm{p}}(\psi) = -\partial^2 \ell_{\mathrm{p}}(\psi)/\partial\psi\,\partial\psi^T$ is the Fisher information function based on the profile log-likelihood function.

The third result (2.11) can be used for model assessment among nested models; for example, the exponential distribution is nested within both the Gamma and Weibull models, and a test based on w of, say, a gamma model with unconstrained shape parameter, and one with the shape parameter set equal to 1, is a test of fit of the exponential model to the data; the rate parameter is the nuisance parameter λ. The use of the log-likelihood ratio to compare two non-nested models, for example a log-normal model to a gamma model, requires a different asymptotic theory (Cox and Hinkley, 1974, Ch. 8).

A related approach to model selection is based on the Akaike information criterion,

$$AIC = -2\ell(\hat{\theta}) + 2d,$$

where d is the dimension of θ. Just as only differences in log-likelihoods are relevant, so are differences in *AIC*: for a sequence of model fits the one with the smallest value of *AIC* is preferred. The *AIC* criterion was developed in the context of prediction in time series, but can be motivated as an estimate of the Kullback-Leibler divergence between a fitted model and a notional 'true' model. The statistical properties of *AIC* as a model selection criterion depend on the context; for example for choosing among a sequence of regression models of the same form, model selection using *AIC* is not consistent (Davison, 2003, Ch. 4.7). Several related versions of model selection criterion have been suggested, including modifications to *AIC*, and a version motivated by Bayesian arguments,

$$BIC = -2\ell(\hat{\theta}) + d\log(n),$$

where n is the sample size for the model with d parameters.

3. More accurate inference

The approximate inference suggested by the approximate pivotal quantities (2.9), (2.10) and (2.11) is obtained by treating the profile log-likelihood function as if it were a genuine log-likelihood function, i.e. as if the true value of λ were $\hat{\lambda}_\psi$. This can be misleading, because it does

not account for the fact that the nuisance parameter has been estimated. One familiar example is inference for the variance in a normal theory linear regression model; the maximum likelihood estimate is

$$\hat{\sigma}^2 = (y - X\hat{\beta})^T (y - X\hat{\beta})/n,$$

which has expectation $(n-k)\sigma^2/n$, where k is the dimension of β. Although this estimator is consistent as $n \to \infty$ with k fixed, it can be a poor estimate for finite samples, especially if k is large relative to n, and the divisor $n-k$ is used in practice. One way to motivate this is to note that $n\hat{\sigma}^2/(n-k)$ is unbiased for σ^2; an argument that generalizes more readily is to note that the likelihood function $L(\beta, \sigma^2; \hat{\beta}, \hat{\sigma}^2)$ can be expressed as

$$L_1\big(\mu, \sigma^2; \hat{\beta}\big) L_2\big(\sigma^2; \hat{\sigma}^2\big),$$

where L_1 is proportional to the density of $\hat{\beta}$ and L_2 is the marginal density of $\hat{\sigma}^2$ or equivalently $(y - X\hat{\beta})^T (y - X\hat{\beta})$. The unbiased estimate of σ^2 maximizes the second component L_2, which is known as the restricted likelihood, and estimators based on it often called "REML" estimators.

Higher order asymptotic theory for likelihood inference has proved to be very useful for generalizing these ideas, by refining the profile log-likelihood to take better account of the nuisance parameter, and has also provided more accurate distribution approximations to pivotal quantities. Perhaps most importantly, for statistical theory, higher order asymptotic theory helps to clarify the role of the likelihood function and the prior in the calibration of Bayesian inference. These three goals have turned out to be very intertwined.

To illustrate some aspects of this, consider the marginal posterior density for ψ, where $\theta = (\psi, \lambda)$:

$$\pi_m(\psi \mid y) = \frac{\int \exp\{\ell(\psi, \lambda; y)\}\pi(\psi, \lambda)\, d\lambda}{\int \exp\{\ell(\psi, \lambda; y)\}\pi(\psi, \lambda)\, d\lambda\, d\psi}. \tag{3.1}$$

Laplace approximation to the numerator and denominator integrals leads to

$$\begin{aligned}
\pi_m(\psi \mid y) &\doteq \frac{(2\pi)^{(d-d_1)/2} \exp\{\ell(\psi, \hat{\lambda}_\psi)\} |j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|^{-1/2} \pi(\psi, \hat{\lambda}_\psi)}{(2\pi)^{d/2} \exp\{\ell(\hat{\psi}, \hat{\lambda})\} |j(\hat{\psi}, \hat{\lambda})|^{-1/2} \pi(\hat{\psi}, \hat{\lambda})} \qquad (3.2)\\
&= \frac{1}{(2\pi)^{d_1/2}} \exp\big\{\ell_{\mathrm{p}}(\psi) - \ell_{\mathrm{p}}(\hat{\psi})\big\} |j_{\mathrm{p}}(\hat{\psi})|^{1/2} \left\{ \frac{|j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|}{|j_{\lambda\lambda}(\hat{\psi}, \hat{\lambda})|} \right\}^{-1/2} \frac{\pi(\psi, \hat{\lambda}_\psi)}{\pi(\hat{\psi}, \hat{\lambda})},\\
&= \frac{1}{(2\pi)^{d_1/2}} \exp\big\{\ell_{\mathrm{a}}(\psi) - \ell_{\mathrm{a}}(\hat{\psi})\big\} \frac{\pi(\psi, \hat{\lambda}_\psi)}{\pi(\hat{\psi}, \hat{\lambda})},
\end{aligned}$$

where $j_{\lambda\lambda}(\theta)$ is the block of the observed Fisher information function corresponding to the nuisance parameter λ, $|j(\hat{\theta})|$ has been computed using the partitioned form to give the second expression in (3.2), and in the third expression

$$\ell_{\mathrm{a}}(\psi) = \ell_{\mathrm{p}}(\psi) - (1/2) \log\big|j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)\big|.$$

When renormalized to integrate to one, this Laplace approximation has relative error $O(n^{-3/2})$ in independent sampling from a model that satisfies various regularity conditions similar to those needed to show the asymptotic normality of the posterior (Tierney and Kadane, 1986).

These expressions show that an adjustment for estimation of the nuisance parameter is captured in $\log|j_{\lambda\lambda}(\cdot)|$, and this adjustment can be included in the profile log-likelihood function, as in the third expression in (3.2), or tacked onto it, as in the second expression. The effect of the prior is isolated from this nuisance parameter adjustment effect, so, for example, if $\hat{\lambda}_\psi = \hat{\lambda}$, and the priors for ψ and λ are independent, then the form of the prior for λ given ψ does not affect the approximation.

The adjusted profile log-likelihood function $\ell_{\mathrm{a}}(\psi)$ is the simplest of a number of modified profile log-likelihood functions suggested in the literature for improved frequentist inference in the presence of nuisance parameters, and was suggested for general use in Cox and Reid (1987), after reparametrizing the model to make ψ and λ orthogonal with respect to expected Fisher information, i.e., $\mathrm{E}\{-\partial^2\ell(\psi,\lambda)/\partial\psi\,\partial\lambda\} = 0$. This reparameterization makes it at least more plausible that ψ and λ could be modelled as *a priori* independent, and also ensures that $\hat{\lambda}_\psi - \hat{\lambda} = O_p(1/n)$, rather than the usual $O_p(1/\sqrt{n})$.

A number of related, but more precise, adjustments to the profile log-likelihood function have been developed from asymptotic expansions for frequentist inference, and take the form

$$\ell_{\mathrm{M}}(\psi) = \ell_{\mathrm{p}}(\psi) + (1/2)\log|j_{\lambda\lambda}(\psi,\hat{\lambda}_\psi)| + B(\psi), \tag{3.3}$$

where $B(\psi) = O_p(1)$; see, for example, DiCiccio and Martin (1993) and Pace and Salvan (2006). The change from $-1/2$ to $+1/2$ is related to the orthogonality conditions; in (3.3) orthogonality of parameters is not needed, as the expression is parameterization invariant.

Inferential statements based on approximations from (2.9)–(2.11), with $\ell_{\mathrm{a}}(\psi)$ or $\ell_{\mathrm{M}}(\psi)$ substituting for the profile log-likelihood function, are still valid and are more accurate in finite samples, as they adjust for errors due to estimation of λ. They are still first-order approximations, although often quite good ones.

One motivation for these modified profile log-likelihood functions, and inference based on them, is that they approximate marginal or conditional likelihoods, when these exist. For example, if the model is such that

$$f(y;\psi,\lambda) \propto g_1(t_1;\psi)g_2(t_2 \mid t_1;\lambda),$$

then inference for ψ can be based on the marginal likelihood for ψ based on t_1, and the theory outlined above applies directly. This factorization is fairly special; more common is a factorization of the form $g_1(t_1;\psi)g_2(t_2 \mid t_1;\lambda,\psi)$: in that case to base our inference on the likelihood for ψ from t_1 would require further checking that little information is lost in ignoring the second term. Arguments like these, applied to special classes of model families, were used to derive the modified profile log-likelihood inference outlined above.

A related development is the improvement of the distributional approximation to the approximate pivotal quantity (2.6). The Laplace approximation (3.2) can be used to obtain the Bayesian pivotal, for scalar ψ,

$$r_B^*(\psi) = r(\psi) + \frac{1}{r(\psi)}\log\left\{\frac{q_B(\psi)}{r(\psi)}\right\} \dot{\sim} N(0,1), \tag{3.4}$$

where

$$r(\psi) = \text{sign}(\hat{\psi} - \psi)\big[2\big\{\ell_{\text{p}}(\hat{\psi}) - \ell_{\text{p}}(\psi)\big\}\big]^{1/2}, \tag{3.5}$$

$$q_B(\psi) = -\ell'_{\text{p}}(\psi) j_{\text{p}}^{-1/2}(\hat{\psi}) \left\{ \frac{|j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|}{|j_{\lambda\lambda}(\hat{\psi}, \hat{\lambda})|} \right\}^{1/2} \frac{\pi(\hat{\psi}, \hat{\lambda})}{\pi(\psi, \hat{\lambda}_\psi)} \tag{3.6}$$

and the approximation in (3.4) is to the posterior distribution of r^*, given y, and is accurate to $O(n^{-3/2})$.

There is a frequentist version of this pivotal that has the same form:

$$r_F^*(\psi) = r(\psi) + \frac{1}{r(\psi)} \log\left\{ \frac{q_F(\psi)}{r(\psi)} \right\} \dot{\sim} N(0, 1), \tag{3.7}$$

where $r(\psi)$ is given by (3.5), but the expression for $q_F(\psi)$ requires additional notation, and indeed an additional likelihood component. In the special case of no nuisance parameters

$$q_F(\theta) = \big\{\ell_{;\hat{\theta}}(\hat{\theta}; \hat{\theta}, a) - \ell_{;\hat{\theta}}(\theta; \hat{\theta}, a)\big\} j^{-1/2}(\hat{\theta}; \hat{\theta}, a) \tag{3.8}$$

$$= \big\{\varphi(\hat{\theta}) - \varphi(\theta)\big\} \varphi_\theta^{-1}(\hat{\theta}) j^{1/2}(\hat{\theta}). \tag{3.9}$$

In (3.8), we have assumed that there is a one-to-one transformation from y to $(\hat{\theta}, a)$, and that we can write the log-likelihood function in terms of $\theta, \hat{\theta}, a$ and then differentiate it with respect to $\hat{\theta}$, for fixed a. Expression (3.9) is equivalent, but expresses this sample space differentiation through a data-dependent reparameterization $\varphi(\theta) = \varphi(\theta; y) = \partial\ell(\theta; y)/\partial V(y)$, where the derivative with respect to $V(y)$ is a directional derivative to be determined.

The details are somewhat cumbersome, and even more so for the case of nuisance parameters, but the resulting r_F^* approximate pivotal quantity is readily calculated in a wide range of models for independent observations $y_1, \ldots, y_n$. Detailed accounts are given in Barndorff-Nielsen and Cox (1994), Pace and Salvan (1997), Severini (2000), Fraser, Reid and Wu (1999) and Brazzale, Davison and Reid (2007, Ch. 8.6); the last emphasizes implementation in a number of practical settings, including generalized linear models, nonlinear regression with normal errors, linear regression with non-normal errors, and a number of more specialized models.

From a theoretical point of view, an important distinction between r_B^* and r_F^* is that the latter requires differentiation of the log-likelihood function on the sample space, whereas the former depends only on the observed log-likelihood function, along with the prior. The similarity of the two expressions suggests that it might be possible to develop prior densities for which the posterior probability bounds are guaranteed to be valid under the model, at least to a higher order of approximation than implied by (2.8), and there is a long line of research on the development of these so-called "matching priors"; see, for example, Datta and Mukerjee (2004).

4. Extending the likelihood function

4.1. Introduction

While the asymptotic results of the last section provide very accurate inferences, they are not as straightforward to apply as the first order results, especially in models with complex dependence. They do shed light on many aspects of theory, including the precise points of difference, asymptotically, between Bayesian and nonBayesian inference. And the techniques used to derive them, saddlepoint and Laplace approximations in the main, have found application in complex models in certain settings, such as the integrated nested Laplace approximation of Rue, Martino and Chopin (2009).

A glance at any number of papers motivated by specific applications, though, will confirm that likelihood summaries, and in particular computation of the maximum likelihood estimator, are often the inferential goal, even as the models become increasingly high-dimensional.

This is perhaps a natural consequence of the emphasis on developing probability models that could plausibly generate, or at least describe, the observed responses, as the likelihood function is directly obtained from the probability model. But more than this, inference based on the likelihood function provides a standard set of tools, whose properties are generally well-known, and avoids the construction of *ad hoc* inferential techniques for each new application. For example, Brown et al. (2004) write "The likelihood framework is an efficient way to extract information from a neural spike train. . . We believe that greater use of the likelihood based approaches and goodness-of-fit measures can help improve the quality of neuroscience data analysis".

A number of inference functions based on the likelihood function, or meant to have some of the key properties of the likelihood function, have been developed in the context of particular applications or particular model families. In some cases the goal is to find 'reasonably reliable' estimates of a parameter, along with an estimated standard error; in other cases the goal is to use approximate pivotal quantities like those outlined in Section 2 in settings where the likelihood is difficult to compute. The goal of obtaining reliable likelihood-based inference in the presence of nuisance parameters was addressed in Section 3. In some settings, families of parametric models are too restrictive, and the aim is to obtain likelihood-type results for inference in semi-parametric and non-parametric settings.

4.2. Generalized linear mixed models

In many applications with longitudinal, clustered, or spatial data, the starting point is a generalized linear model with a linear predictor of the form $X\beta + Zu$, where X and Z are $n \times k$ and $n \times q$, respectively, matrices of predictors, and u is a q-vector of random effects. The marginal distribution of the responses requires integrating over the distribution of the random effects u, and this is often computationally infeasible. Many approximations have been suggested: one approach is to approximate the integral by Laplace's method (Breslow and Clayton, 1993), leading to what is commonly called penalized quasi-likelihood, although this is different from the penalized versions of composite likelihood discussed below. The term quasi-likelihood in the context of generalized linear models refers to the specification of the model through the mean

function and variance function only, without specifying a full joint density for the observations. This was first suggested by Wedderburn (1974), and extended to longitudinal data in Liang and Zeger (1986) and later work, leading to the methodology of generalized estimating equations, or GEE. Renard, Molenberghs and Geys (2004) compared penalized quasi-likelihood to pairwise likelihood, discussed in Section 4.3, in simulations of multivariate probit models for binary data with random effects. In general penalized quasi-likelihood led to estimates with larger bias and variance than pairwise likelihood.

A different approach to generalized linear mixed models has been developed by Lee and Nelder; see, for example, Lee and Nelder (1996) and Lee, Nelder and Pawitan (2006), under the name of h-likelihood. This addresses some of the failings of the penalized quasi-likelihood method by modelling the mean parameters and dispersion parameters separately. The h-likelihood for the dispersion parameters is motivated by REML-type arguments not unrelated to the higher order asymptotic theory outlined in the previous section. There are also connections to work on prediction using likelihood methods (Bjørnstad, 1990). Likelihood approaches to prediction have proved to be somewhat elusive, at least in part because the 'parameter' to be predicted is a random variable, although Bayesian approaches are straightforward as no distinction is made between parameters and random variables.

4.3. Composite likelihood

Composite likelihood is one approach to combining the advantages of likelihood with computational feasibility; more precisely it is a collection of approaches. The general principle is to simplify complex dependence relationships by computing marginal or conditional distributions of some subsets of the responses, and multiplying these together to form an inference function.

As an *ad hoc* solution it has emerged in several versions and in several contexts in the statistical literature; an important example is the pseudo-likelihood for spatial processes proposed in Besag (1974, 1975). In studies of large networks, computational complexity can be reduced by ignoring links between distant nodes, effectively treating sub-networks as independent. In Gaussian process models with high-dimensional covariance matrices, assuming sparsity in the covariance matrix is effectively assuming subsets of variables are independent. The term composite likelihood was proposed in Lindsay (1988), where the theoretical properties of composite likelihood estimation were studied in some generality.

We suppose a vector response of length q is modelled by $f(y;\theta), \theta \in \mathbb{R}^d$. Given a set of events $\mathcal{A}_k, k = 1, \ldots, K$, the composite likelihood function is defined as

$$CL(\theta; y) = \prod_{k=1}^{K} f(y \in \mathcal{A}_k; \theta), \tag{4.1}$$

and the composite log-likelihood function is

$$c\ell(\theta; y) = \sum_{k} \log f(y \in \mathcal{A}_k; \theta). \tag{4.2}$$

Because each component in the sum is the log of a density function, the resulting score function $\partial c\ell(\theta; y)/\partial\theta$ has expected value 0, so has at least one of the properties of a genuine log-likelihood function.

Relatively simple and widely used examples of composite likelihoods include independence composite likelihood,

$$c\ell_{\text{ind}}(\theta; y) = \sum_{r=1}^{q} \log f_1(y_r; \theta),$$

pairwise composite likelihood

$$c\ell_{\text{pair}}(\theta; y) = \sum_{r=1}^{q} \sum_{s>r} \log f_2(y_r, y_s; \theta),$$

and pairwise conditional composite likelihood

$$c\ell_{\text{cond}}(\theta; y) = \sum_{r=1}^{q} \log f(y_r \mid y_{(-r)}; \theta), \tag{4.3}$$

where $f_1(y_r; \theta)$ and $f_2(y_r, y_s; \theta)$ are the marginal densities for a single component and a pair of components of the vector observation, and the density in (4.3) is the conditional density of one component, given the remainder.

Many similar types of composite likelihood can be constructed, appropriate to time series, or spatial data, or repeated measures, and so on, and the definition is usually further extended by allowing each component event to have an associated weight w_k. Indeed one of the difficulties of studying the theory of composite likelihood is the generality of the definition.

Inference based on composite likelihood is constructed from analogues to the asymptotic results for genuine likelihood functions. Assuming we have a sample $\underline{y} = (y^{(1)}, \ldots, y^{(n)})$ of independent observations of y, the composite score function,

$$u_{CL}(\theta; \underline{y}) = \sum_{i=1}^{n} \sum_{k} \partial \log f\big(y^{(i)} \in \mathcal{A}_k; \theta\big)/\partial\theta, \tag{4.4}$$

is used as an estimating function to obtain the maximum composite likelihood estimator $\hat{\theta}_{CL}$, and under regularity conditions on the full model, with $n \to \infty$ and fixed K, we have, for example,

$$(\hat{\theta}_{CL} - \theta)^T G(\hat{\theta}_{CL})(\hat{\theta}_{CL} - \theta) \xrightarrow{\mathcal{L}} \chi^2_d, \tag{4.5}$$

where

$$G(\theta) = H(\theta) J^{-1}(\theta) H(\theta) \tag{4.6}$$

is the $d \times d$ Godambe information matrix, and

$$J(\theta) = \text{var}\big\{u_{CL}(\theta; Y)\big\}, \qquad H(\theta) = \text{E}\big\{-(\partial/\partial\theta) u_{CL}(\theta; Y)\big\},$$

are the variability and sensitivity matrix associated with u_{CL}.

The analogue of (2.5) is

$$2\{c\ell(\hat{\theta}_{CL}) - c\ell(\theta)\} \xrightarrow{\mathcal{L}} \sum_{i=1}^{d} \lambda_i \chi^2_{1i}, \tag{4.7}$$

where λ_i are the eigenvalues of $J^{-1}(\theta)H(\theta)$.

Neither of these results is quite as convenient as the full likelihood versions, and in particular contexts it may be difficult to estimate $J(\theta)$ accurately, but there are a number of practical settings where these results are much more easily implemented than the full likelihood results, and the efficiency of the methods can be quite good.

A number of applied contexts are surveyed in Varin, Reid and Firth (2011). As just one example, developed subsequently, Davison, Padoan and Ribatet (2012) investigate pairwise composite likelihood for max-stable processes, developed to model extreme values recorded at a number D of spatially correlated sites. Although the form of the D-dimensional density is known, it is not computable for $D > 3$, although expressions are available for the joint density at each pair of sites. Composite likelihood seems to be particularly important for various types of spatial models, and many variations of it have been suggested for these settings.

In some applications, particularly for time series, but also for space-time data, a sample of independent observations is not available, and the relevant asymptotic theory is for $q \to \infty$, where q is the dimension of the single response. The asymptotic results outlined above will require some conditions on the decay of the dependence among components as the 'distance' between them increases. Asymptotic theory for pairwise likelihood is investigated in Davis and Yau (2011) for linear time series, and in Davis, Klüppelberg and Steinkohl (2012) for max-stable processes in space and time.

Composite likelihood can also be used for model selection, with an expression analogous to *AIC*, and for Bayesian inference, after adjustment to accommodate result (4.7). *Statistica Sinica* **21**, #1 is a special issue devoted to composite likelihood, and more recent research is summarized in the report on a workshop at the Banff International Research Station (Joe, 2012).

4.4. Semi-parametric likelihood

In some applications, a flexible class of models can be constructed in which the nuisance 'parameter' is an unknown function. The most widely-known example is the proportional hazards model of Cox (1972) for censored survival data; but semi-parametric regression models are also widely used, where the particular covariates of interest are modelled with a low-dimensional regression parameter, and other features expected to influence the response are modelled as 'smooth' functions. Cox (1972) developed inference based on a partial likelihood, which ignored the aspects of the likelihood bearing on the timing of failure events, and subsequent theory based on asymptotics for counting processes established the validity of this approach. In fact, Cox (1972)'s partial likelihood can be viewed as an example of composite likelihood as described above, although the theory for general semi-parametric models seems more natural.

Murphy and van der Vaart (2000) showed that partial likelihood can be viewed as a profile likelihood, maximized over the nuisance function, and discussed a class of semi-parametric models for which the profile likelihood continues to have the same asymptotic properties as the usual parametric profile likelihood; the contributions to the discussion of their results provide further insight and references to the extensive literature on semi- and non-parametric likelihoods. There is, however, no guarantee that asymptotic theory will lead to accurate approximation for finite samples; it would presumably have at least the same drawbacks as profile likelihood in the parametric setting. Improvements via modifications to the profile likelihood, as described above in the parametric case, do not seem to be available in these more general settings.

Some semi-parametric models are in effect converted to high-dimensional parametric models through the use of linear combinations of basis functions; thus the linear predictor associated with a component y_i might be $\beta_0 + \beta_1 x_i + \sum_{j=1}^{J} \gamma_j B(z_i)$, or $\beta_0 + \beta_1 x_i + \sum_{j=1}^{J} \gamma_{1j} B_j(z_{1i}) + \cdots + \sum_{j=1}^{J} \gamma_{kj} B_j(z_{ki})$. The log-likelihood function for models such as these is often regularized, so that $\ell(\beta, \gamma)$ is replaced by $\ell(\beta, \gamma) + \lambda p(\gamma)$, where $p(\cdot)$ is a penalty function such as $\Sigma \gamma_{kj}^2$ or $\Sigma |\gamma_{kj}|$, and λ a tuning parameter. Many of these extensions, and the asymptotic theory associated with them, are discussed in van der Vaart (1998, Ch. 25). Penalized likelihood using squared error is reviewed in Green (1987); the L_1 penalty has been suggested as a means of combining likelihood inference with variable selection; see, for example, Fan and Li (2001).

Penalized composite likelihoods have been proposed for applications in spatial analysis (Divino, Frigessi and Green, 2000; Apanasovich et al., 2008; Xue, Zou and Cai, 2012), Gaussian graphical models (Gao and Massam, 2012), and clustered longitudinal data (Gao and Song, 2010).

The difference between semi-parametric likelihoods and nonparametric likelihoods is somewhat blurred; both have an effectively infinite-dimensional parameter space, and as discussed in Murphy and van der Vaart (2000) and the discussion, conditions on the model to ensure that likelihood-type asymptotics still hold can be quite technical.

Empirical likelihood is a rather different approach to non-parametric models first proposed by Owen (2001); a recent discussion is Hjort, McKeague and Van Keilegom (2009). Empirical likelihood assumes the existence of a finite-dimensional parameter of interest, defined as a functional of the distribution function for the data, and constructs a profile likelihood by maximizing the joint probability of the data, under the constraint that this parameter is fixed. This construction is particularly natural in survey sampling, where the parameter is often a property of the population (Chen and Sitter, 1999; Wu and Rao, 2006). Distribution theory for empirical likelihood more closely follows that for usual parametric likelihoods.

4.5. Simulation methods

Simulation of the posterior density by Markov chain Monte Carlo methods is widely used for Bayesian inference, and there is an enormous literature on various methods and their properties. Some of these methods can be adapted for use when the likelihood function itself cannot be computed, but it is possible to simulate observations from the stochastic model; many examples arise in statistical genetics. Simulation methods for maximum likelihood estimation in genetics was proposed in Geyer and Thompson (1992); more recently sequential Monte Carlo

methods (see, for example, Sisson, Fan and Tanaka (2007)) and ABC (approximate Bayesian computation) methods (Fearnhead and Prangle, 2012; Marin et al., 2011) are being investigated as computational tools.

5. Conclusion

A reviewer of an earlier draft suggested that a great many applications, especially involving very large and/or complex datasets, take more algorithmic approaches, often using techniques designed to develop sparse solutions, such as wavelet or thresholding techniques, and that likelihood methods may not be relevant for these application areas.

Certainly a likelihood-based approach depends on a statistical model for the data, and for many applications under the general rubric of machine learning these may not be as important as developing fast and reliable approaches to prediction; recommender systems are one such example.

There are however many applications of 'big data' methods where statistical models do provide some structure, and in these settings, as in the more classical application areas, likelihood methods provide a unifying basis for inference.

Acknowledgements

This research was partially supported by the Natural Sciences and Engineering Research Council. Thanks are due to two reviewers for helpful comments on an earlier version.

References

Apanasovich, T.V., Ruppert, D., Lupton, J.R., Popovic, N., Turner, N.D., Chapkin, R.S. and Carroll, R.J. (2008). Aberrant crypt foci and semiparametric modeling of correlated binary data. *Biometrics* **64** 490–500, 667. MR2432419

Barndorff-Nielsen, O.E. and Cox, D.R. (1994). *Inference and Asymptotics*. *Monographs on Statistics and Applied Probability* **52**. London: Chapman & Hall. MR1317097

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **36** 192–236. With discussion by D.R. Cox, A.G. Hawkes, P. Clifford, P. Whittle, K. Ord, R. Mead, J.M. Hammersley, and M.S. Bartlett and with a reply by the author. MR0373208

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

Bjørnstad, J.F. (1990). Predictive likelihood: A review. *Statist. Sci.* **5** 242–265. With comments and a rejoinder by the author. MR1062578

Brazzale, A.R., Davison, A.C. and Reid, N. (2007). *Applied Asymptotics: Case Studies in Small-Sample Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **23**. Cambridge: Cambridge Univ. Press. MR2342742

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalised linear models. *J. Amer. Statist. Assoc.* **88** 9–25.

Brown, E.N., Barbieri, R., Eden, U.T. and Frank, L.M. (2004). Likelihood methods for neural spike train data analysis. In *Computational Neuroscience*. *Chapman & Hall/CRC Math. Biol. Med. Ser.* 253–286. Boca Raton, FL: Chapman & Hall/CRC. MR2029664

Chen, J. and Sitter, R.R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statist. Sinica* **9** 385–406. MR1707846

Cox, D.R. (1972). Regression models and life-tables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **34** 187–220. With discussion by F. Downton, Richard Peto, D.J. Bartholomew, D.V. Lindley, P.W. Glassborow, D.E. Barton, Susannah Howard, B. Benjamin, John J. Gart, L.D. Meshalkin, A.R. Kagan, M. Zelen, R.E. Barlow, Jack Kalbfleisch, R.L. Prentice and Norman Breslow, and a reply by D.R. Cox. MR0341758

Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics*. London: Chapman & Hall. MR0370837

Cox, D.R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **49** 1–39. With a discussion. MR0893334

Datta, G.S. and Mukerjee, R. (2004). *Probability Matching Priors*: *Higher Order Asymptotics*. *Lecture Notes in Statistics* **178**. New York: Springer. MR2053794

Davis, R.A., Klüppelberg, C. and Steinkohl, C. (2012). Statistical inference for max-stable processes in space and time. Preprint, available at arXiv:1204.5581v1, accessed on August 6, 2012.

Davis, R.A. and Yau, C.Y. (2011). Comments on pairwise likelihood in time series models. *Statist. Sinica* **21** 255–277. MR2796862

Davison, A.C. (2003). *Statistical Models*. *Cambridge Series in Statistical and Probabilistic Mathematics* **11**. Cambridge: Cambridge Univ. Press. MR1998913

Davison, A.C., Padoan, S.A. and Ribatet, M. (2012). Statistical modelling of spatial extremes. *Statist. Sci.* **27** 161–186.

DiCiccio, T.J. and Martin, M.A. (1993). Simple modifications for signed roots of likelihood ratio statistics. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **55** 305–316. MR1210437

Divino, F., Frigessi, A. and Green, P.J. (2000). Penalized pseudolikelihood inference in spatial interaction models with covariates. *Scand. J. Statist.* **27** 445–458.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **74** 419–474. MR2925370

Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc. A* **222**, 309–368.

Fisher, R.A. (1956). *Statistical Methods and Scientific Inference*. Edinburgh: Oliver and Boyd. Third edition 1973. MR0346955

Fraser, D.A.S., Reid, N. and Wu, J. (1999). A simple general formula for tail probabilities for frequentist and Bayesian inference. *Biometrika* **86** 249–264. MR1705367

Gao, X. and Massam, H. (2012). Composite likelihood estimation of high dimensional Gaussian graphical models with symmetry. Presented at BIRS Workshop on Composite Likelihood, April, 2012 (see Joe (2012)).

Gao, X. and Song, P.X.K. (2010). Composite likelihood Bayesian information criteria for model selection in high-dimensional data. *J. Amer. Statist. Assoc.* **105** 1531–1540. Supplementary materials available online. MR2796569

Geyer, C.J. and Thompson, E.A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **54** 657–699. With discussion and a reply by the authors. MR1185217

Green, P.J. (1987). Penalized likelihood for general semi-parametric regression models. *Internat. Statist. Rev.* **55** 245–259. MR0963142

Hjort, N.L., McKeague, I.W. and Van Keilegom, I. (2009). Extending the scope of empirical likelihood. *Ann. Statist.* **37** 1079–1111. MR2509068

Joe, H. (2012). Report on the Workshop on Composite Likelihood. Available at http://www.birs.ca/events/2012/5-day-workshops/12w5046.

Lee, Y. and Nelder, J.A. (1996). Hierarchical generalized linear models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58** 619–678. With discussion. MR1410182

Lee, Y., Nelder, J.A. and Pawitan, Y. (2006). *Generalized Linear Models with Random Effects*: *Unified Analysis via H-Likelihood*. *Monographs on Statistics and Applied Probability* **106**. Boca Raton, FL: Chapman & Hall/CRC. MR2259540

Liang, K.Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22. MR0836430

Lindsay, B.G. (1988). Composite likelihood methods. In *Statistical Inference from Stochastic Processes* (*Ithaca*, *NY*, 1987). *Contemp. Math.* **80** 221–239. Providence, RI: Amer. Math. Soc. MR0999014

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

Murphy, S.A. and van der Vaart, A.W. (2000). On profile likelihood. *J. Amer. Statist. Assoc.* **95** 449–485. With comments and a rejoinder by the authors. MR1803168

Owen, A. (2001). *Empirical Likelihood*. London: Chapman & Hall/CRC.

Pace, L. and Salvan, A. (1997). *Principles of Statistical Inference From a Neo-Fisherian Perspective*. *Advanced Series on Statistical Science & Applied Probability* **4**. River Edge, NJ: World Scientific. MR1476674

Pace, L. and Salvan, A. (2006). Adjustments of the profile likelihood from a new perspective. *J. Statist. Plann. Inference* **136** 3554–3564. MR2256276

Renard, D., Molenberghs, G. and Geys, H. (2004). A pairwise likelihood approach to estimation in multilevel probit models. *Comput. Statist. Data Anal.* **44** 649–667. MR2026438

Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **71** 319–392. MR2649602

Severini, T.A. (2000). *Likelihood Methods in Statistics*. *Oxford Statistical Science Series* **22**. Oxford: Oxford Univ. Press. MR1854870

Sisson, S.A., Fan, Y. and Tanaka, M.M. (2007). Sequential Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA* **104** 1760–1765 (electronic). MR2301870

Tierney, L. and Kadane, J.B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.* **81** 82–86. MR0830567

van der Vaart, A.W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge: Cambridge Univ. Press. MR1652247

Varin, C., Reid, N. and Firth, D. (2011). An overview of composite likelihood methods. *Statist. Sinica* **21** 5–42. MR2796852

Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **61** 439–447. MR0375592

Wu, C. and Rao, J.N.K. (2006). Pseudo-empirical likelihood ratio confidence intervals for complex surveys. *Canad. J. Statist.* **34** 359–375. MR2328549

Xue, L., Zou, H. and Cai, T. (2012). Non-concave penalized composite conditional likelihood estimation of sparse Ising models. *Ann. Statist.* **40** 1403–1429.