Approximate likelihoods

N. Reid*

Abstract. In complex models likelihood functions may be difficult to compute, or depend on assumptions about high order dependencies that may be difficult to verify. A number of methods have been devised to compute inference functions either meant to approximate the true likelihood function, or to provide inferential summaries that balance statistical efficiency with ease of computation. Examples include variational approximations, composite likelihood, quasi-likelihood, indirect inference, and Laplace-type approximations. This paper surveys various approximations to likelihood and likelihood inference, with a view to identifying common themes and outstanding problems.

Mathematics Subject Classification (2000). Primary 62F99; Secondary 62E20.

Keywords. Asymptotic theory, composite likelihood, indirect inference, quasi-likelihood, simulation.

1. Introduction

Statistical inference is broadly concerned with the problem of learning from data; in particular, data obtained from a system subject to random fluctuations. This problem is made somewhat more concrete by considering a family of probability models, which it is hoped describe at least some of the essential features of the system. The question then becomes how to reason inductively, using an observed set, or sets, of data to draw conclusions about the probability model, or key features of the probability model, in order to advance understanding of the system. The theory of statistical inference studies the mathematical aspects of this inductive reasoning to develop strategies, and to ensure that these strategies make efficient use of the data at hand, give answers that with high probability are not likely to mislead, and provide a means of attaching an estimate of the uncertainty of the answers.

Probability models are defined by the probability distribution attached to various sets in the space on which they are defined, and in statistical inference problems are very often specified in a relatively small class of probability distributions parameterized by one or more unknown constants. The likelihood function is a function of these unknown parameters, and for a given set of data depends only on these parameters and the observed data. Inference based on the likelihood function

*This research is partially supported by the Natural Sciences and Engineering Research Council of Canada and the Canada Research Chairs Program.

Proceedings of the ICIAM,
Beijing, China, 2015

is very widely used in applications, and plays a key role in statistical theory.

As scientific problems and technological advances provide data of ever-increasing size and complexity of structure, probability models also become increasingly high-dimensional and complex. This has meant in practice that it is often not feasible to base inference on the likelihood function, and a wide number of simplifications, both computational and conceptual, have been proposed. In this paper we review several of these approaches.

In the next section we give a brief overview of likelihood inference to set notation, introduce some examples from the literature, and outline the key ideas in the theory of likelihood inference. In later sections we consider a variety of approximate methods based on likelihood inference, emphasizing the connections between and among them, and highlighting areas where further research is needed.

2. Models and likelihood

2.1. The likelihood function. We assume our statistical model includes specification of an observable random vector $Y = (Y_1, \ldots, Y_n)$ and a family of probability models for Y indexed by a parameter θ taking values in some parameter space Θ. We further assume that this family of probability models can be represented by a density function $f(y;\theta)$ with respect to some dominating measure, typically Lebesgue or counting measure. We will generally restrict attention to models in which the parameter space Θ is finite-dimensional, and indeed a subset of $\mathbb{R}^d$. In more general, non-parametric, models Θ is an infinite-dimensional space, for example the space of "all smooth functions". These models are more difficult to analyze from a likelihood point of view; a brief discussion is given in §5. Particularly relevant for complex data is an asymptotic regime in which both d and n go to infinity, although in most of our discussion we will assume d is fixed and $n \to \infty$.

Given an observed value $y = (y_1, \ldots, y_n)$ from a model with density $f(y;\theta)$ the likelihood function for θ is defined as the equivalence class of functions of θ:

$$L(\theta; y) = c(y) f(y;\theta), \tag{1}$$

with $c(\cdot)$ an arbitrary function (Fisher, 1922). It is often convenient to work with the log-likelihood function, or equivalence class of functions,

$$\ell(\theta; y) = \log L(\theta; y) + a(y). \tag{2}$$

The value of the likelihood function at any given value θ_0 can only be assessed relative to other values of θ, and it is often convenient to standardize the likelihood by its maximum value $L(\hat\theta; y)$, where $\hat\theta = \hat\theta(y) = \arg\sup_\theta L(\theta; y)$. Fisher (1956, Ch. III) suggested using likelihood ratios relative to the maximum to declare regions of the parameter space 'very plausible', 'somewhat plausible, and 'highly implausible' according to somewhat arbitrary cut-off values for the ratio of 3 and 10. This direct use of the likelihood function for inference has been developed

much further by Royall and others; see for example Royal (1997), or Burnham & Anderson (2002).

Likelihood methods emphasize the role of probability modelling, and in particular parametric models, in the study of random phenomena. The notation used in (1) emphasizes the inductive problem of reasoning from y back to θ, and thus identifying which values of θ, or which members of the family $\{f(y;\theta);\theta \in \Theta\}$ are consistent with the observed data y. In a Bayesian framework, where the modelling includes specification of a prior density $\pi(\theta)$, the likelihood function converts a prior density $\pi(\theta)$ to a posterior density $\pi(\theta \mid y)$ via Bayes' theorem. The widespread use of likelihood in applications is possibly due to the fact that it enables relatively straightforward computation of a set of summary statistics whose probability distributions can be analysed, so in this way provides a common inferential framework independent of particular applications.

Example 1: Generalized Linear Models. Suppose $y = (y_1, \ldots, y_n)$ are independent observations with density

$$f(y_i \mid x_i; \beta, \phi) = \exp[\{y_i\theta_i - b(\theta_i)\}/(\phi a_i) + c(y_i, \phi)], \tag{3}$$

for suitably defined functions $b(\cdot)$ and $c(\cdot,\cdot)$, where a_i are assumed known, and θ_i is related to β via a linear regression: $\theta_i = x_i^{\mathrm{T}}\beta$, or in some cases $h(\theta_i) = x_i^{\mathrm{T}}\beta$. The parameter θ of (2) is here (β, ϕ). In (3) x_i is a known d-vector of explanatory variables associated with the ith observation.

This is called a generalized linear model, and widely used in applied work to model data from the binomial, Poisson, normal, and gamma distributions. The log-likelihood function is

$$\ell(\beta, \phi; y) = \sum_{i=1}^{n} [\{y_i x_i^{\mathrm{T}}\beta - b(\theta_i)\}/(\phi a_i) + c(y_i, \phi)], \tag{4}$$

and typically is easy to optimize over β by using an iterative root finder for the *score equation* $\partial\ell(\beta, \phi; y)/\partial\beta = 0$. The special nature of the dependence on ϕ means the maximum likelihood estimator of β is the same for any ϕ.

Instead of a single response for each i, we may be interested in a series of responses, as arise for example in measuring a set of subjects at a number of different time points. Then $y_i = (y_{i1}, \ldots, y_{im_i})$ is a vector of responses on the same subject and these responses may be correlated. One way to model this correlation is to posit an unobserved latent random variable u_i for each subject. If the density of y_{ij}, conditional on u_i, follows a generalized linear model,

$$f(y_{ij} \mid u_i; \beta, \phi) = \exp[\{y_{ij}\theta_{ij} - b(\theta_{ij})\}/(\phi a_i) + c(y_{ij}, \phi)], \tag{5}$$

and we assume

$$\theta_{ij} = x_{ij}^{\mathrm{T}}\beta + z_{ij}^{\mathrm{T}}u_i, \text{ or } h(\theta_{ij}) = x_{ij}^{\mathrm{T}}\beta + z_{ij}^{\mathrm{T}}u_i,$$

where x_{ij} and z_{ij} are vectors of explanatory variables, then to compute the likelihood for the observed data we need to integrate out the random effects u_i over their

distribution. For example if the model for u_i is a k-dimensional normal distribution with mean 0 and covariance matrix Σ, the log-likelihood function is

$$\ell(\beta,\Sigma,\phi;y)=\sum_{i=1}^{n}\left(\frac{y_i^{\mathrm{T}}X_i\beta}{a_i\phi}-\frac{1}{2}\log|\Sigma| \right. \\ \left. +\log\int_{\mathbb{R}^k}\exp\left\{\frac{y_i^{\mathrm{T}}Z_iu_i-1_i^{\mathrm{T}}b(X_i\beta+Z_iu_i)}{a_i\phi}\ -\frac{1}{2}u_i^{\mathrm{T}}\Sigma^{-1}u_i\right\}du_i\right), \quad (6)$$

where for the ith observation, X_i is the $m_i\times d$ matrix of explanatory variables x for the fixed effects, and Z_i the $m_i\times k$ matrix of explanatory variables for the random effects.

A version of this model, with $a_i\phi=1$, is treated in Ormerod & Wand (2010) to illustrate the variational approximation discussed in §3.2 below. By considering observations with a more complex structure than simply independent, identically distributed, the log-likelihood function becomes correspondingly more complex: (6) requires k-dimensional integration for each value β and this may be prohibitive for standard likelihood computations.

Example 2: Poisson auto-regression. A similar approach can be used to model a single time series of observations from a Poisson distribution: suppose y_i takes values on the non-negative integers, with density function

$$f(y_i\mid\alpha_i;\theta)=\exp(y_i\log\mu_i-\mu_i)/y_i!,\quad \log\mu_i=\beta+\alpha_i. \quad (7)$$

We can model an autoregressive behaviour for an observed time series $y_1,\ldots,y_n$ by modelling the unobserved α as

$$\alpha_i=\rho\alpha_{i-1}+\epsilon_i,\quad \epsilon_i\sim_{\mathrm{ind}}N(0,\sigma^2),\ |\rho|<1,$$

leading to the likelihood function

$$L(\theta;y_1,\ldots,y_n)=\int\left(\prod_{i=1}^{n}f(y_i\mid\alpha_i;\theta)\right)f(\alpha;\theta)d\alpha, \quad (8)$$

where $\theta=(\beta,\rho,\sigma^2)$. This example was discussed in Davis & Yau (2011), who studied various versions of composite likelihood in time series settings; see §3.1.

Example 3: Multivariate extremes: Suppose we have n observations of, for example, windspeed, at each of D locations in space. In a study of extreme values, we may be interested in the joint distribution of the maximum windspeed at these D locations. Davison et al. (2012) show that after suitable centering and scaling, these component wise maxima $Z_1,\ldots,Z_D$, say, follow the multivariate extreme value distribution:

$$\Pr(Z_1\le z_1,\ldots,Z_D\le z_D)=\exp\{-V(z_1,\ldots,z_D)\}, \quad (9)$$

where $V(\cdot)$ is the so-called spectral measure. For $D=2$ there are parametrized versions proposed in the literature: see Davison et al. (2012) and Davison & Huser

(2015); for $D > 2$ the situation is much more complex. Computing the likelihood function is conceptually straightforward, as the density of Z is given by the D-dimensional derivative of (9), but for moderately large D this is in fact not computable, and the lack of families of parametrized models for $D > 2$ complicates the situation. Davison & Huser (2015) use pairwise likelihood approximations, as described in §3.1.

Example 4: $M/G/1$ queue: Suppose we observe the time of departure y_i of customer i, from a queue with a single server. The arrival times V_i are assumed to follow a dependent exponential process: $V_i = V_{i-1} + E_i$, where E_i are independent exponential random variables with mean θ_3, and $V_0 \equiv 0$. The distribution of y_i depends on the arrival time and the number of customers in the queue. Assuming the V_i are unobserved, the likelihood function is an integral over the distribution of these:

$$L(\theta; y) = \int \cdots \int f(v_1 \mid \theta) \prod_{i=2}^{n} f(v_i \mid v_{i-1}; \theta) \prod_{i=1}^{n} f(y_i \mid v_i, x_{i-1}; \theta) dv_1 \cdots dv_n,$$

where $x_i = \Sigma_{j=1}^{i} y_j$. This model, with the distribution of Y_i, given X_i, V_i assumed to be uniformly distributed over an unknown interval, was used in Heggland & Frigessi (2004) in a discussion of indirect inference, and in Fearnhead & Prangle (2010) to illustrate the use of approximate Bayesian computation; see §4.2. Shestopaloff & Neal (2013) consider calculation of the likelihood and posteriors based on this likelihood using Markov chain Monte Carlo methods.

Example 5: Ising Model: The Ising model is a Markov random field for a binary vector $y = (y_1, \ldots, y_K)$, with $y_i = \pm 1$, where y_i records a binary property of node i in a graph with vertex set $V = \{1, \ldots, K\}$ and edge set $E \subset V \times V$. The density is

$$f(y; \theta) = \exp\Big(\sum_{(i,j) \in E} \theta_{ij} y_j y_j \Big) / Z(\theta),$$

and the parameter θ_{ij} measures the strength of the interaction between nodes i and j. The partition function $Z(\theta)$ is the normalizing constant for this density, in principle determined by summing over all binary vectors y, but in practice typically not computable. Estimation of sparse Ising models using an approximate likelihood and regularization is treated in Ravikumar et al. (2010) and Xue et al. (2012); see §3.1.

The restricted Boltzmann machine is a Markov random field for an observed output $y = (y_1, \ldots, y_n)$ that depends on some number, usually large, of hidden units h, with likelihood function

$$L(\theta; y) = \frac{1}{Z(\theta)} \sum_{h} \exp(h^{\mathrm{T}} W y + \alpha^{\mathrm{T}} h + \beta^{\mathrm{T}} y), \quad \theta = (W, \alpha, \beta). \tag{10}$$

As in the Ising model, the partition function $Z(\theta) = \Sigma_{y,h} \exp(h^{\mathrm{T}} W y + \alpha^{\mathrm{T}} h + \beta^{\mathrm{T}} y)$ is usually not computable, and with many hidden units the summation in the numerator is also prohibitively complex.

2.2. Likelihood Inference. In settings where we can compute the likelihood function, and various summary quantities based on the likelihood function, it is of interest to study the distribution of these quantities from the model $f(y;\theta_0) = f(y_1,\ldots,y_n;\theta)$, where θ_0 is the notional 'true' value of θ that generated the data, as $n \to \infty$. To study this we define the *score function* $u(\theta;y) = \partial\ell(\theta;y)/\partial\theta$, and the observed and expected Fisher information functions:

$$j(\theta;y) = -\frac{\partial^2\ell(\theta;y)}{\partial\theta\partial\theta^{\mathrm{T}}}, \quad i(\theta) = E\{-\frac{\partial^2\ell(\theta;y)}{\partial\theta\partial\theta^{\mathrm{T}}}\}. \tag{11}$$

The maximum likelihood estimator is often obtained by solving the score equation

$$u(\theta;y) = 0; \tag{12}$$

if this equation has more than one solution then more detailed assessment of the model is needed, but in regular models the solution of this equation will determine the maximum likelihood estimate, or at least determine a consistent sequence of parameter estimates. When the components of y are independent, the log-likelihood function at any fixed value of θ is a sum of independent random variables, as is $u(\theta;y)$, and under some conditions on the model the central limit theorem for $u(\theta;y)$ leads to the following convergence in distribution results, as $n \to \infty$:

$$i^{-1/2}(\theta)u(\theta) \quad \stackrel{\mathcal{D}}{\to} \quad N(0,I), \tag{13}$$

$$i^{1/2}(\theta)(\hat\theta - \theta) \quad \stackrel{\mathcal{D}}{\to} \quad N(0,I), \tag{14}$$

$$2\{\ell(\hat\theta) - \ell(\theta)\} \quad \stackrel{\mathcal{D}}{\to} \quad \chi^2_d, \tag{15}$$

where I is the $d \times d$ identify matrix, and 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 $f(y;\theta_0)$. For practical use it is convenient to define the approximate *pivotal quantities*

$$s(\theta) = j^{-1/2}(\hat\theta)u(\theta), \tag{16}$$

$$q(\theta) = j^{1/2}(\hat\theta)(\hat\theta - \theta), \tag{17}$$

$$r(\theta) = \mathrm{sign}(\hat\theta - \theta)[2\{\ell(\hat\theta) - \ell(\theta)\}]^{1/2}, \tag{18}$$

$$w(\theta) = 2\{\ell(\hat\theta) - \ell(\theta)\} \stackrel{\cdot}{\sim} \chi^2_d, \tag{19}$$

the distribution of each of the first three being approximately standard normal, but $r(\theta)$ defined only when $d = 1$. We could, for example, plot $\Phi\{q(\theta)\}$ as a function of $\theta \in \mathbb{R}$, where $\Phi(\cdot)$ is the standard normal distribution function, thus obtaining approximate p-values for testing any value of θ for fixed y. The approach to inference based on these pivotal quantities avoids the somewhat artificial distinction between point estimation and hypothesis testing.

When $\theta \in \mathbb{R}^d$, it is often useful to separate parameters of interest ψ, from so-called nuisance parameters λ, and analogous versions of the above limiting results

in this setting lead to the approximate pivotal quantities

$$s(\psi) = j_{\mathrm{p}}^{-1/2}(\hat{\psi})\ell'_{\mathrm{p}}(\psi), \tag{20}$$

$$q(\psi) = j_{\mathrm{p}}^{1/2}(\hat{\psi})(\hat{\psi} - \psi), \tag{21}$$

$$w(\psi) = 2\{\ell_{\mathrm{p}}(\hat{\psi}) - \ell_{\mathrm{p}}(\psi)\}, \tag{22}$$

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, 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 pivotal quantity has an approximate $\chi^2_{d_1}$ distribution, where d_1 is the dimension of ψ, and (22) 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 & Hinkley, 1974, Ch. 9.4).

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

$$AIC = -2\ell(\hat{\theta}) + 2d, \tag{23}$$

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.

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

$$\pi(\theta \mid y) = \frac{\exp\{\ell(\theta; y)\}\pi(\theta)}{\int \exp\{\ell(\phi; y)\}\pi(\phi)d\phi}. \tag{24}$$

The posterior density for inference about a parameter of interest ψ is

$$\pi_m(\psi \mid y) = \frac{\int \exp\{\ell(\psi, \lambda; y)\}\pi(\psi, \lambda)d\lambda}{\int \exp\{\ell(\phi; y)\}\pi(\phi)d\phi}; \tag{25}$$

in principle ψ need not be a component of the parameter vector in the model, but can be any function of θ. Bayesian inference is conceptually straightforward, given a prior density, although in high-dimensional models the integral must be approximated. Two very useful methods of approximation to the posterior density are Laplace approximation, and Markov chain Monte Carlo simulation from the posterior; the latter in particular has enabled the application of Bayesian inference in models of considerable complexity. Difficulties with Bayesian inference include the specification of a prior density, and the meaning of probabilities for parameters of a mathematical model.

The limiting results in (13), (14) and (15) provide finite-sample approximations that are routinely used for inference about θ or ψ. An analogous limiting normality result for the posterior density shows that for a fixed prior density, as $n \to \infty$, the methods lead to the same limit theory; i.e. the effect of the prior is "washed out" by increasing amounts of data. To study the divergence of Bayesian and likelihood-based inference it is necessary to consider the next terms in the relevant asymptotic expansions. The simplest of these is the Laplace approximation to the posterior marginal density; there are analogous higher-order approximations to s, q, and w, reviewed for example in Brazzale et al. (2007).

In some treatments of the theory of inference there is emphasis on optimality properties of the relevant procedures, for fixed sample size if possible, but more usually in some appropriate asymptotic sense. From this point of view, the maximum likelihood estimator is "best asymptotically normal", in the sense that it is consistent for θ_0, and its asymptotic variance, the inverse Fisher information, is the smallest possible.

The Neyman factorization criterion, and more abstract treatments (Barndorff-Nielsen et al., 1976; Fraser & Naderi 2007), establish that the likelihood map is a minimal sufficient statistic, and thus captures all the information in the data about the parameter. While this doesn't imply that using approximate pivotal quantities such as $q(\theta)$ or $w(\theta)$ are the optimal use of the likelihood function, their generality and simplicity mean they are extremely useful in applications.

The examples in §2.1 use just moderately complex models, and modern applications such as the study of social networks, of disease processes with space-time dependence, of the classification of images, and so on are very widespread and often extremely complex. There is an urgent need to develop a set of inferential tools that form the common starting point analogous to those provided by the classical results. As a result there are many approximate likelihoods, or inference functions, being developed for more complex situations. In the next sections we consider some of the approximate likelihoods under active development, using our examples as illustrations.

3. Simplified Likelihoods

3.1. Composite Likelihood. One method of constructing inference functions based on a probability model that has proved to be both useful and flexible

is the method of *composite likelihood*, also referred to as *pseudo-likelihoood*. A composite likelihood is typically formed by ignoring some of the complex dependencies in the full joint model. For example, we might replace the joint density of a k-dimensional vector, $f(y_{i1}, \ldots, y_{ik}; \theta)$ by a product of pairwise joint densities $f_2(y_{ij}, y_{ik}; \theta)$, or a product of conditional densities $f(y_{ij} \mid y_{i(j)}; \theta)$. Depending on the setting either of these might offer considerable computational simplification.

With an independent sample of such k-vectors, the composite pairwise likelihood is

$$CL_{\text{pair}}(\theta; y) \propto \prod_{i=1}^{n} \prod_{j<k} f_2(y_{ij}, y_{ik}; \theta),$$

and a similar pairwise conditional likelihood could be defined. Often the composite likelihood function is described more generally via a set of subsets $\{\mathcal{A}_k\}$; with a slight abuse of notation:

$$CL(\theta; y) \propto \prod_{i=1}^{n} \prod_{k} f(y_i \in \mathcal{A}_k; \theta).$$

More flexibility is incorporated by permitting the components to be weighted:

$$CL(\theta; y) \propto \prod_{i=1}^{n} \prod_{k} \{f(y_i \in \mathcal{A}_k); \theta\}^{w_k}. \tag{26}$$

The simplest possible composite likelihood is to treat components as completely independent, leading to

$$CL_{\text{ind}}(\theta; y) \propto \prod_{i=1}^{n} \prod_{j=1}^{k} f_1(y_{ij}; \theta); \tag{27}$$

while this is often too simple for statistical applications where some covariance between components is key, it is sometimes useful in other contexts.

Composite likelihood was first introduced as pseudo-likelihood by Besag (1975) in the context of models for spatially correlated data on a grid; Besag's pseudo-likelihood considered the conditional distribution at each point in space, conditioned on the nearest neighbours. Variants of this have been widely used in spatial data modelling. Lindsay (1988) studied the general theoretical properties of (26), and coined the name composite likelihood to describe such objects. A key feature of composite likelihood is that the related composite score equation,

$$\frac{\partial}{\partial \theta} \log CL(\theta; y) \equiv \frac{\partial}{\partial \theta} \text{c}\ell(\theta; y) = 0, \tag{28}$$

is an unbiased[1] estimating equation. Under some regularity conditions this ensures that the resulting point estimator, $\hat{\theta}_{CL}$ is consistent for θ. It is not asymptotically

[1] meaning its expected value in the true model is 0

efficient, since the model is misspecified, but an expression for its asymptotic variance is obtained by Taylor series expansion of the score equation. The asymptotic variance is of the typical 'sandwich' form that arises in the theory of estimating equations. It depends on the quantities

$$H(\theta) = -\mathrm{E}\{\frac{\partial^2 \mathrm{c}\ell(\theta;y)}{\partial\theta\partial\theta^{\mathrm{T}}}\} \quad \text{and} \quad J(\theta) = \mathrm{E}\{\frac{\partial \mathrm{c}\ell(\theta;y)}{\partial\theta}\frac{\partial \mathrm{c}\ell(\theta;y)}{\partial\theta^{\mathrm{T}}}\}, \tag{29}$$

and a typical result, analogous to (14), is

$$\sqrt{n}(\hat{\theta}_{CL} - \theta) \xrightarrow{\mathcal{D}} N\{0, G^{-1}(\theta)\}, \quad G(\theta) = H(\theta)J^{-1}(\theta)H(\theta);$$

the function $G(\theta)$ is often called the Godambe information (Godambe, 1960).

This asymptotic theory requires smoothness on the underlying density, and assumes that the dimension of the parameter space stays fixed as the sample size increases. It also requires some conditions ensuring that the component densities can identify the parameter θ that is assumed to govern the true full model. It is not too difficult to construct examples where specification of the component densities does not generate a family of multivariate densities compatible with these. For composite conditional likelihood the Hammersley-Clifford theorem can be invoked to establish this consistency, but for composite marginal likelihood this seems to need to be checked on a case by case basis.

A likelihood-ratio type result for inference, analogous to (15), is

$$2\{\mathrm{c}\ell(\hat{\theta}_{CL};y) - \mathrm{c}\ell(\theta;y)\} \xrightarrow{\mathcal{D}} \sum_{j=1}^{d} \lambda_j \chi^2_{1j},$$

where χ^2_{1j} are independent random variables each following a chi-squared distribution with 1 degree of freedom and λ_j are the eigenvalues of $J^{-1}(\theta)H(\theta)$. If the composite log-likelihood was a genuine log-likelihood function, then $H(\theta)$ and $J(\theta)$ would be equal, and equal to the expected Fisher information function $i(\theta)$.

Analogous results are available for the case where $\theta = (\psi, \lambda)$ with λ treated as a nuisance parameter: see for example Varin et al. (2011, §2).

Example 2 (cont.):

The joint likelihood based on a sample of size n from the Poisson autoregressive model involves integrating out the latent auto-regressive process as given in (8) above. A composite likelihood function well-suited to the autoregressive setting is

$$CL(\theta; y_1, \ldots, y_n) = \prod_{i=1}^{n-1} \int\int f(y_i \mid \alpha_i; \theta) f(y_{i+1} \mid \alpha_{i+1}; \theta) d\alpha_i d\alpha_{i+1};$$

note that in this time series regime we have a single dependent vector of increasing length, so special arguments will be needed to ensure that the maximum composite likelihood estimator is consistent. Davis & Yau (2011) study this *consecutive pairs* likelihood for several models, including this Poisson auto-regression. They show for

the current example $\hat{\theta}_{CL}$ is consistent, asymptotically normal, and its asymptotic variance is estimable. In their simulations the mean-squared error of $\hat{\theta}_{CL}$ was competitive with alternate approximations to the likelihood based on numerical integration.

Example 3 (cont.):

Pairwise composite likelihood is widely used in multivariate extreme values, and is illustrated in Davison et al. (2012) and Davison & Huser (2015) on a number of different models, each using a different process representation $V(\cdot)$ in (9). One feature of their approach is to use a model selection criterion based on composite likelihood; this is constructed similarly to the model-selection criterion (23) and is defined as

$$CLIC = -2\mathrm{c}\ell(\hat{\theta}_{CL}) + 2\mathrm{trace}(\hat{H}^{-1}\hat{J}), \tag{30}$$

where $\hat{H}$ and $\hat{J}$ are empirical estimates of the information quantities defined in (29) and (35). Davison et al. (2012) use this measure to choose between various models for the joint distribution of annual rainfall extremes.

Example 5 (cont.):

The Ising model requires computation of the partition function $Z(\theta)$, and for very complex graphical systems with many nodes this is computationally prohibitive. Ravikumar et al. (2012) suggest a composite likelihood version based on conditional densities from neighbourhood contributions:

$$f(y_j \mid y_{(-j)}; \theta) = \frac{\exp(2y_j \sum_{k\neq j} \theta_{jk} y_k)}{\exp(2y_j \sum_{k\neq j} \theta_{jk} y_k) + 1}, \tag{31}$$

which eliminates $Z(\theta)$. Given a sample $y^{(1)}, \ldots, y^{(n)}$, they study estimation of θ based on a penalized version of the composite conditional likelihood:

$$\arg\max_{\theta} \left\{ \sum_{i=1}^{n} \ell_j(\theta; y^{(i)}) - \sum_j \sum_k P_\lambda(|\theta_{jk}|) \right\}, \tag{32}$$

where each component ℓ_j is the log of a conditional density as at (31). They showed that maximization of the this penalized composite log-likelihood leads to consistent estimation of the set of non-zero θ's, i.e. consistent estimation of the edge set of the relevant graphical model, even as p increases with n in a regime where $p >> n$; see for example the remarks following Corollary 1 in Ravikumar et al. (2010). A related penalized composite likelihood estimation algorithm was developed in Xue et al. (2012).

3.2. Variational Approximations. Variational methods were developed to approximate posterior distributions; for a review see Titterington (2004), for example. The main idea is to approximate the posterior $f(\theta \mid y)$ by a function $q(\theta)$[2], that is simple to compute and close to the true posterior density. One simple

[2]the dependence on y is suppressed

version of $q(\theta)$ pretends the components are independent:

$$q(\theta) = \prod_{j=1}^{d} q_j(\theta_j); \tag{33}$$

another is to model $q(\theta)$ by a simple parametric family. These two approaches are reviewed in Ormerod & Wand (2010, §2.2,3). The quality of the approximation is typically measured by the Kullback-Leibler divergence

$$KL\{q||f(\cdot \mid y)\} = \int q(\beta) \log\{q(\beta)/f(\beta \mid y) d\beta, \tag{34}$$

and minimizing this over q is equivalent to

$$\max_q \int q(\beta) \log\{f(y,\beta)/q(\beta)\} d\beta;$$

$f(y,\beta)$ is the joint density $f(y \mid \beta)\pi(\beta)$. To use this in the context of complex likelihoods with latent random variables, we need to approximate $\log \int f(y \mid u;\theta) f(u) du = \log f(y;\theta)$, say, which we can express as

$$\log f(y;\theta) = \int q(u) \log\{f(y,u;\theta)/q(u)\} du + KL\{q||f(u \mid y)\}, \tag{35}$$

from which we have

$$\log f(y;\theta) \geq \int q(u) \log\{f(y,u;\theta)/q(u)\} du;$$

here u are random effects, such as in (5), or the latent autoregression in Example 2, or the unobserved arrival times in Example 4.

Example 1 (cont.)

As an example of a variational method in which $q(\theta)$ is modelled by a parametric family, Ormerod & Wand (2012) consider the generalized linear mixed model (5) with $\phi a_i = 1$, and assume a normal distribution for the random vector u_i, as is very common in practice. The exact log-likelihood is

$$\begin{aligned}
&\sum_{i=1}^{m} \Bigg(y_i^{\mathrm{T}} X_i \beta - \frac{1}{2} \log|\Sigma| \\
&\quad + \log \int_{\mathbb{R}^k} \exp\{y_i^{\mathrm{T}} Z_i u_i - 1_i^{\mathrm{T}} b(X_i\beta + Z_i u_i) - \frac{1}{2} u_i^{\mathrm{T}} \Sigma^{-1} u_i\} du_i \Bigg) \qquad (36) \\
&\quad = \sum_{i=1}^{m} \Bigg(y_i^{\mathrm{T}} X_i \beta - \frac{1}{2} \log|\Sigma| \\
&\quad + \log \int_{\mathbb{R}^k} \exp\{y_i^{\mathrm{T}} Z_i u_i - 1_i^{\mathrm{T}} b(X_i\beta + Z_i u_i) - \frac{1}{2} u_i^{\mathrm{T}} \Sigma^{-1} u_i\} \frac{\phi_{\Lambda_i}(u-\mu_i)}{\phi_{\Lambda_i}(u-\mu_i)} du_i \Bigg) \qquad (37)
\end{aligned}$$

and the variational argument above shows that

$$\begin{aligned}\ell(\beta,\Sigma) &\geq \sum_{i=1}^{m}\left(y_i^{\mathrm{T}}X_i\beta - \frac{1}{2}\log|\Sigma|\right)\\ &\quad+\sum_{i=1}^{m}E_{u\sim N(\mu_i,\Lambda_i)}\left(y_i^{\mathrm{T}}Z_iu - 1_i^{\mathrm{T}}b(X_i\beta+Z_iu) - \frac{1}{2}u^{\mathrm{T}}\Sigma^{-1}u - \log\{\phi_{\Lambda_i}(u-\mu_i)\}\right)\\ &\equiv \ell(\beta,\Sigma,\mu,\Lambda),\end{aligned} \tag{38}$$

which simplifies the k-dimensional integration in the original likelihood to k one-dimensional integrations. The lower bound depends on two new variational parameters, μ and Λ, as well as the original parameters of the model, β and Σ. The variational estimates of β and Σ, the quantities of inferential interest, are obtained by maximizing (38) over all its arguments. It is not clear whether the resulting estimators are consistent, and asymptotically normal, nor how their efficiency compares to the maximum likelihood estimators, although some results for a Poisson mixed model are given in Hall et al. (2011a,b). Tan & Nott (2013, 2014) emphasize the construction of algorithms for solving for the variational estimates and using variational methods for model selection.

Variational methods have found more direct application in Bayesian inference, where the assumption of independence expressed in (33) is often used. Several examples involving latent random effects are described in Ormerod & Wand (2010, §2); and the approach is reviewed in the Bayesian setting in Titterington (2004, §3.2) and Bishop (2006, Ch. 10).

Using (33) in the context of likelihood inference essentially replaces $\ell(\theta;y)$ with a simpler function of θ. In contrast the composite log-likelihood replaces $\ell(\theta;y)$ with a simpler function of y; the simplest of which is the independence composite likelihood (27). In applications of composite likelihood pairwise marginal likelihood is more widely used, as it often seems to capture key aspects of the dependence among the components of y. The comparison of composite likelihood methods to variational approximations, and the combination of both approaches, is starting to appear in the machine learning literature, often with emphasis on various types of graphical models (Lyu, 2011; Zhang & Schneider, 2012; Matias & Robin, 2014). It would be of interest to make the connections more explicit; perhaps insights from the variational apporach could inform the choice of component densities for the construction of composite likelihood.

3.3. Laplace approximation and INLA. One of the simplest approximations to a log-likelihood function that involves high-dimensional integration is Laplace approximation. Suppose the log-likelihood function is of the form

$$\ell(\theta,\alpha;y) = \log\int f(y\mid u;\theta)g(u;\alpha)du = \log\int\exp\{Q(u;y,\theta,\alpha)\}du, \tag{39}$$

which would apply to a model with latent random variables, u, or to a Bayesian posterior calculation, with $g(u;\alpha)$ playing the role of a prior for the 'nuisance

parameters' u.[3] The Laplace approximation to this is obtained by expanding $Q(\cdot; y; \theta, \alpha)$ in a Taylor series about its maximum over u, say $\tilde{u}$:

$$\ell_{\text{Lap}}(\theta, \alpha; y) = Q(\tilde{u}; y, \theta, \alpha) - \frac{1}{2} \log |Q''(\tilde{u}; y, \theta, \alpha)|, \tag{40}$$

where the dependence of $\tilde{u}$ on θ, α and y is suppressed.

A simplification of this Laplace approximation, combined with a normal assumption for the distribution of the random effects, leads to an approximate log-likelihood function of the form

$$\log f(y \mid u; \theta) - \frac{1}{2} u^{\mathrm{T}} \Sigma^{-1}(\alpha) u, \tag{41}$$

to be jointly maximized over θ, u, and any unknown parameters in Σ, the covariance matrix for u (Breslow & Clayton, 1993; Green, 1987). This is often used in generalized linear mixed models as a version of an approximate likelihood. As pointed out by Molenberghs & Verbeke (2006, Ch. 14), this log-likelihood function can be viewed as an approximating linear mixed model, where the nonlinear function $g\{\mathrm{E}(y)\}$ that defined a generalized linear model is linearized about some fixed value.

One major limitation of the Laplace approximation is that it relies on the log-likelihood function having a unique maximum over u, and being 'smooth' around that maximum. While in principle it should be possible in a multi-modal setting to develop local approximations around each mode, this detracts from the simplicity of the approach. Various extensions of the simple Laplace approximation have been considered for complex and high-dimensional models. Shun & McCullagh (1995) show that the Laplace approximation remains valid if the dimension of the integral grows with n, at rate $o(n^{1/3})$. Raudenbush et al. (2000) considered expansions of linear mixed models to higher order, providing a more accurate approximate likelihood than ℓ_{PQL}.

Rue et al. (2009) proposed in a Bayesian context double use of the Laplace approximation, called integrated nested Laplace approximation (INLA), as an alternative to Markov chain Monte Carlo simulation. Suppose we have for each observation y_i a model with density $f(y_i \mid \theta_i)$, a prior $\pi(\theta_i \mid \vartheta)$ depending on additional hyper-parameters, and a prior $\pi(\vartheta)$ for these hyperparameters. The marginal posterior for θ_i is

$$\pi(\theta_i \mid y) = \int \pi(\theta_i \mid \vartheta, y) \pi(\vartheta \mid y) d\vartheta;$$

the INLA methods uses a Laplace approximation to each of the densities in the integrand.

[3] The unknown parameters α in the desnity of g are in Bayesian inference usually called hyper-parameters.

4. Estimating equations

4.1. Quasi-likelihood. A different approach to inference in complex models is to focus directly on estimation of one or more model parameters, along with estimation of their standard error. The score equation (12) is an example of an estimating equation, as is the composite score equation (28). In generalized linear models, where typically

$$\mathrm{E}(y_i;\theta) = \mu_i(\theta), \quad \mathrm{Var}(y_i;\theta) = \phi V(\mu_i), \tag{42}$$

with ϕ a dispersion parameter, and $\mu_i(\cdot)$ and $V(\cdot)$ known functions, the score equation is

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)(\partial\mu_i/\partial\beta)}{\phi a_i V(\mu_i)} = 0, \tag{43}$$

which shows that the mean/variance relationship is the key to determining the maximum likelihood estimate of β, and that this estimate does not depend on ϕ. For example, for data from a Poisson distribution with mean μ_i, $V(\mu_i) = \mu_i$. However, if we used a model in which $\mathrm{E}(y_i) = \mu_i$ and $\mathrm{var}(y_i) = \phi\mu_i$, which allows for an inflation of the Poisson variance, then the estimating equation for the regression parameter would still satisfy (43). This new model is called an over-dispersed Poisson, and the model is usually referred to as "quasi-Poisson". Solving the quasi-score equation (43) gives estimates of the parameters in the mean function that are consistent and asymptotically normal.

The quasi-log-likelihood associated with the estimating equation (43) is

$$Q(\beta; y) = \sum_{i=1}^{n} \int_{y_i}^{\mu_i} \frac{y_i - u}{\phi V(u)} du, \tag{44}$$

and has some of the properties of a genuine log-likelihood: for example the first and second Bartlett identities hold, so no 'sandwich' form of the asymptotic variance is needed, in contrast with mis-specified models. See, for example, Davison (2003; Ch. 10.6).

Quasi-likelihood inference has the flavour of composite likelihood, but instead of simplifying the log-likelihood function by ignoring some dependencies, we simplify the estimating equation by modelling only the structure of the mean and the variance function. Because (43) has expected value 0 as long as the mean function is correct, the resulting quasi-likelihood estimator is consistent for β, and under mild conditions is asymptotically normal (McCullagh, 1983).

Quasi-likelihood and composite likelihood for spatial point processes are compared in Guan et al. (2015).

This estimating equation approach was generalized to dependent data by Liang & Zeger (1986). In the generalized linear mixed model (5), the random effects induce correlation between two observations on the same subject. Thus we have

$$\mathrm{E}(y_i) = \mu_i, \quad \mathrm{var}(y_i) = V(\mu_i, \alpha), \tag{45}$$

say, where V_i is now an $m_i \times m_i$ covariance matrix for the vector y_i, and α are

parameters in the covariance function. The analogue to (43) is

$$\sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta} \right)^{\mathrm{T}} V^{-1}(\mu_i, \alpha)(y_i - \mu_i) = 0, \tag{46}$$

where some estimate of α will be needed for its solution. The solution of (46) will under mild conditions be consistent for β, even if the variance function is misspecified. However, there is no unique inference function analogous to $Q(\beta; y)$ that can be represented as an integral in this more general setting.

In generalized linear mixed models, the score equations obtained from the Laplace approximation to the log-likelihood (39) have the form of a quasi-log-likelihood function with a penalty term, so (39) is often called penalized quasi-likelihood.

Chapter 14 of Molenberghs and Verbeke (2006) provides an excellent overview of various approximations to the log-likelihood for generalized linear mixed models.

4.2. Indirect inference. Indirect inference provides another approach to inference using estimating equations, and has similarities both to composite likelihood estimation, quasi-likelihood inference, and the simulation method known as approximate Bayesian computation (ABC), described briefly in the next section. Recall that the maximum composite likelihood estimator $\hat{\theta}_{CL}$ is consistent, under regularity conditions on the model, essentially because the composite score equation has expected value 0; estimating equations with this property are called unbiased, or sometimes unbiased for 0.

Suppose in contrast we have a d-dimensional estimating function that is *biased*, i.e. we solve the equation

$$g(y_1, \ldots, y_n; \theta) = 0,$$

where $\mathrm{E}_\theta\{g(Y; \theta)\} \neq 0$. Denote the solution, assumed unique, by $\tilde{\theta}_n$. Under smoothness conditions on g and the model for $y_1, \ldots, y_n$, the usual series expansion will lead to $\tilde{\theta}_n \xrightarrow{p} \theta^*$, where θ^* is the value at which $\mathrm{E}_\theta\{g(Y; \theta^*)\} = 0$. This defines a mapping from the notional true value, θ, to θ^*, which we write as

$$\theta^* = \tilde{k}(\theta), \quad \theta = k(\theta^*),$$

with the inverse function $k(\cdot)$ called the *bridge* function (Jiang & Turnbull, 2004). The estimator

$$\hat{\theta}_n = k(\tilde{\theta}_n),$$

is thus corrected for the bias in the estimating equation, after the fact, so-to-speak. The bridge function connects the wrong limit θ^* to the desired limit θ. Yi & Reid (2010) illustrate this approach in the context of data with missing responses, showing how the missing data model affects the resulting inference. Jiang & Turnbull (2004) consider the more general case where the estimating equation may have more components than the dimension d of θ, in which case the function $\tilde{k}$ will not be invertible. They define an indirect likelihood function based on a normal approximation to the distribution of g, and show how the

usual likelihood pivots (16), (17) and (19) can be used with the indirect likelihood function.

Using a biased estimating equation is a simple form of the method of indirect inference. We could more generally build indirect inference from a set of functions of the data vector, $s_1(y), \ldots, s_q(y)$. These statistics might for example be the maximum likelihood estimate of the parameter θ^* in a simple model, known to be incorrect, but easy to use.

Indirect inference is widely used in econometrics, where the models of interest are often specified by a dynamical system of the form

$$y_i = \Psi_i(y_{i-1}, x_i, \epsilon_i; \theta), \tag{47}$$

say, where θ is the d-dimensional parameter of interest, ϵ_i are independent noise terms, and x_i are potential explanatory variables. While the likelihood function based on a sample $y_1, \ldots, y_n$ might not be computable, the form of (47) suggests that it may well be feasible to simulate from the system. The following discussion is adapted from Smith (2008), which reviews indirect inference for dynamical systems.

A simpler model, known to be wrong, but easy to use, is for example a one-step Markov model

$$y_i \sim f(y_i \mid y_{i-1}, x_i; \theta^*), \tag{48}$$

where $\theta^* \in \mathbb{R}^q$ may or may not be of the same dimension as θ; the key point is that the maximum likelihood estimate $\hat{\theta}^*$ is relatively easy to compute.

We can now consider simulating samples from (47) under various values of θ, and identifying 'good' values of θ as those providing simulated samples that lead to the same $\hat{\theta}^*$:

(i) simulate $y^m = (y_1^m, \ldots, y_n^m), m = 1, \ldots, M$ from (47) at a given value θ

(ii) compute the average log-likelihood for the simpler model

$$\sum_m \sum_i \log f(y_i^m \mid y_{i-1}^m, x_i; \theta^*),$$

and find the maximum likelihood estimate $\hat{\theta}^*(\theta)$

(iii) the estimate $\hat{\theta}$ is the value closest to $\hat{\theta}^*$ (in some metric).

If $q = d$, then step (iii) is simply a matter of inverting the bridge function $\hat{\theta}^*(\theta)$. In many applications in econometrics, $q > d$, so the indirect estimate of θ is defined by minimizing a loss function, such as

$$\{\hat{\theta}^*(\theta) - \hat{\theta}^*\}^{\mathrm{T}} W \{\hat{\theta}^*(\theta) - \hat{\theta}^*\},$$

for a choice of weight matrix W. This loss function is equivalent to using Jiang & Turnbull's (2004) indirect likelihood function. Under the usual regularity conditions these estimates of θ are consistent and asymptotically normal, but the asymptotic variance will be larger than the inverse Fisher information, and typically take the sandwich form.

The use of an auxiliary parametric model is just one version of indirect inference. A similar approach could be taken by computing the sample moments of data simulated from the complex model, and choosing the parameter values that for which the simulated moments are closest to the observed sample moments. This is a simulation version of the so-called generalized method of moments, which has been widely applied: see for example Rodriguez-Iturbe et al. (1987) for an application to hydrology.

5. Discussion

Indirect inference involves simulating from the true model, under various values of θ, and comparing the resulting estimates under the wrong model. The method of approximate Bayesian computation has much the same structure, but in the context of obtaining samples from the posterior distribution for θ, in settings where again the model governing the data can be used to generate samples, but the computation of the log-likelihood function is not feasible. As described in Marin et al. (2011), given a data set y from our complex model we proceed as follows:

(i) simulate a candidate parameter value from the prior density $\pi(\theta)$, say θ'

(ii) simulate data z from the model $f(\cdot;\theta')$

(iii) if $z = y$, then θ' is an observation from the posterior density $\pi(\cdot \mid y)$

If y has high dimension, or has a continuous density, then requiring $z = y$ is much too strong, so (iii) is typically replaced by

(iiia) summarize z and y by a set of statistics, $s(z)$ and $s(y)$, say

(iiib) accept θ' if $\rho\{s(z), s(y)\} < \epsilon$, for some distance function ρ.

There are a great many variations of the ABC algorithm, many using improved methods of sampling by incorporating ideas from the theory of Markov chain Monte Carlo sampling. Cox & Kartsonaki draw parallels between ABC methods and indirect inference, and develop an alternative method based on principles of fractional factorial design.

Example 5 (cont.): The $M/G/1$ queue is used to illustrate the ABC method in Fearnhead & Prangle (2011), and compared there to the indirect inference approach. Simulating the data is relatively straightforward, as the model is defined in steps:

$$V_1 \sim \text{Exp}(\theta_3),$$
$$V_i \mid V_{i-1} \sim V_{i-1} + \text{Exp}(\theta_3),$$
$$Y_i \mid x_{i-1}, V_i u \sim \text{Uniform}(\theta + 1 + \max(0, V_i - X_{i-1}, \theta_2 + \max(0, V_i - X_{i-1}),$$

where $X_i = \Sigma_{j=1}^{i} Y_j$. In this model the 'general' service time is actually uniform on the interval θ_1, θ_2.

For ABC simulation, Fearnhead & Prangle (2011) use quantiles of the departure times as summary statistics. The indirect inference approach uses $\bar{y}$, $y_{(1)}$ and $\hat{\theta}_2$ from the steady-state version of the model. Comparisons in numerical work show that the ABC method gives more accurate estimates of θ_1, whereas indirect inference gives more accurate estimates of θ_2, and both methods perform similarly for estimation of θ_3.

Both ABC and indirect inference need a set of parameter values from which to simulate, θ' or θ. They both require a set of auxiliary functions, $s(\cdot)$ or $\hat{\theta}^*(\cdot)$. In indirect inference $\hat{\theta}^*$ is the bridge to the parameters θ of real interest.

Simulation approaches to the computation of the maximum likelihood estimator, based on Markov chain Monte Carlo methods, are proposed in Geyer & Thompson (1992), and analysed further in Geyer (1994). Okabayashi & Geyer (2012) note difficulties with convergence, and propose an alternative optimization method, as well as a simulation-based approximation to it. In their §5.2 they discuss the Ising model of Example 5, although without an assumption of sparsity.

Optimization theory and methods play an important role in likelihood-related inference, although the theoretical advances in optimization are not always reflected in the statistical sciences literature. Okabayashi & Geyer, for example, is mainly focussed on ensuring convergence of the algorithm to the maximum likelihood estimate, with special emphasis on situations where 'good' starting values for the parameters are not available. Grosse (2015) studies the optimization algorithm based on natural gradient ascent, also known as Fisher scoring, for finding the maximum likelihood estimate in the restricted Bolzmann machine (10). Fisher scoring finds the maximum likelihood estimate by the updating the equation $\theta \leftarrow \theta + i^{-1}(\theta)u(\theta)$, which can be viewed as a steepest descent algorithm using the expected Fisher information as a measure of distance on the parameter space. In RBMs with many hidden layers, this matrix is very difficult to invert, and Grosse suggests approximations to the inverse based on Gaussian graphical models.

As models become increasingly complex, the log-likelihood function may well be non-convex, in which case many algorithms converge either to a local maximum, or a local minimum, and the behaviour of the optimization algorithm depends strongly on the starting values. The most common approach in these settings is to seek a closely related problem for which the objective function is convex. Loh & Wainwright (2015) show that this may not always be necessary, in the sense that the statistical noise in the resulting estimate is larger than the distance between local modes: the statistical error in this sense dominates the optimization error, and any local maximum is a suitable estimator. This seems an important and interesting discussion that is only starting.

The study of likelihood methods and approximate likelihood methods for complex models is an area of active research. The program "Intractable Likelihoods", supported by the Engineering and Physical Sciences Research Council in the U.K. (iLike, 2015), is investigating a number of the methods discussed in this paper, along with many others.

The discussion here has restricted attention to parametric models, and most of the results presented are dependent on the model being correctly specified. Likelihood approaches that do not require the specification of a parametric model are also widely studied. One class of such models, usually called semi-parametric models, specify a parametric model for the parameters of interest, say a small subset of regression coefficients, and but allow the nuisance parameter to be an unknown function. The most well-known example of such a semi-parametric model is Cox's (1972) proportional hazards model for failure time data. Murphy and van der Vaart (2000) studied profile likelihood inference for a class of semi-parametric models; see also van der Vaart (1998, Ch. 25). Another approach to non-parametric likelihood inference is the empirical likelihood of Owen (1990; 2001); see also Hjort et al. (2009) for several extensions. Yet another approach is based on bootstrap resampling (Davison et al., 1992; Pawitan, 2000).

Penalized log-likelihoods (Green, 1987) have also been used to fit semi-parametric models, with the nonparametric component modelled typically by a linear combination of a large number of basis functions, and the problem is regularized by a constraint on the size of the linear coefficients. Regularized likelihood inference has been the subject of intense investigation for high-dimensional models, both as an inference method and as a method of model selection (Fan & Li, 2001).

We have not included in this survey all the "likelihood-like" inference functions that have been proposed: those omitted include h-likelihood (Nelder & Lee, 1992), local likelihood (Hastie et al. 2008, §6.5), weighted likelihood (Hu & Zidek, 2002; Plante, 2008), sieve likelihood (van der Waart, Ch. 25; Geman & Hwang, 1982), and surely others. Fisher's (1922) prescient but naively simple idea to concentrate on the θ-section of the density $f(y;\theta)$ continues to drive research in statistical theory and methods, nearly 100 years later.

Dutta, S. & Mondal, D. (2015). An h-likelihood method for spatial mixed linear models based on intrinsic auto-regressions. *J. Roy. Statist. Soc.* B, **77**, 699-726.
REML for spatial mixed linear models. spatial agric field trials

Acknowledgements

I am indebted to Don Fraser, Yang Ning, Cristiano Varin and Grace Yi for ongoing collaborations, and helpful discussions of likelihood-based inference.

References

Gaun, Y., Jalilian, A. & Waagepetersen, R. (2015). Quasi-likelihood for spatial point processes. *J. Roy. Stat. Soc.* B **77**, 677–698.

[1] Barndorff-Nielsen, O.E., Hoffman-Jørgensen, J. & Pederson, K. (1976). On the minimal sufficiency of the likelihood function. *Scand. J. Statist.* **3**, 37-38.

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

[3] Bishop, C. (2006). *Pattern Recognition and Machine Learning.* Springer, New York.

[4] Brazzale, A.R., Davison, A.C. & Reid, N. (2007). *Applied Asymptotics.* Cambridge University Press, Cambridge.

[5] Breslow, N.E. & Clayton, D. G. (1993). Approximate inference in generalizsed linear mixed models. *J. Am. Statist. Assoc.* **88**, 9–25.

[6] Burnham, K.P. & Anderson, D.R. (2002). *Model selection and multimodel inference: a practical information-theoretic approach.* Springer, New York.

[7] Cox, D.R. (1972). Regression models and life tables. J. Roy. Statist. Soc. B **34**, 187220.

[8] Cox, D.R. & Hinkley, D.V. (1974). *Theoretical Statistics*. Chapman & Hall, London.

[9] Cox, D.R. & Kartsonaki, C. (2012). The fitting of complex parametric models. *Biometrika* **99**, 741–747.

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

[11] Davison, A.C. (2003). *Statistical Models*. Cambridge University Press, Cambridge.

[12] Davison, A.C., Hinkley, D.V. & Worton, B.J. (1992). Bootstrap likelihoods. *Biometrika***79**, 113–130.

[13] Davison, A.C., Padoan, S.A. & Ribatet, M. (2012). Statistical modeling of spatial extremes. *Statistical Science* **27**, 161–186.

[14] Davison, A.C. & Huser, R.(2015). Statistics of extremes *Annual Reviews* **2**, to appear.

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

[16] Fearnhead, P. & Prangle, D. (2012). Approximate likelihood methods for estimating local recombination rates *J. R. Statist. Soc.* B **64**, 657–680.

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

[18] Fisher, R.A. (1956). *Statistical Methods and Scientific Inference.* Oliver & Boyd, Edinburgh.

[19] Fraser, D.A.S. & Naderi, A. (2007). Minimal sufficient statistics emerge from the observed likelihood function. *Int. J. Statist. Sci.* **6** 55–61.

[20] Geman, S. & Hwang, C.-R. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *Ann. Statist.* **10**, 401–414.

[21] Geyer, C. (1994). On the convergence of Monte Carlo maximum likelihood calculations. *J. Roy. Statist. Soc.* B **56**, 261–274.

[22] Geyer, C. & Thompson, E.A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *J. R. Statist. Soc.* B **54**, 657–699.

[23] Godambe, V.P. (1960). An optimum property of regular maximum likelihood estimation. *Ann. Math. Statist.* **31**, 1208–1212.

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

[25] Grosse, R. (2015). Scaling up natural gradient by sparsely factorizing the Fisher information matrix. preprint.

[26] Hall, P., Ormerod, J.T. & Wand, M.P. (2011a). Theory of Gaussian variational approximation for a Poisson linear mixed model. *Stat. Sinica* **21**, 369–389.

[27] Hall, P., Pham, T., Wand, M.P. & Wang S.S.J. (2011b). Asymptotic normality and valid inference for Gaussian variational approximation. *Ann. Statist.* **39** 2502–2532.

[28] Hastie, T., Tibshirani, R. & Friedman, J. (2008). *Elements of Statistical Learning.* 2nd ed. Springer, New York.

[29] Heggland, K. & Frigessi, A. (2004). Estimating functions in indirect inference. *J. R. Statist. Soc.* B **66**, 447–462.

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

[31] Hjort N. L., McKeague I. W. & van Keilegom, I. (2009). Extending the scope of empirical likelihood. *Ann. Statist.* **37**, 1079–1111.

[32] Hu, F. & Zidek, J.V. (2002). The weighted likelihood. *Can. J. Statist.* **30**, 347–371.

[33] iLike (2015). Intractable Likelihoods. `http://www.i-like.org.uk/`, accessed on June 2, 2015.

[34] Jiang, W. & Turnbull, B. (2004). The indirect method: inference based on intermediate statistics: a synthesis and examples. *Statistical Science* **19**, 239–263 .

[35] Liang, K.-Y. & Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

[36] Lindsay, B. (1988). Composite likelihood methods. *Contemp. Math.* **80**, 220–239.

[37] Loh, P. and Wainwright, M. (2015). Regularized M-estimators with nonconvexity. *J. Machine Learning Res.* **16**, 559–616.

[38] Lyu, S. (2011). Unifying non-maximum likelihood learning objectives with minimum KL contraction. in *Advances in Neural Information Processing Systems* **24**, J. Shawe-Taylor and R.S. Zemel, eds., 64–72.

[39] Marin, J.-M., Pudlo, P., Robert, C.P. & Ryder, R.J. (2011). Approximate Bayesian computational methods. *Stat. & Computing* **22**, 1167–1180.

[40] Matias, C. & Robin, S. (2014). Modelling heterogeneity in random graphs through latent space models: a selective review. `http://arxiv.org/1402.4296.pdf`, accessed on June 2, 2015.

[41] McCullagh, P. (1983). Quasi-likelihood functions. *Ann. Statist.* **11**, 59–67.

[42] Molenberghs, G. & Verbeke, G. (2006). *Discrete Longitudinal Data* Springer, New York.

[43] Murphy, S. A. & van der Vaart, A. W. (2000). On profile likelihood (with discussion). *J. Am. Statist. Assoc.* **95**, 449–465.

Lee, Y., Nelder, J.A¿ & Pawitan, Y. (2006). *Generalized Lienar Models with Random Effects Unified Analysis via H-likelihood.* Chapman & Hall/CRC, Boca Raton.

[44] Nelder, J.A. and Lee, Y. (1992). Likelihood, quasi-likelihood and pseudolikelihood: some comparisons. *J. R. Statist. Soc.* B **54**, 273–284.

[45] Okabayashi, S. & Geyer, C.J. (2012). Long range search for maximum likelihood in exponential families. *Elect. J. Statist.* **6**, 123–147.

[46] Ormerod, & Wand, M. (2010). Explaining variational approximations. *Am. Stat.* **64**, 140–153.

[47] Ormerod, & Wand, M. (2012). Gaussian variational approximate inference for generalized linear mixed models. *J Comp Graph Statist***21**, 2–17.

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

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

[50] Plante, J.-F. (2008). Nonparametric adaptive likelihood weights. *Canad. J. Statist.* **36**, 443–461.

[51] Pawitan, Y. (2000). Computing empirical likelihood from the bootstrap. *Statist. Prob. Letters* **47**, 337–345.

[52] Raudenbush, S.W., Yang, M.-L. & Yosef, M. (2000). Maximum likelihood for generalized linear models with nested and random effects via high-order, multivariate Laplace approximations. *J. Comp. Graph. Statist.* **9**, 141–157

[53] Ravikumar, P., Wainwright, M.J. & Lafferty, J. (2010). High-dimensional Ising model selection using ℓ_1 regularized regression. *Ann. Statist.* **38**, 1287–1319.

[54] Robin, S. (2011). `http://carlit.toulouse.inra.fr/AIGM/pub/Reunion_nov2012/MSTGA-1211-Robin.pdf` online talk

[55] Rodriguez-Iturbe, I., Cox, D.R. & Isham, V. (1987). Some models for rainfall based on stochastic point processes. *Proc. R. Soc. Lond.* **410**, 26988.

[56] Royall, R.J. (1997). *Statistical Evidence: A Likelihood Paradigm.*. Chapman & Hall/CRC, London.

[57] Rue, H. & Martino, S. (2009). Approximation Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Statist. Soc.* B **71**, 319–392.

[58] Shalizi, C. (2013). Notebooks. `http://vserver1.cscs.lsa.umich.edu/crshalizi/notebooks/` indirect inference

[59] Shestopaloff, A. & Neal, R. (2013). On Bayesian inference for the M/G/1 queue with efficient MCMC sampling. `http://www.cs.toronto.edu/fadford/ftp/queue-mcmc.pdf`, accessed on June 2, 2015.

[60] Shun, Z. & McCullagh, P. (1995). Laplace approximation of high dimensional integrals *J. R. Statist. Soc.* B *57*, 749–760.

[61] Smith, A.A. (2008). Indirect inference. in *New Palgrave Dictionary of Economics* 2nd ed. `http://www.dictionaryofeconomics.com/`, accessed on June 1, 2015.

[62] Tan, & Nott, (2013). Variational inference for generalized linear mixed models using partially noncentered parametrizations. *Statist. Sci.* **28**, 168–188.

[63] Tan & Nott (2014). Variational approximation for mixtures of linear models. *J. Comp. Graph. Statist.* **23**, 564–585.

[65] Titterington, D.M. (2006). Bayesian methods for neural networks and related models. *Statistical Science* **19**, 128–139.

[65] van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

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

[67] Xue, L., Zou, H. & Cai, T. (2012). Nonconcave penalized composite conditional likelihood of sparse Ising models. *Ann. Statist.* **40**, 1403–1429.

[68] Yi, G. & Reid, N. (2010). A note on misspecified estimating equations. *Statistica Sinica* **20**, 1749–1769.

[69] Zhang, Y. & Schneider, J. (2012). A composite likelihood view for multi-label classification. in Proceedings of the 15th International Conference on Artificial Intelligence and Statistics. http://www.jmlr.org/proceedings/papers/v22/, accessed on June 2, 2015.

Department of Statistical Sciences
University of Toronto
100 St. George St.
Toronto Canada M5S 3G3
E-mail: reid@utstat.utoronto.ca