Likelihood inference

Nancy Reid*

The essential role of the likelihood function in both Bayesian and non-Bayesian inference is described. Several topics related to the extension of likelihood-based methodology to more complex settings are reviewed, including modifications to profile likelihood, composite and pseudo-likelihoods, quasi-likelihood, semi-parametric and non-parametric likelihoods, and empirical likelihood. © 2010 John Wiley & Sons, Inc. *WIREs Comp Stat* 2010 2 517–525

INTRODUCTION

The likelihood function for a parametric model is proportional to the density function of the model, but is considered as a function of the parameters in the model, with the data held fixed. In machine learning applications, where inference about the model parameters is often less important than prediction of new instances, the negative of the log of the likelihood function can serve as a useful loss function. The likelihood function has proved to be such a powerful tool for inference that it has been extended and generalized to semi-parametric models and non-parametric models, and various pseudo-likelihood functions have been proposed for more complex models. This article will review some of the extensions to likelihood and likelihood-based inference that have been developed for the analysis of large or complex data sets.

NOTATION AND EXAMPLES

We start with a given parametric model, $f(y;\theta)$, the probability density function for a random variable Y. At least initially we assume that y is a vector of n components $y_1,\ldots,y_n$, $y_i \in \mathbb{R}$, and $\theta \in \Omega$. In regular statistical models, Ω is very often taken to be $\mathbb{R}^d$ or a subset of $\mathbb{R}^d$.

The likelihood function for this parametric model is

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

viewed as a function of θ, for fixed y. While some authors define the likelihood function without the arbitrary function $c(y)$, this definition shows explicitly that the value of the likelihood function is only meaningful in relative terms. It is usually more convenient to work with the log-likelihood function

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

this is particularly useful when the components of y are independent.

Example 1 If $Y = (Y_1,\ldots,Y_n)$ are independent and identically distributed normal random variables, with mean μ and variance σ^2, then

$$\ell(\theta; y) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum(y_i-\mu)^2,$$

where $\theta = (\mu,\sigma^2)$, $\Omega = \mathbb{R}\times\mathbb{R}^+$, and following definition (2), we can ignore the constant term $-(n/2)\log(2\pi)$. This example can be generalized in many ways, for example by assuming the Y_i are independent, with means μ_i; if further $\mu_i = x_i^T\beta$ where x_i is a $q\times 1$ vector of known values associated with the ith component, then we have a standard linear regression model with log-likelihood function

$$\ell(\theta; y) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum\left(y_i - x_i^T\beta\right)^2,$$

where $\theta = (\beta,\sigma^2)$.

There are many extensions of this simple regression model. One important class of extensions is to fitting a model where the mean is specified only as a 'smooth function' of the covariates:

$$y_i = m(x_i) + \epsilon_i.$$

In applications, the smooth function is often modeled using a set of basis functions. For example, with a

*Correspondence to: reid@utstat.utoronto.ca

Department of Statistics, University of Toronto, Toronto, Canada M5S 3G3

DOI: 10.1002/wics.110

single covariate ($q = 1$) we might write

$$m(x) = \sum_{j=1}^{J} \phi_j B_j(x) \qquad (3)$$

for some functions $B_j(\cdot)$; both J and B_j are to be specified. A popular choice for B_j is the set of B-spline basis functions; alternatives include sine and cosine functions, or wavelet bases. With more than one covariate, one popular choice is to model the mean with q distinct smooth functions, and the result is called an additive model. For small values of q, usually at most 3, another possibility is to use spline basis functions in 2 or 3 dimensions, such as thin plate splines. A good general reference is Wood.[1]

Example 2 Suppose each y_i is itself a vector of length k, and an observed value of y_i has a 1 in the cth place and zeroes elsewhere, if the ith data point is in class c. A general model for this is a multinomial, and a sample of size n from the multinomial distribution has log-likelihood

$$\ell(\theta; y) = \sum_{i=1}^{n} \sum_{c=1}^{k} y_{ic} \log(p_c),$$

where $0 \le p_c \le 1$, $\sum_c p_c = 1$, and $\sum_{ic} y_{ic} = n$. In machine learning this is often called the negative cross-entropy function; see, for example, Ref 2 (Ch. 7). As in Example 1, we could model the vector of probabilities p through some input variables x and unknown parameters β, or with a smooth function $m(\cdot)$.

Example 3 If the sequence $y = (y_1, \ldots, y_n)$ is observed sequentially in time, then the model for y can be written as a product of conditional densities:

$$f(y_1, \ldots, y_n) = f(y_1) \cdot f(y_2 \mid y_1) \ldots f(y_n \mid y_{n-1}, \ldots, y_1).$$

For example, if we assume a Markov structure then this can be simplified to

$$f(y_1, \ldots, y_n) = f(y_1) \prod_{i=2}^{n} f(y_i \mid y_{i-1}).$$

An unknown vector of parameters θ could enter as part of the modeling of these conditional densities. As an example, we might have $y_i = (1 - \rho)\mu + \rho y_{i-1} + \epsilon_i$, with the innovations ϵ_i taken to be independent. This is an autoregressive model of order 1; the likelihood is fully specified by assuming a model, such as Gaussian, for the distribution of ϵ, and a model for the starting value y_0. Extensions incorporating additional parameters with covariates x_i could also be introduced.

Example 4 The log-likelihood function for a non-homogeneous Poisson process evolving in time is given by

$$\sum_{i=1}^{n} \log\{\lambda(y_i)\} - \int_0^{\infty} \lambda(u) du, \quad 0 < y_1 < \cdots < y_n, \qquad (4)$$

where $\lambda(\cdot)$ is the rate function for the process, and events are observed to occur at times $y_1, \ldots, y_n$. Parameters θ are introduced into the log-likelihood function by specifying a parametric form for the rate function, such as $\lambda(t) = \lambda$, or $\lambda(t) = \exp\{x(t)^T \beta\}$. This formulation can be greatly extended, for example to data measured in space, rather than time, in which case (4) becomes

$$\sum_{i=1}^{n} \log\{\lambda(y_i)\} - \Lambda(\mathcal{S}),$$

where $(y_1, \ldots, y_n)$ now take values in a set $\mathcal{S}$, for example a set of latitude and longitude points in a spatial area, and $\Lambda(\mathcal{S}) = \int_{\mathcal{S}} \lambda(s) ds$.

A number of derived quantities are routinely used in parametric inference. The score function $\ell'(\theta) = \partial \ell(\theta; y)/\partial \theta$ is typically used to obtain the maximum likelihood estimator, which in regular models satisfies

$$\ell'(\hat{\theta}; y) = 0.$$

The negative second derivative of the log-likelihood is called the (Fisher) information function and the observed and expected Fisher information are, respectively,

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

where the expectation is over the distribution of $y = (y_1, \ldots, y_n)$.

In models where some components of θ are of direct interest and others are nuisance parameters, it is usual to write $\theta = (\psi, \lambda)$, where ψ is the parameter of interest, and to partition $j(\theta)$ and $i(\theta)$ accordingly, for example

$$i(\theta) = \begin{pmatrix} i_{\psi\psi}(\theta) & i_{\psi\lambda}(\theta) \\ i_{\lambda\psi}(\theta) & i_{\lambda\lambda}(\theta) \end{pmatrix}.$$

General introductions to the definition of the likelihood function and its use ininference are given by Fisher,[3] Edwards[4] and Azzalini.[5] A large number of relevant and interesting models are discussed by Davison[6] (Ch. 4 and 6), Cox and Hinkley[7] (Ch. 2), and Barndorff-Nielsen and Cox[8] (Ch. 2). Example 4 above has been drawn from the work of Davison[6] (Ch. 6.5, where several other examples are presented).

LIKELIHOOD INFERENCE: BAYES AND FREQUENTIST

In fairly wide generality the following convergence results can be derived:

$$\ell'(\theta)^T\{j(\hat\theta)\}^{-1}\ell'(\theta) \quad \overset{\mathcal{L}}{\to} \quad \chi^2_d, \tag{5}$$

$$(\hat\theta-\theta)^T j(\hat\theta)(\hat\theta-\theta) \quad \overset{\mathcal{L}}{\to} \quad \chi^2_d, \tag{6}$$

$$2\{\ell(\hat\theta)-\ell(\theta)\} \quad \overset{\mathcal{L}}{\to} \quad \chi^2_d, \tag{7}$$

where the limit is taken as n the dimension of y, goes to ∞.[1] In (5)–(7), χ^2_d is the chi-squared distribution on d degrees of freedom, where d is the dimension of θ. One of the necessary conditions to obtain these results is that a central limit theorem is available for the $\ell'(\theta; y)$, which is a sum of n quantities if the components of y are independent. Also needed is the convergence (in probability) of the maximum likelihood estimator $\hat\theta$ to the true value θ, and this can often be difficult to establish for some models; in many discussions it is simply assumed to be true.

Similar results are available for inference about component parameters: writing $\theta = (\psi, \lambda)$, and denoting by $\hat\lambda_\psi$ the constrained maximum likelihood estimate of λ for ψ fixed,

$$\sup_\lambda \ell(\psi,\lambda; y) = \ell(\psi,\hat\lambda_\psi; y) = \ell_{\mathrm{P}}(\psi), \tag{8}$$

we have, for example,

$$\ell_{\mathrm{P}}'(\psi)^T j^{\psi\psi}(\hat\theta)\ell_{\mathrm{P}}'(\psi) \quad \overset{\mathcal{L}}{\to} \quad \chi^2_q, \tag{9}$$

$$(\hat\psi-\psi)^T\{j^{\psi\psi}(\hat\theta)\}^{-1}(\hat\psi-\psi) \quad \overset{\mathcal{L}}{\to} \quad \chi^2_q, \tag{10}$$

$$2\{\ell(\hat\psi,\hat\lambda)-\ell(\psi,\hat\lambda_\psi)\} \quad \overset{\mathcal{L}}{\to} \quad \chi^2_q, \tag{11}$$

where q is the dimension of ψ. The function $\ell_{\mathrm{P}}(\psi)$ defined in (8) is the *profile*, or *concentrated*, log-likelihood function.

The first-order approximations suggested by these limiting results, such as

$$\hat\theta \ \dot\sim\ N\{\theta, j^{-1}(\hat\theta)\}, \tag{12}$$

$$\hat\psi \ \dot\sim\ N\{\psi, j^{\psi\psi}(\hat\theta)\}, \tag{13}$$

$$\pm\sqrt{2\{\ell(\hat\psi,\hat\lambda)-\ell(\psi,\hat\lambda_\psi)\}} \ \dot\sim\ N(0,1), \tag{14}$$

are widely used in practice for inference about θ. Most statistical packages now have general purpose routines for calculating these approximations. The third approximation applies only for $q = 1$, and the sign is usually taken as $\mathrm{sign}(\hat\psi - \psi)$.

Bayesian inference based on the likelihood function is quite straightforward in principle: a prior probability distribution for θ, denoted $\pi(\theta)$ is combined with the likelihood function using the rules of conditional probability to form the posterior density for θ,

$$\pi(\theta \mid y) = \frac{L(\theta; y)\pi(\theta)}{\int L(\theta; y)\pi(\theta)d\theta}. \tag{15}$$

Inference for a sub-parameter, $\psi(\theta)$, say, is obtained from the marginal density for ψ:

$$\pi_m(\psi \mid y) = \int_{\psi(\theta)=\psi} \pi(\theta \mid y)d\theta, \tag{16}$$

and a point estimate for ψ might be the mean or the mode of this marginal posterior. Marginal posterior probability statements are also readily obtained, so, for example, a posterior interval with probability $(1-\alpha)$ is given by $(\psi_{\mathrm{L}}, \psi_{\mathrm{U}})$, obtained by finding ψ_{L} and ψ_{U} so that

$$\int_{\psi=\psi_{\mathrm{L}}}^{\psi=\psi_{\mathrm{U}}} \pi_m(\psi \mid y)d\psi;$$

this interval is not unique, and one choice often recommended is to also require the interval to have highest posterior density.

The integrals needed for computation of (15) and (16) can be approximated by various methods, including Laplace's approximation or quadrature rules. In high-dimensional cases, samples from the posterior density can be obtained using Markov chain Monte Carlo (MCMC) sampling. This involves constructing a Markov chain with stationary distribution proportional to $\pi(\theta \mid y)$ and simulating samples from the stationary distribution by running the Markov chain for a sufficiently long time. There is a very large literature on techniques for MCMC sampling and convergence properties: two book-length treatments are Refs 9 and 10.

In most scientific applications of Bayesian methods it is of interest to understand the behavior of Bayesian inference under the probability distribution given by the model $f(y; \theta)$, that is, for a fixed value of θ and random sampling from y. This provides a means for studying, for example, whether a posterior marginal probability interval for ψ has validity under

[1] More generally the limit can be taken as the expected Fisher information in y increases. Recall that $\hat\theta = \hat\theta(y) = \hat\theta(y_1, \ldots, y_n)$.

the sampling model. Under conditions on the model and the prior it can be shown that the posterior density for $\theta - \hat{\theta}$ is asymptotically normal with mean 0 and variance given by the inverse of the observed Fisher information[11] (Ch. 4); informally this is described as the prior is 'swamped by the data'. To assess the influence of the prior on the posterior, from the point of view of asymptotic theory, it is necessary to use results on higher order approximations, based on series expansions of the distribution of the maximum likelihood estimator, in the non-Bayesian setting, and series expansions to posterior integrals, in the Bayesian setting. For a debate on the relevance of model-based assessments of posterior probabilities, see, for example, Refs 12, 13 and the discussion of these papers.

COMPUTATION OF LIKELIHOOD QUANTITIES

The maximum likelihood estimator is in regular models obtained from the root(s) of the score equation

$$\ell'(\theta; y) = 0,$$

which is usually solved iteratively, using a method like Newton–Raphson or gradient descent. The method of Fisher scoring uses Newton–Raphson with the second derivative replaced by its expected value. If the equation has multiple roots, the maximum likelihood estimator is found, in principle, by finding all the roots and choosing the one with the largest likelihood.

In the class of generalized linear models there is enough smoothness in the model to ensure that the score equation above has a unique root, which is indeed the maximum likelihood estimator, and this solution can be found by an iteratively reweighted least squares fit. This allows many of the techniques of linear regression to be extended to this class of nonlinear models, with the log-likelihood function replacing squared error.

Example 5 Suppose the response Y_i is binomial, with sample size n_i and probability of success p_i, and that we have a sample of k independent observations from this model. The log-likelihood function is

$$\ell(p) = \sum_{i=1}^{k} \{y_i \log(p_i) + (n_i - y_i) \log(1 - p_i)\}.$$

With no further information linking the observations, the maximum likelihood estimate of p is simply vector of observed proportions $(y_1/n_1, \ldots, y_k/n_k)$. With a number of covariates $x_{i1}, \ldots, x_{iq}$, potentially associated with each y_i, a possible model for p is the logistic model

$$\log \frac{p_i}{1 - p_i} = x_i^T \beta,$$

which gives

$$\ell(\beta) = \sum_{i=1}^{k} y_i x_i^T \beta - n_i \log\{1 + \exp(x_i^T \beta)\},$$

and score equation

$$\sum (y_i/n_i) x_i^T = \sum p_i(\beta) x_i^T,$$

where $p_i(\beta) = \exp(x_i^T \beta)/\{1 + \exp(x_i^T \beta)\}$. This can be framed as a weighted least squares problem, with the weights depending on β. An initial guess of β provides the starting weights, and at each step the least squares equation is solved and the weights updated, until convergence. The special case of binary data, with $n_i = 1$ is often useful in practice, although diagnostics and fitting in this situation need more care.[14]

Example 6 In a feed-forward neural network with one hidden layer, the model can be expressed as a nonlinear regression model[2] (Ch. 10):

$$\begin{aligned} Y_i &\sim Bin(n_i, p_i), \\ \log \frac{p_i}{1 - p_i} &= \beta_0 + Z_i^T \beta, \\ Z_{im} &= \frac{\exp(x_i^T \alpha_m)}{1 + \exp(x_i^T \alpha_m)}, \end{aligned}$$

where the Z's are not observed. This has some similarities to the logistic regression model, and using the cross-entropy loss to fit this model is the same as maximum likelihood estimation. However, the log-likelihood function is essentially over-parameterized, so has several local maxima. Standard algorithms do not automatically attempt to find all the roots and then choose the one with largest likelihood. The approach recommended by Venables and Ripley[14] (Ch. 9) is to use a number of random starting points, thus fitting several neural networks, and to average the predictions.

A widely used algorithm for computing maximum likelihood estimates, developed for models which allow for missing data, is the EM algorithm of Dempster et al.[15] This algorithm iterates between estimating the missing observations and maximizing the likelihood function for the complete data. See Ref 6 (Ch. 5) for an introduction and Ref 16 or Ref 17 for more detailed discussion and further references.

LIKELIHOOD FOR MODEL SELECTION

The χ^2 approximation derived from the asymptotic result (7) or (11) provides a test of the hypothesis $\theta = \theta_0$ or $\psi = \psi_0$; for example (11) can be used to assess whether some components in a logistic regression model are significantly different from zero. The model with $\psi = \psi_0$ is nested in the original model $f(y;\theta)$ by assumption, in the sense that the parameter space is a subset of Ω. In treatments of generalized linear models, twice the difference of log-likelihoods comparing the parametric model with a non-parametric competitor is called the *deviance*, and the contribution from the *i*th of *n* independent observations from the model is called the deviance contribution from y_i. These deviance contributions play the role of residuals in some diagnostic methods, and choosing among nested generalized linear models is often done by analysis of deviance, repeatedly using result (11).

Example 5 (*cont*) The fully non-parametric fit of the binomial model is $\hat{p}_i = y_i/n_i, i = 1, \ldots, k$, and the deviance comparing this model to the logistic regression model is

$$D = 2\sum_{i=1}^{k}[\ell(\hat{p}_i; y_i) - \ell\{p_i(\hat{\beta}); y_i\}] = \sum_{i=1}^{k} d_i\{y_i, p_i(\hat{\beta})\}.$$

In a normal theory linear regression, the deviance is simply the negative sum of squared residuals.

We might hope to choose the best model among a set of competing models by finding the one that has the largest value of the maximized log-likelihood function. However, this will always choose the most complex model, as we can always do at least slightly better on the data set we are fitting by making the model more complex. This is a familiar problem in linear regression, where adding additional covariates is guaranteed to reduce the residual sum of squares.

A commonly used approach to this problem is to add to the log-likelihood function a penalty for model complexity. The most widely used version is Akaike's Information Criterion, defined as

$$\mathrm{AIC} = -2\log\ell(\hat{\theta}; y) + 2p, \tag{17}$$

where p is the number of parameters estimated in $\ell(\theta;\cdot)$. Models with smaller values of AIC are preferred over models with larger values, and the term $2p$ is a penalty for fitting models with larger number of parameters. This expression for AIC is derived by Davison[6] (Ch. 4) as an estimate of the Kullback–Leibler divergence between the fitted model $f(y;\hat{\theta})$ and the true model g: in his derivation the fitted model need not be nested within the true model.

The KL-divergence arises in a number of contexts in statistical inference and in information theory. In particular, writing $\hat{G}_n(y)$ for the empirical distribution function based on a sample $y_1, \ldots, y_n$ from a distribution G, we can see that the maximum likelihood estimator $\hat{\theta}$ is that value that minimizes the KL-divergence between $f(y;\theta)$ and $d\hat{G}_n(y)$, where $d\hat{G}_n$ puts mass $1/n$ at each of the n observations.

It is very common in regression-type settings, such as (3), to use AIC as a tool for model choice. It is known, however, to actually be inconsistent for this purpose, and various modifications have been suggested. The original derivation of AIC was in the time series context, where the focus on prediction from the fitted model arises somewhat more naturally. A good recent reference is Claeskens and Hjort.[18] There are several other model selection criteria similar to AIC; in particular a Bayesian version due to Schwarz called BIC replaces $2p$ in (17) with $\log(n)p$.[19]

MODIFIED PROFILE LIKELIHOOD

In this section we re-visit the approximations given at (9), (10) and (11). The assumption is that we have a model with a fairly high-dimensional parameter, θ, but that many components of θ are nuisance parameters, incorporated to make the model more realistic, but not of particular interest in themselves. Thus we partition θ as (ψ, λ), with ψ the parameters of interest and λ the nuisance parameters. It is intuitively clear that the profile log-likelihood is too concentrated around its maximum point, $\hat{\psi}$, because we have not allowed for errors of estimation of the nuisance parameters λ, so, for example, the curvature of ℓ_{P} at $\hat{\psi}$ is likelihood an overly optimistic estimate of the precision of the maximum likelihood estimator $\hat{\psi}$.

Example 1 (*cont.*) If the model is $y_i = x_i^T\beta + \epsilon_i$, where x_i is a $q \times 1$ vector of known covariate values, and ϵ_i is assumed to follow a $N(0,\psi)$ distribution, the maximum likelihood estimate of ψ is

$$\hat{\psi} = \frac{1}{n}\Sigma(y_i - x_i^T\hat{\beta})^2, \tag{18}$$

which tends to be too small, as it does not allow for the fact that q unknown parameters (the components of β) have been estimated. In this example, there is a simple improvement, based on the result that the likelihood function for (β, ψ) factors into

$$L_1(\beta, \psi; \bar{y})L_2\{\psi; \Sigma(y_i - x_i^T\hat{\beta})^2\}. \tag{19}$$

The factor $L_2(\psi)$ is proportional to the marginal density of the residuals, $\Sigma(y_i - x_i^T\hat{\beta})^2$, and basing inference for ψ only on this marginal likelihood leads, for example, to the maximum marginal likelihood estimate

$$\hat{\psi}_m = \frac{1}{n-q}\Sigma(y_i - x_i^T\hat{\beta})^2, \tag{20}$$

an unbiased estimate of ψ. The estimate based on the marginal likelihood of the residuals is often called the restricted maximum likelihood (REML) estimate, and REML methods are particularly important in estimating variance components in linear models with random effects. A book-length discussion is available in Ref 20.

The theory of higher order approximations has been used to derive a general adjustment to the profile likelihood or log-likelihood function, which takes the form

$$\ell_{\mathrm{A}}(\psi) = \ell_{\mathrm{P}}(\psi) + \frac{1}{2}\log|j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)| + B(\psi), \tag{21}$$

where $j_{\lambda\lambda}$ is defined by the partitioning of the observed information function, and $B(\psi)$ is a further adjustment function that is $O_p(1)$. Several versions of $B(\psi)$ have been suggested in the statistical literature: the main goal is to adjust the profile log-likelihood for errors in estimation of the nuisance parameters λ, essentially by finding an approximation to the factorization (19). Marginal likelihoods for scale parameters in linear regression models with non-normal errors are discussed by Fraser.[21] Barndorff-Nielsen[22] suggested a general form for $B(\psi)$ based on higher order approximations and Fraser[23] proposed a closely related version that can be calculated without explicit specification of approximately ancillary statistics.

In the special case that ψ is orthogonal to the nuisance parameter λ with respect to expected Fisher information, that is $i_{\psi\lambda}(\theta) = 0$, a simplification of $\ell_{\mathrm{A}}(\psi)$ is available as

$$\ell_{\mathrm{CR}}(\psi) = \ell_{\mathrm{P}}(\psi) - \frac{1}{2}\log|j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|, \tag{22}$$

which was introduced by Cox and Reid.[24] The change of sign on $\log|j_{\lambda\lambda}|$ comes from the orthogonality equations. In independent, identically distributed sampling, $\ell_{\mathrm{P}}(\psi)$ is $O_p(n)$, i.e. is the sum of n bounded random variables, whereas $\log|j_{\lambda\lambda}|$ is $O_p(1)$. A drawback of ℓ_{CR} is that it is not invariant to one-to-one reparametrizations of λ, all of which are orthogonal to ψ. In contrast $\ell_{\mathrm{A}}(\psi)$ is invariant to transformations $\theta = (\psi, \lambda)$ to $\tilde{\theta} = \{\psi, \eta(\psi, \lambda)\}$, sometimes called interest-respecting transformations. Inference based on various versions of $\ell_{\mathrm{A}}(\cdot)$, that is for various choices of $B(\cdot)$, is discussed by DiCiccio et al.,[25] Chang and Mukerjee,[26] and references therein.

A theory of higher order approximations to likelihood-based quantities, refining approximations such as (12), (13), and (14), has been developed in a long series of papers beginning with Refs 27–30; these in turn built on the saddlepoint approximation of Daniels[31] and Edgeworth expansions.[8] Concise accounts of the theory are available in several books including Refs 8, 32, and 33. A number of applications of higher order asymptotics are presented by Brazzale et al.[34]

EXTENSIONS OF LIKELIHOOD

There are many great likelihood-type functions that have been suggested for inference in setting with complex data. One of the most important is the partial likelihood for censored survival data.[35,36]

Example 7 Suppose we have a response y_i on each of n individuals, where y_i is either a true failure time or a censored failure time, along with an indicator variable that identifies the uncensored observations. A model closely related to the non-homogeneous Poisson process of Example 4 is to assume that the failure rate for the ith individual takes the form

$$\lambda(t_i) = \exp(x_i^T\beta)\lambda_0(t_i), \tag{23}$$

where x_i is a vector of covariates associated with the individual, and $\lambda_0(t)$ is a baseline failure rate, left unspecified. The k observed failure times are ordered as $y_{(1)} < \cdots < y_{(k)}$, and we use $\mathcal{R}_i$ to denote the risk set of individuals available to fail at time $y_{(i)}$, that is all individuals whose observed values of y, either censored or uncensored, are greater than $y_{(i)}$. Cox[35] suggested that inference for β be based on the partial likelihood

$$\prod_{i=1}^{k}\frac{\exp(x_{(i)}^T\beta)}{\sum_{j\in\mathcal{R}_i}\exp(x_j^T\beta)},$$

where $x_{(i)}$ is the vector of covariates for the individual with observed time $y_{(i)}$. This ignores the part of the likelihood that records information between failure times. It is not a marginal or conditional likelihood except in special cases, but inference based on the partial likelihood has many of the properties of inference based on the full likelihood function, including consistency and asymptotic normality, with asymptotic covariance consistently estimated by the second derivative of the log of the partial likelihood function. This has been extended to many types of processes evolving in time, and many types of incompletely observed data.

Model (23) is a semi-parametric model, and general likelihood theory for such models can be accessed through Refs 37 and 38.

Many other likelihood-like functions can be constructed using the density of just part of the data. Besag[39] proposed a pseudo-likelihood function for spatial data, composed of the product of the conditional densities of each point, conditioned on its immediate neighbors. This was one of a class of such likelihoods now often referred as composite likelihoods, after Lindsay.[40]

Example 8 One way to model correlated binary data is to start with an unobserved latent variable modeled, for example, as

$$\begin{aligned} z_{ir} &= x_{ir}^T\beta + w_{ir}^T b_i + \epsilon_{ir}, \quad b_i \sim N(0, \Sigma_b), \\ &\quad \epsilon_{ir} \sim N(0,1), \end{aligned}$$

where $r = 1, \ldots, n_i$ indexes observations in a cluster, $i = 1, \ldots, n$ indexes clusters, and x_{ir} and w_{ir} are covariates associated with the rth individual in the ith cluster. If we observe $y_{ir} = 1$ if $z_{ir} \geq 0$, the joint likelihood for y is

$$L(\theta; y) = \prod_{i=1}^{n} \log \int_{-\infty}^{\infty} \prod_{r=1}^{n_i} p_{ir}^{y_{ir}} (1 - p_{ir})^{1-y_{ir}} \phi(b_i, \Sigma_b) db_i,$$

where $p_{ir} = \Phi(x_{ir}^T\beta + w_{ir}^T b_i)$, $\Phi(\cdot)$ is the standard normal distribution function, and $\phi(\cdot; \mu, \Sigma)$ is the normal density function with mean vector μ and covariance matrix Σ. The integral in the likelihood is difficult to evaluate for models with random effect b_i of dimension more than two or three, and an alternative investigated by Renard et al.[41] in this setting is the joint likelihood of all possible pairs of observations within each cluster. This pairwise likelihood is an example of a composite likelihood. Renard et al.[41] show that inference based on the pairwise likelihood is quite efficient relative to that based on full likelihood. There is a large literature on the relative efficiency of composite likelihood methods; see Ref 42.

A somewhat different approach to the likelihood-based analysis of complex data is based on the quasi-likelihood of Wedderburn.[43] This approach starts by specifying parametric forms for the mean and variance of the response, for example

$$E(y_i \mid x_i) = \mu(x_i^T\beta), \quad \text{var}(y_i \mid x_i) = \varphi V(\mu_i),$$

where $\mu(\cdot)$ and $V(\cdot)$ are known functions, and φ is an additional scale parameter for the variance function. Inference for β is based on the estimating equation

$$\sum_{i=1}^{n} V(\mu_i)^{-1/2}\{y_i - \mu(x_i^T\beta)\} = 0,$$

which would be the score equation for a generalized linear model with these first two moments, if such a model existed. The theory of quasi-likelihood inference is developed by McCullagh.[44] This was extended to the analysis of longitudinal data by Liang and Zeger[45] under the description generalized estimating equations, or GEE. Liang and Zeger proposed using what they called a 'working covariance' function for $V(\cdot)$ and showed that the estimates of the parameters in the mean were consistent even if the working covariance function was not correct. At the time of writing the relationship between GEE methods and composite likelihood methods is not clear.

If the mean function is modeled with both fixed and random effects, as in Example 8, then this quasi-likelihood approach also involves integration. Breslow and Clayton[46] show that Laplace approximation to this integral leads to a version of penalized quasi-likelihood for generalized linear mixed models. Green[47] gives a general discussion of inference based on penalized likelihood functions, in the context where the parameter governing the distribution of the ith observation can be expressed as

$$\theta_i = x_i^T\beta + m(w_i)$$

with $m(\cdot)$ a 'smooth' function of the form (3). A different approach to quasi-likelihood estimation of variance components has been developed by Nelder and Lee; see, for example, Refs 48 and 49.

Owen[50] initiated a literature on a type of non-parametric likelihood called empirical likelihood. In the simplest case where $y_1, \ldots, y_n$ are i.i.d. from a density f, the usual non-parametric likelihood puts mass $1/n$ on each of the observations. This is not exactly a likelihood function in the strict sense since the density is not dominated by a sigma-finite measure. Owen showed that if we assume that all the possible densities for f have a common parameter, for example the mean, μ, then the empirical maximum likelihood estimator, which maximizes

$$\prod_{i=1}^{n} p_i, \quad \text{subject to } \sum p_i y_i = \mu, \text{and } \sum p_i = 1$$

is consistent and asymptotically normal, and further that likelihood ratio tests of the form (7) or (14) can be based on the empirical likelihood.

Empirical likelihood enables the use of likelihood-based arguments in a non-parametric setting. It has been extended and generalized considerably since Owen's original paper: see, for example, Ref 51.

In Owen's[50] empirical likelihood, the emphasis is on inference for a small or at least finite number of parameters that are assumed to have an appropriate interpretation without specifying the parametric form of the model. Another version of non-parametric likelihood inference is the theory of using likelihood-like arguments with parameters that are functions, for example maximum likelihood estimation of a log-concave density from an independent sample from such a density. The theory for this is considerably more complex: for recent results on consistency of such estimators see Ref 52 and references therein. Some of the theoretical work is closely related to that for semi-parametric models such as the proportional hazards model of Example 7, and a good introduction is Ref 53 (Ch. 21).

CONCLUSION

The likelihood function and derived quantities based on the likelihood function are the basis for all statistical inference based on mathematical modeling. Derived quantities based on the likelihood function provide estimates of unknown parameters, estimates of uncertainty, and methods for testing hypotheses and selecting models. The large number of extensions to likelihood suggested for tackling particular complex models arising in applications are a testament to the central role of likelihood and ideas based on likelihood in statistical inference.

REFERENCES

1. Wood S. *Generalized Additive Models: An Introduction with R*. New York: Chapman & Hall/CRC; 2006.
2. Hastie T, Tibshirani RJ, Friedman J. *The Elements of Statistical Learning*. 2nd ed. New York: Springer-Verlag; 2009.
3. Fisher RA. *Statistical Methods and Scientific Inference*. Edinburgh: Oliver & Boyd; 1956.
4. Edwards AF. *Likelihood (Expanded Edition)*. Baltimore: Johns Hopkins University Press; 1992.
5. Azzalini A. *Statistical Inference*. London: Chapman & Hall; 1998.
6. Davison AC. *Statistical Models*. Cambridge: Cambridge University Press; 2003.
7. Cox DR, Hinkley DV. *Theoretical Statistics*. London: Chapman & Hall; 1974.
8. Barndorff-Nielsen OE, Cox DR. *Inference and Asymptotics*. London: Chapman & Hall; 1994.
9. Casella G, Robert CP. *Monte Carlo Statistical Methods*. New York: Springer-Verlag; 1999.
10. Gilks WR, Richardson S, Spiegelhalter D. *Markov Chain Monte Carlo in Practice*. New York: Chapman & Hall/CRC; 1996.
11. Berger JO. *Statistical Decision Theory and Bayesian Analysis*. New York: Springer-Verlag; 1985.
12. Berger JO. The case for objective Bayes analysis. *Bayesian Stat* 2006, 1:385–402, doi:10.1214/06-BA115.
13. Goldstein M. Subjective Bayesian analysis: principles and practice. *Bayesian Stat* 2006, 1:403–420, doi:10.1214/06-BA116.
14. Venables WN, Ripley BD. *Modern Applied Statistics with S*. New York: Springer-Verlag; 2003.
15. Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc B* 1977, 39:1–38.
16. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. 2nd ed. New York: John Wiley & Sons; 2002.
17. McLachlan GJ, Krishnan T. *The EM Algorithm and Extensions*. New York: John Wiley & Sons; 2007.
18. Claeskens G, Hjort NL. *Model Selection and Model Averaging*. Cambridge: Cambridge University Press; 2008.
19. Kass RE, Wasserman L. Formal rules for selecting prior distributions: a review and annotated bibliography. *J Am Stat Assoc* 1996, 91:1343–1370.
20. Searle SR, Casella G, McCulloch CE. *Variance Components*. New York: John Wiley & Sons; 1992.
21. Fraser DAS. *Inference and Linear Models*. New York: McGraw-Hill; 1979.
22. Barndorff-Nielsen OE. On a formula for the distribution of the maximum likelihood estimator. *Biometrika* 1983, 70:343–365.
23. Fraser DAS. Likelihood for component parameters. *Biometrika* 2003, 90:327–339.
24. Cox DR, Reid N. Parameter orthogonality and approximate conditional inference (with discussion). *J R Stat Soc B* 1987, 49:1–39.
25. Diciccio TJ, Martin MA, Stern SE, Young GA. Information bias and adjusted profile likelihoods. *J R Stat Soc B* 1996, 58:189–203.

26. Chang H, Mukerjee R. Probability matching property of adjusted likelihoods. *Stat Probab Lett* 2006, 76:838–842.
27. Barndorff-Nielsen OE. Conditionality resolutions. *Biometrika* 1980, 67:293–310.
28. Cox DR. Local ancillarity. *Biometrika* 1980, 67:279–286.
29. Durbin J. Approximations for densities of sufficient statistics. *Biometrika* 1980, 67:311–333.
30. Hinkley DV. Likelihood as approximate pivotal. *Biometrika* 1980, 67:287–292.
31. Daniels HE. Saddlepoint approximations in statistics. *Ann Math Stat* 1954, 46:631–650.
32. Pace L, Salvan A. *Principles of Statistical Inference: From a Neo-Fisherian Perspective*. Singapore: World Scientific; 1997.
33. Severini TA. *Likelihood Methods in Statistics*. Oxford: Oxford University Press; 2001.
34. Brazzale AR, Davison AC, Reid N. *Applied Asymptotics*. Cambridge: Cambridge University Press; 2007.
35. Cox DR. Regression models and life tables (with discussion). *J R Stat Soc B* 1972, 34:187–220.
36. Cox DR. Partial likelihood. *Biometrika* 1975, 62:269–276.
37. Murphy SA, van der Vaart AW. On profile likelihood (with discussion). *J Am Stat Assoc* 2000, 95:449–485.
38. Murphy SA, van der Vaart AW. Semiparametric likelihood ratio inference. *Ann Stat* 1997, 25:1471–1509.
39. Besag JE. Spatial interaction and the statistical analysis of lattice systems (with discussion). *J R Stat Soc B* 1974, 34:192–236.
40. Lindsay BG. Composite likelihood methods. *Contemp Math* 1988, 80:220–239.
41. Renard D, Molenberghs G, Geys H. A pairwise likelihood approach to estimation in multilevel probit models. *Comput Stat Data Anal* 2004, 44:649–667.
42. Varin C. On composite marginal likelihoods. *Adv Stat Anal* 2008, 92:1–28.
43. Wedderburn RWM. Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika* 1974, 61:439–447.
44. McCullagh P. Quasi-likelihood functions. *Ann Stat* 1983, 11:59–67.
45. Liang K-Y, Zeger S. Longitudinal data analysis using generalized linear models. *Biometrika* 1986, 73:13–22.
46. Breslow N, Clayton D. Approximate inference in generalized linear mixed models. *J Am Stat Assoc* 1993, 88:9–25.
47. Green PJ. Penalized likelihood for general semi-parametric regression models. *Int Statist Rev* 1987, 55:245–259.
48. Lee Y, Nelder JA. Hierarchical generalised linear models: a synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika* 2001, 88:987–1006.
49. Nelder JA, Lee Y. Likelihood, quasi-likelihood and pseudolikelihood: some comparisons. *J R Stat Soc B* 1992, 54:273–284.
50. Owen AB. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* 1988, 75:237–249.
51. Hjort NL, McKeague IW, van Keilegom I. Extending the scope of empirical likelihood. *Ann Stat* 2009, 37:1079–1111.
52. Balabdaoui F, Rufibach K, Wellner JA. Limit distribution theory for maximum likelihood estimation of a log-concave density. *Ann Stat* 2009, 37:1299–1331.
53. van der Vaart AW. *Asymptotic Statistics*. Cambridge: Cambridge University Press; 1998.