Marginal and Conditional Likelihoods

Author(s): J. D. Kalbfleisch and D. A. Sprott

Source: *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, Sep., 1973, Vol. 35, No. 3, Dedicated to the Memory of P. C. Mahalanobis (Sep., 1973), pp. 311-328

Published by: Indian Statistical Institute

Stable URL: https://www.jstor.org/stable/25049882

MARGINAL AND CONDITIONAL LIKELIHOODS

By J. D. KALBFLEISCH[1]

State University of New York at Buffalo

and

D. A. SPROTT[2]

University of Waterloo

SUMMARY. Marginal and conditional likelihoods are illustrated on various problems that have occurred in the literature. They can thus be compared with the more conventional exact and asymptotic procedures that have been used on these problems. Of particular interest is the question of what, if any, information of the data is sacrificed in the use of marginal and conditional likelihoods. This question is discussed and some aspects of it are illustrated on the examples presented. More needs to be done however in this respect.

1. Introduction

Kalbfleisch and Sprott (1970) outlined and discussed several methods of eliminating a set of nuisance parameters β from a likelihood function so that inferences can be made about the parameter θ of interest. In some instances, when the problem possesses certain structures, the methods of marginal and conditional likelihood are available and these were discussed in some detail. Roughly speaking, marginal likelihood and conditional likelihood involve separating the information in the data into two parts by means of ancillary and sufficient statistics for β respectively. The one part of the data yields a likelihood for the parameter θ of interest which does not involve the nuisance parameter β; its interpretation is therefore straightforward and free from difficulties arising from our ignorance about β. The remaining part of the data yields a joint likelihood for β and θ, but any information contained in θ is inextricably tied up with the unknown β. In certain special cases, this second factor can be said to contain no information about θ when β is completely unknown. The first factor then contains all the information in the sample about θ and is called the marginal or conditional likelihood of θ according as the above separation is accomplished via ancillary or sufficient statistics. In other more general cases the second factor will contain information about θ. Confining attention to the first factor for inferences about θ sacrifices this information and the resulting likelihoods are called approximate marginal or conditional likelihoods. An important consideration in this case is to evaluate the amount of information lost in ignoring the second factor.

This work was supported in part by :

[1]PHS Training Grant No. CA 05227-01, from the National Cancer Institute and by

[2]National Research Council of Canada.

The main purpose of this paper is to give diverse examples of the applications of exact and approximate marginal and conditional likelihoods to problems that have been discussed in the literature. In this way, not only can the performance of these likelihoods be examined, but the results can be compared with those of other more conventional procedures. Intrinsic in the concept of exact and approximate marginal and conditional likelihoods is the concept of "no (or little) information about θ in the absence of knowledge of β". A secondary purpose of this paper is to examine this concept in several examples and to discuss methods by which its applicability can be checked in specific examples. The second section deals specifically with conditional likelihoods and the problem of assessing the information lost through ignoring the residual factor. This has particular relevance to other conditional methods of testing and estimation which have been proposed (see for example Cox, 1970, p. 46).

For completeness, the definitions of the marginal likelihood function, the conditional likelihood function and the maximum relative likelihood function are briefly stated here. These formulae, along with their justifications and the assumptions involved, are more extensively discussed in Kalbfleisch and Sprott (1970).

1.1. *The conditional likelihood function.* Suppose that the random variate $x = (x_1, \ldots, x_n)$ has probability element given by $f(x; \theta, \beta)dx$ and that for a given value of θ, $T_\theta = (T_1, \ldots, T_k)$ is sufficient for the estimation of β where $T_i = T_i(x_1, \ldots, x_n, \theta)$ can be a function of θ. The conditional likelihood function of θ is

$$C(\theta) \propto f(x; \theta, \beta)dx / f(T_\theta; \theta, \beta)dT_\theta = f(x; \theta \mid T_\theta)(dx/dT_\theta), \qquad \ldots \quad (1)$$

provided the marginal distribution of T_θ contains no information about θ when β is unknown. If $z_1 = z_1(x_1), \ldots, z_n = z_n(x_n)$ are the Cartesian coordinates of a point in Euclidean n-space, this expression becomes

$$C(\theta) \propto f(x_1, \ldots, x_n; \theta \mid T_\theta) / \sqrt{|J\,J'|} \qquad \ldots \quad (2)$$

where J is the $k \times n$ Jacobian matrix $(\partial T_i / \partial z_j)$. The derivaion of (2) is considered in the appendix. If $T_1, \ldots, T_k$ are not functions of θ, then (1) may be interpreted directly since J does not depend on θ.

1.2. *The marginal likelihood function.* Suppose that $a_\theta = (a_1, \ldots, a_{n-r})$ is, for a given value of θ, ancillary for β. The marginal likelihood of θ is

$$M(\theta) \propto f(a_\theta; \theta)da_\theta, \qquad \ldots \quad (3)$$

provided the conditional distribution of the data given a_θ contains no information on θ when β is unknown. If

$$z_i = z_i(x_i) = z_i(u_1, \ldots, u_r, a_1, \ldots, a_{n-r}) \quad i = 1, \ldots, n, \qquad \ldots \quad (4)$$

is a non-singular tranfsormation from $z_1, \ldots, z_n$ to $u_1, \ldots, u_r, a_1, \ldots, a_{n-r}$, then under the assumption that $z_1, \ldots, z_n$ are the Cartesian coordinates of a point in Euclidean n-space, (3) becomes

$$M(\theta) \propto f(a_\theta; \theta) \sqrt{|K'K|} / |L| \qquad \ldots (5)$$

where $|L|$ is the Jacobian of the transformation (4) and K is the $n \times r$ matrix $(\partial z_i / \partial u_j)$. Fraser (1967) gave a definition of marginal likelihood and considered the first examples of this kind. Additional examples are found in Fraser (1968), and Kalbfleisch and Sprott (1970). Again, if $a_1, \ldots, a_{n-r}$ are not functions of θ, (3) may be interpreted directly since then K and L do not depend on θ.

1.3. *The maximum relative likelihood function.* The maximum relative likelihood function of θ arising from the joint likelihood $l(\theta, \beta; x)$ of θ, β is

$$R_M(\theta) = \sup_{\beta} l(\theta, \beta; x) / \sup_{\theta, \beta} l(\theta, \beta; x). \qquad \ldots (6)$$

Sections 2 and 3 give simple examples of conditional and marginal likelihoods where the sufficient and the ancillary statistics are not functions of θ. Section 4 considers the more complicated case. Comparisons with the maximum relative likelihood (6) are also made.

In the above two definitions, the coordinates $(z_1, \ldots, z_n)$ of the sample point are specified as being the Cartesian coordinates of a point in Euclidean space. This specification defines the metric on the sample space as being Euclidean with the element of distance, ds, given by

$$ds^2 = \Sigma dz_i^2.$$

In Section 4, the results depend on what metric is specified or, equivalently, what variates are assumed to form Cartesian coordinates in a Euclidean space.

2. Conditional likelihood

Example 2.1 : *A model in virology.* A Poisson model arises in virology from which exact conditional likelihoods can be derived. It is required to estimate h "the hit number" from the observed plaque counts $n_0, n_1, \ldots, n_k$ where n_i has a Poisson distribution with mean $\beta\theta^i$ and $\theta = d^{-h}$. Here d is the known dilution factor and $i = 0, 1, \ldots, k$ are the dilution levels (cf. Alling, 1971).

The joint distribution of $n_0, n_1, \ldots, n_k$ is

$$\prod_{i=0}^{k} (\beta\theta^i)^{n_i} \exp(-\beta\theta^i)/n_i!.$$

The distribution of $n = \Sigma n_i$ is Poisson with mean $\phi = \beta(1-\theta^{k+1})/(1-\theta)$ so that the conditional distribution of $n_0, n_1, \ldots, n_k$ given n is

$$n!(1-\theta)^n\theta^T/[(1-\theta^{k+1})^n\Pi n_i!] \quad \ldots (7)$$

where $T = \Sigma i n_i$.

Since the marginal distribution of n depends only on ϕ and since ϕ and θ are functionally dependent (i.e. ϕ is a 1 : 1 function of β for each θ so that θ is not identifiable) it follows that the marginal distribution of n contains no information about θ when β is completely unknown. Thus (7) is a complete summary of the information about θ contained in the data and is proportional to the conditional likelihood of θ. A similar example arising in time dependent Poisson Processes has been discussed by Cox and Lewis (1966, p. 46).

Example 2.2 : *Approximate conditional likelihoods in contingency tables.* The 2×2 contingency table provides a simple example of (approximate) conditional likelihoods. Suppose that treatment A is tested n times giving x successes and $(n-x)$ failures while treatment B gives y successes in m trials. If the probabilities of success are β and $\theta\beta/(1-\beta+\theta\beta)$ for treatments B and A respectively, then for any specified θ (the odds ratio), the marginal total $t = x+y$ is sufficient for β. The conditional distribution of x given t is a function of θ only and is available for inference about θ.

Unlike Example 2.1, the use of this conditional distribution for inference about θ entails some loss of information about θ. Indeed, the marginal distribution of t, which is the residual factor not used in the inference, is

$$f(t;\theta,\beta) = \frac{\beta^t(1-\beta)^{m+n-t}}{(1-\beta+\theta\beta)^n}\sum_i \binom{n}{i}\binom{m}{t-i}\theta^i. \quad \ldots (8)$$

Here the residual factor (8) is not a function of a single parameter $\phi(\theta,\beta)$ as was the case in Example 2.1. It would therefore appear that there is information about θ in (8) which is ignored in using only the conditional distribution; the likelihood arising from this conditional distribution is therefore called an approximate conditional likelihood.

In order to determine the amount of information lost, it is necessary to examine the residual factor (8). Information about θ is tied up with information about the unknown β and it is difficult to quantify or ascertain what is lost. One possibility is to examine the maximum relative likelihood of θ arising out of (8), i.e.

$$R_M(\theta) \propto \frac{\hat{\beta}^t(1-\hat{\beta})^{m+n-t}}{(1-\hat{\beta}+\theta\hat{\beta})^n}\sum_i \binom{n}{i}\binom{m}{t-i}\theta^i \quad \ldots (9)$$

where $\hat{\beta} = \hat{\beta}(\theta)$ is the maximum likelihood estimate of β for the specified θ. The approximate conditional likelihood is

$$C(\theta) \propto \theta^x / \sum_i \binom{n}{i} \binom{m}{t-i} \theta^i. \qquad \ldots \quad (10)$$

The conditional relative likelihood function

$$CR(\theta) = C(\theta) / \sup_\theta C(\theta)$$

and the residual maximum relative likelihood function of θ (9) are compared (Fig. 1) for the data cited by Bliss (1967, p. 65, Table 4.8).

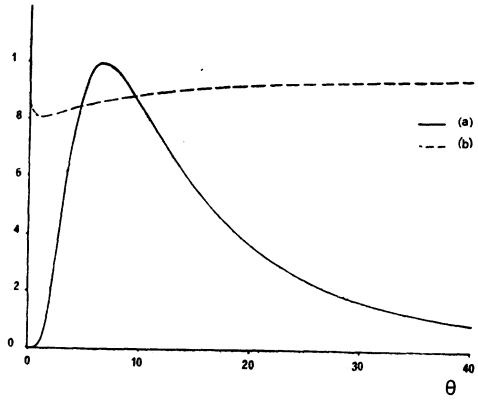

Figure 1. (a) Conditional Relative Likelihood
(b) Residual maximum relative likelihood arising from the 2×2 contingency table data cited in Example 2.2.

The maximum relative likelihood arising from the residual term is nearly constant over the range of θ and quite close to 1. The "absence of knowledge of β" could be interpreted to mean that we are free to pick or examine any value of β we please, no one value having precedence over any other. From $R_M(\theta)$, it can be seen that for each θ we can select a value of β (namely $\hat{\beta}(\theta)$) that will make that value of θ plausible given only the residual term (9). From this point of view, $R_M(\theta)$ and hence (8) provides little information concerning θ in the absence of knowledge of β and essentially nothing is lost through restricting attention to the approximate conditional likelihood (10).

The "play the winner rule" (Zelen, 1969) provides a similar example involving waiting time (negative binomial) distributions. The results in this case differ from those obtained in the 2×2 contingency table and this underlines the fact that the specification of the sample space plays an important role in these methods unlike the conventional use of the entire likelihood function.

Example 2.3 : *Approximate conditional likelihoods arising from a negative binomial distribution.* Suppose that $x_{i1}, \ldots, x_{im_i}$ is a sample from a negative binomial distribution with parameters (β_i, θ). The distribution of $x_{i1}, \ldots x_{im_i}$ is

$$\beta_i^{t_i} (1-\beta_i)^{m_i \theta} \prod_{j=0}^{\infty} \binom{\theta+j-1}{j}^{a_{ij}}$$

where a_{ij} is the frequency with which j is observed, and $t_i = \Sigma j a_{ij}$, $m_i = \Sigma a_{ij}$. For any specified θ, t_i is sufficient for β_i and has a negative binomial distribution with parameters $(\beta_i, m_i\,\theta)$. If n such samples are taken from n negative binomial distributions, the conditional distribution of the data given $t_1, \ldots, t_n$ is

$$\prod_{i=1}^{n}\left[\prod_{j=1}^{m_i}\binom{\theta+j-1}{j}^{a_{ij}}\right]\Big/\binom{m_i\theta+t_i-1}{t_i} \qquad \ldots\ (11)$$

which is proportional to the approximate conditional likelihood of θ, $C(\theta)$. Numerical examples of (11) have been given by Sprott (1970).

The use of (11) for inference about θ involves the sacrifice of the information on θ in the residual term arising from the marginal distribution of $t_1, \ldots, t_n$,

$$\prod_{i=1}^{n}\binom{m_i\theta+t_i-1}{t_i}\beta_i^{t_i}(1-\beta_i)^{m_i\theta}. \qquad \ldots\ (12)$$

As before, some idea of the information lost through the use of (11) is obtained by comparing the graphs of the approximate conditional relative likelihood $CR(\theta)$ from (11) with the maximum relative likelihood function $R_M(\theta)$ arising from (12). This is done (Fig. 2) for the data of Fisher (1941, Example 1) in which $n = 1$.

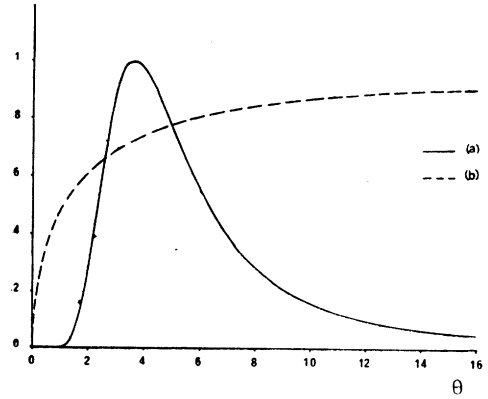

Figure 2. (a) Conditional Relative Likelihood
(b) Residual maximum relative likelihood arising from the negative binomial distribution data cited in Example 2.3.

From this it would seem that $R_M(\theta)$ does not contradict $CR(\theta)$ and further $R_M(\theta)$ does not vary rapidly when θ is not too small. Again, any specified value of θ (not too small) can be made very plausible by appropriate choices of β_i in (12).

This example and the preceding ones are special cases of the power series distribution, where $x_{i1}, \ldots, x_{im_i}$ are independent observations on a discrete random variate x_i with probability function

$$f(x_i;\, \beta_i, \theta) = a_{ix_i}(\theta)\,\beta_i^{x_i}/\psi_i(\beta_i, \theta) \quad x_i = 0, 1, 2, \ldots$$

where

$$\psi_i(\beta_i, \theta) = \sum_{r=0}^{\infty} a_{ir}(\theta)\,\beta_i^r, \quad 0 < \beta_i < \infty$$

and $a_{ir}(\theta) \geqslant 0$ for all θ and r, $(i = 1, 2, \ldots, n)$. This is the generalized power series distribution (see Patil and Joshi, 1968, page 23). The probability of the sample $(x_{ij}; j = 1, \ldots, m_i;\ i = 1, \ldots, n)$ is

$$\prod_{i=1}^{n}\Big[\prod_{j=1}^{m_i} a_{ij}(\theta)\Big]\,\beta_i^{t_i}/\psi_i^{m_i}(\beta_i, \theta)$$

so that $t_1, \ldots, t_n$ is sufficient for $\beta_1, \ldots, \beta_n$ where $t_i = \Sigma_j x_{ij}$. Here in general one may obtain a conditional likelihood function of θ by considering the distribution of the data given $t_1, \ldots, t_n$. This yields a class of models similar to the negative binomial discussed above. The next example considers matched pairs from a power series distribution from which a number of cases of importance arise.

Example 2.4 : *Matched pairs.* Suppose that x_i and y_i are independent random variates with power series distributions

$$f(x_i; \beta_i) = a_{x_i}\beta_i^{x_i}/\psi(\beta_i)$$

$$f(y_i; \theta, \beta_i) = a_{y_i}(\theta\beta_i)^{y_i}/\psi(\theta\beta_i)$$

respectively, $i = 1, \ldots, n$ independently. This is a matched pairs problem with β_i the parameter specific to the pair and θ giving the common ratio between members of each pair. It is easily seen that $t_i = x_i + y_i$ is sufficient for β_i when θ is known and that its distribution is

$$g(t_i; \theta, \beta_i) = c_{t_i}(\theta)\beta_i^{t_i}/\psi(\beta_i)\psi(\theta\beta_i) \qquad \ldots \quad (13)$$

where $$c_{t_i}(\theta) = \sum_{r=0}^{t_i} a_r\, a_{t_i - r}\theta^r.$$

The conditional probability of the data given $t_1, \ldots, t_n$ is

$$\prod_{i=1}^{n} a_{y_i}\, a_{t_i - y_i}\theta^{y_i}/c_{t_i}(\theta) \qquad \ldots \quad (14)$$

which is a function of θ only. In using (14) to make an inference about θ, whatever information is contained in (13) about θ is ignored.

In some very special instances it can be shown that (13) contains no information about θ when β_i is unknown. Consider for example the Poisson distribution where $\psi(\beta_i) = e^{\beta_i}$. In this case, the conditional likelihood is (from 14)

$$c(\theta) \propto \prod_{i=1}^{n} \binom{t_i}{y_i} \theta^{\Sigma y_i}/(1+\theta)^{\Sigma t_i} \qquad \ldots \quad (15)$$

while the residual term has components (from 13)

$$[\beta_i(1+\theta)]^{t_i} e^{-\beta_i(1+\theta)}/t_i! \qquad \ldots \quad (16)$$

which involves the parameters only through $\phi_i = \beta_i(1+\theta)$. Thus, as in Example 2.1, (16) contains no information about θ when β_i is unknown and (15) is an exact conditional likelihood of θ. This example was cited by Cox (1958) for the case $n = 1$.

More usually, as in Examples 2.2 and 2.3, (13) will contain some information about θ and (14) will be only an approximate conditional likelihood. Consider for example, the case of matched pairs from a binomial distribution with $x_i = 0$ or 1, (success or failure) where β_i is the odds of success vs. failure for the first member of the pair while $\theta\beta_i$ is the odds for the second (and so $\psi(\beta) = (1+\beta)$). In this case (14) is

$$\prod_{i \in S} \{\theta^{y_i}/1+\theta\} \qquad \ldots \quad (17)$$

where $S = \{i \mid t_i = 1\}$. This distribution has been suggested for inference about θ by several authors (see for example Cox, 1970, p. 56). The residual term (13) being ignored has i-th component.

$$\frac{1}{(1+\beta_i)(1+\theta\beta_i)}, \qquad t_i = 0, \qquad \ldots \quad (18a)$$

$$\frac{\beta_i(1+\theta)}{(1+\beta_i)(1+\theta\beta_i)}, \qquad t_i = 1, \qquad \ldots \quad (18b)$$

$$\frac{\theta\beta_i^2}{(1+\beta_i)(1+\theta\beta_i)}, \qquad t_i = 2. \qquad \ldots \quad (18c)$$

In order to examine what information is lost through ignoring the residual term, we can, as before, examine the maximum relative likelihood function arising from the term being ignored. In this case, it is easily shown that this maximum relative likelihood function is unity for cases (a) and (c), while for (b) is $(1+\theta)/(1+\sqrt{\theta})^2$. Thus the entire maximum relative likelihood arising from (18) is

$$R_M(\theta) = \left\{\frac{(1+\theta)}{(1+\sqrt{\theta})^2}\right\}^{T_1} \qquad \ldots \quad (19)$$

where T_1 is the number of t_i's observed equal to 1. This function attains a minimum at $\theta = 1$ of $(1/2)^{T_1}$ and maximum value of 1 at $\theta = 0$ and $\theta = \infty$.

A typical result of a matched pairs experiment of this type will therefore give an approximate conditional likelihood from (17) with a unique maximum at

$$\hat{\theta} = \sum_{i \varepsilon S} y_i / (T_1 - \sum_{i \varepsilon S} y_i)$$

while the residual term is U-shaped with maxima at 0 and ∞. It might appear, therefore, that (19) could contradict (17) and this underlines the difficulty in determining what information is being lost in using an approximate conditional likelihood. In the present case, it might be noted that to estimate θ with any precision one would want an experiment which gave several "discordant" pairs, i.e. an experiment in which T_1 was reasonably large. If the curves arising from (19) are examined for different values of T_1 one finds that possible curves are very limited; a fairly substantial change in T_1 changes the maximum relative likelihood function (19) only very slightly. To this extent, knowledge about T_1 contributes little to our knowledge about θ in the sense that large changes in T_1 would produce only small changes in the maximum relative likelihood function or in our inference on θ from (19). This suggests that in evaluating the information lost, not only is the observed $R_M(\theta)$ a relevant consideration, but so also is the variability of the family of possible $R_M(\theta)$'s.

This use of the maximum relative likelihood function $R_M(\theta)$ arising from (18) may seem to the reader rather artificial. Our motivation in these considerations arises from analogous calculations in a normal distribution. Suppose a single observation x is selected from $N(\beta, \sigma^2)$ where $-\infty < \beta < \infty$ and $0 < \varepsilon < \sigma < \infty$, and we wish to determine the amount of information on σ^2 contained in this datum. The maximum relative likelihood function of σ^2 is

$$R_M(\sigma^2) = \frac{\varepsilon}{\sigma} .$$

Although this gives a fairly sharp zoning of values for σ^2, the family of possible $R_M(\sigma^2)$'s is very limited, consisting in fact of only one possibility. The fact that the shape of $R_M(\sigma^2)$ is known even before x is observed makes $R_M(\sigma^2)$ uninformative about σ^2 (and in fact demonstrates the inappropriateness of its use in estimating σ^2). In example 2.4, the family of possible $R_M(\theta)$'s is very limited and the argument is an approximation to that used in the normal case. It is of interest to note that had it been decided to terminate the experiment when the T_1-th discordant pair was observed, the normal situation would be exactly reproduced; in this case, the shape of $R_M(\theta)$ arising from the residual term would be known before the data were recorded.

Examples 2.3 and 2.4 have the characteristic that introducing additional observations also introduces additional parameters; the parameters in these models have been called incidental and structural by Neyman and Scott (1948). Anderson (1970) has defined the conditional likelihood estimate as the value of θ which maximizes an exact or approximate conditional likelihood function. He has shown in

models of this kind that these estimates (for the structural parameters) are consistent and asymptotically normally distributed under regularity conditions. It should be noted that the maximum likelihood estimate itself need not be consistent in such cases.

3. EXAMPLES OF MARGINAL LIKELIHOODS

Marginal likelihoods are more restrictive than conditional likelihoods, but as will be seen, in some instances are intuitively more appealing. They are more restrictive because they depend on the existence of ancillary statistics which usually are not present in discrete models. The examples in Section 2 could not be treated by marginal likelihoods for this reason. However, in the more complicated continuous examples discussed in Section 4, marginal likelihoods often produce more intuitively appealing results than the analogous conditional likelihoods.

3.1 : *The guaranteed exponential distribution.* Suppose $x_1, \ldots, x_n$ is a random sample from a distribution with probability density function

$$f(x; \theta, \beta) = \frac{1}{\theta} \exp\{-(x-\beta)/\theta\} \quad \beta < x < \infty$$

where $0 < \theta < \infty$ and $-\infty < \beta < \infty$. If $x_{(i)}$ is the i-th order statistic, it is well known that

$$a_i = x_{(i)} - x_{(1)} \quad i = 2, \ldots, n$$

are $(n-1)$ functionally independent ancillary statistics for θ. Further, the group of location transformations acts transitively on $x_{(1)}$ while leaving the subspace defined by $a_2, \ldots, a_n$ fixed. The homomorphic group acting on the parameter space is transitive on β while leaving θ fixed. By the definition of Barnard (1963), $a_2, \ldots, a_n$ are "sufficient for θ in the absence of knowledge of β" or equivalently the conditional distribution of $x_{(1)}$ given $a_2, \ldots, a_n$ contains no information on θ when β is unknown. The marginal likelihood arises, therefore, through the distribution of $a_2, \ldots, a_n$ and is

$$M(\theta) \propto \frac{1}{\theta^{n-1}} \exp\left\{-\frac{\Sigma a_i}{\theta}\right\}. \qquad \ldots \quad (20)$$

This same result can be obtained as a conditional likelihood of θ.

Approximate marginal likelihoods are often advantageous. For example, in most practical instances, β is known to be positive and this restriction removes the group structure. The use of (20) to estimate θ will entail some loss of information. Here again, the residual term should be examined to ascertain that the information lost is not too great or in contradiction to the marginal likelihood.

This example could be extended to the case where samples are drawn from each of n guaranteed exponentials with parameters (θ, β_i) $i = 1, \ldots, n$. Similar examples involving incidental and structural parameters occur in estimating the varince σ^2 from pairs (x_i, y_i) of $N(\mu_i, \sigma^2)$ variates, and in estimating ρ from a sample of k-variate normal vectors X with distribution $N(\mu, \sigma' \rho \sigma)$ where μ is $(k \times 1)$, ρ is $k \times k$ symmetric with $\rho_{ii} = 1$ and $\sigma = \text{diag}(\sigma_1, \ldots, \sigma_k)$ (c.f. Barnard, 1969). Another interesting example of marginal likelihood where the ancillaries are not functions of θ is discussed by Kalbfleisch and Prentice (1972). In what follows, examples of marginal and conditional likelihoods in which the statistics involve the parameter θ are discussed.

4. Conditional and marginal likelihoods : differentials involving the parameters

Example 4.1 : *Weibull distribution.* Suppose that x has pobability density function

$$f(x; \theta, \beta) = \theta\beta x^{\theta-1} \exp\{-\beta x^{\theta}\} \quad 0 < x < \infty \qquad \ldots \quad (21)$$

where $0 < \theta < \infty$ and $0 < \beta < \infty$. Such a distribution could be considered to have arisen from a family of parametric transformations of an observed response with an underlying exponential distribution similar to those studied by Box and Cox (1964) for an underlying normal distribution. In the model (21), both marginal and conditional likelihoods can be obtained for the parameter θ.

Suppose that $x_1, \ldots, x_n$ are n independent observations on x; their joint probability element is

$$\theta^n \beta^n \Pi x_i^{\theta-1} \exp(-\beta T_\theta) \Pi dx_i \qquad \ldots \quad (22)$$

where $T_\theta = \Sigma x_i^\theta$ is sufficient for β for any given θ. The sufficient statistic is a function of the parameter θ of interest and in order to obtain the conditional likelihood function of θ it is necessary to specify the metric of the sample space as in Section 1.1. In this case, it seems appropriate in applying (2) to take $z_i = z_i(x_i) = \log x_i$, $i = 1, \ldots, n$ as forming the Euclidean space. This preserves the symmetries of the problem, under which the group $x_i \to x_i^c$, $\theta \to \theta/c$ leaves the problem invariant. It is easily verified that, under this assumption, the conditional likelihood is from (2)

$$C(\theta) \propto \theta^{n-1} \Pi x_i^{\theta-1} / (T_\theta^{n-1} \sqrt{\Sigma x_i^{2\theta}}). \qquad \ldots \quad (23)$$

In a similar manner, the marginal likelihood of θ can be obtained by noting that the random variates

$$a_i = x_i^\theta / T_\theta \quad i = 1, \ldots, n$$

constitute $n-1$ functionally independent ancillary statistics for β for any given θ. The marginal likelihood again requires the specification of the metric as above and the marginal likelihood function is from (5)

$$M(\theta) \propto \theta^{n-1} \Pi x_i^{\theta-1} / T_\theta^n. \qquad \ldots \quad (24)$$

It can be argued that no information is lost in restricting attention to (23) or to (24). It is easily seen, in this example, that the ancillaries $a_2, \ldots, a_n$ are statistically independent of the random variate T_θ. Thus, both the marginal and conditional likelihoods involve ingoring the information contained in the marginal distribution of T_θ. If we consider the group of transformation $x_i \rightarrow x_i^c$ the homomorphic group acting on the parameter space takes $\beta \rightarrow c^\theta \beta$. Thus β and T_θ can be made arbitrary while θ remains fixed. It follows, then, that T_θ can contain no information on θ. The conditional and marginal likelihoods in (23) and (24) are therefore exact.

The marginal likelihood (24) in this example seems intuitively more appealing than the conditional likelihood because of its greater simplicity and because the same result (24) is obtained over a wide class of metrics. For example if $z_i = g(x_i)$, $i = 1, \ldots, n$ are assumed to form the Cartesian coordinates, the result remains unchanged provided g is strictly monotone.

The maximum relative likelihood function, from (6), is proportional to

$$\theta^n \Pi x_i^{\theta-1} / T_\theta^n \qquad \ldots \quad (25)$$

in this example. The functions (23), (24) and (25) are compared (Fig. 3) using the data (Table 1) of Plait (1962). In this plot, each likelihood is standardized to have maximum value unity.

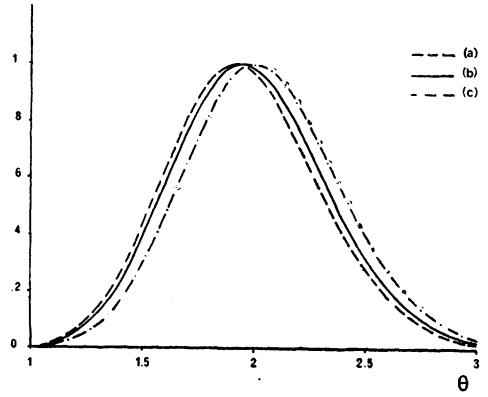

Figure 3. (a) Conditional Relative Likelihood
(b) Marginal Relative Likehood
(c) Maximum relative likelihood arising from the Weibull distribution data of Example 4.1.

It appears that (24) cannot be obtained as a conditional likelihood by a suitable choice of metric; thus, unlike the examples considered by Kalbfleisch and Sprott (1970), the marginal and conditional likelihoods are not the same in this example. The difference arises since in the former case the differentials are interpreted by considering the volume element fixed in the subspace orthogonal to that defined by

$$a = \text{const} \quad i = 1, \ldots, n \qquad \ldots \quad (26)$$

TABLE 1. OBSERVATIONS FROM A WEIBULL DISTRIBUTION

1.90	3.34	3.65	4.20	4.72	5.89	6.10	6.62	7.92	8.40
8.50	9.00	9.60	11.02	11.95	12.40	13.03	13.42	18.07	20.63

while in the latter, the volume element in the subspace defined by

$$T_\theta = \text{constant} \qquad \ldots \quad (27)$$

is considered fixed. The two likelihoods will be the same whenever the subspace defined by (26) and (27) are orthogonal. In the present example, this is not the case.

Example 4.2 : *Non-linear regression.* Suppose y_α has the normal distribution

$$N\left(\sum_{i=1}^{k} \beta_i f_{i\alpha}, \sigma^2\right) \qquad \ldots \quad (28)$$

where $$f_{i\alpha} = f_{i\alpha}(x_{i\alpha}, \theta).$$

For a specified θ, the sufficient statistics for β, σ^2 are the maximum likelihood estimates

$$B_\theta' = (b_1^{(\theta)}, \ldots, b_k^{(\theta)}), \quad S_\theta^2 = (Y - F_\theta' B_\theta)'(Y - F_\theta' B_\theta)$$

where $$B_\theta = A_\theta^{-1} C_\theta, \; A_\theta = F_\theta F_\theta', \; C_\theta = F_\theta Y \text{ and } F_\theta = (f_{i\alpha}).$$

Then J of (2) is the $(k+1)\times n$ matrix

$$J = \begin{pmatrix} \partial B_\theta / \partial Y \\ \partial S_\theta^2 / \partial Y \end{pmatrix} = \begin{pmatrix} A_\theta^{-1} F_\theta \\ 2(Y' - B_\theta' F_\theta) \end{pmatrix}$$

so that

$$|J\,J'| = 4(Y' - B_0' F_\theta) Y / |A_\theta| = 4 S_\theta^2 / |A_\theta|.$$

Here again, a group invariance argument simlar to that used in Example 4.1 shows that no information is lost in the estimation of θ through ignoring the joint marginal distribution of B_θ, S_θ^2. By substitution in (2), the conditional likelihood is found to reduce to

$$C(\theta) \propto \left(\frac{1}{S_\theta}\right)^{n-k-1} \qquad \ldots \quad (29)$$

This model (28) also gives rise to a marginal likelihood function of θ. The observations can be expressed in the form

$$Y = F_\theta' B_\theta + D_\theta S_\theta$$

where $$D_\theta = (Y - F_\theta' B_\theta)/S_\theta$$

is the vector of standardized residuals. D_θ is ancillary for β, σ^2 and again, group invariant arguments indicate that the distribution of B_θ and S_θ^2 given D_θ contains no information about θ when β and σ^2 are unknown. The marginal likelihood therefore arises from the marginal distribution of D_θ. The matrix K in (5) is

$$K = (F_\theta', D_\theta)$$

and since

$$D_\theta' D_\theta = 1 \text{ and } F_\theta D_\theta = 0$$

it follows that

$$\sqrt{|K'K|} = \sqrt{|A_\theta|}.$$

By substitution into (5), it is easily seen that the marginal likelihood reduces to

$$M(\theta) \propto \left(\frac{1}{S_\theta}\right)^{n-k-1}$$

In this example, the maximum relative likelihood function of θ (from 6) is proportional to $\left(\frac{1}{S_\theta}\right)^n$. The difference between this and the marginal and conditional likelihoods would be large if the number of parameters k eliminated is large compared to the number of observations n. The marginal and conditional likelihoods take account of the loss of degrees of freedom in estimating β; the maximum relative likelihood ignores this and so can be overprecise.

A special case of the above model was considered by Williams (1962) in which $f_{1\alpha} = 1$ (all α), $f_{2\alpha} = \exp(-\theta x_\alpha)$. Another example of the above model is

$$f_{1\alpha} = 1 \quad f_{2\alpha} = x_\alpha - \lambda \quad f_{3\alpha} = (x_\alpha - \lambda) \quad \tanh\left(\frac{x_\alpha - \lambda}{\theta}\right)$$

discussed by Bacon and Watts (1971) using Bayesian methods. The conditional and marginal likelihood of (θ, λ) is similar to the posterior distribution they obtain for (θ, λ) without the determinantal factor. Since, as they state, this determinantal factor is almost constant, the posterior distribution exhibits essentially the same features as the conditional likelihood. For their example 1, the maximum of the posterior density occurs at $\theta = .61$, $\lambda = .054$, while the corresponding conditional maximum likelihood estimate is $\theta = .622$, $\lambda = .047$.

ACKNOWLEDGEMENTS

We would like to thank Professor D. A. S. Fraser for having read a previous version of the manuscript and for substantial suggestions for revision. We would also like to thank Professor G. A. Barnard for extensive discussions and suggestions and Mr. C. Minder for checking the numerical computations and graphs.

Appendix

Suppose the square of the differential of arc ds^2 in an n-dimensional space can be written as a positive definite quadratic form of the independent differentials $dx_1, dx_2, \ldots, dx_n$ of the coordinates :

$$ds^2 = \Sigma g_{ij} dx_i dx_j = dX' \, g \, dX. \qquad \ldots \quad \text{(A-1)}$$

Here, $g = (g_{ij})$ is a positive definite symmetric matrix and $dX' = (dx_1, dx_2, \ldots, dx_n)$. The volume element in this space is defined as

$$dV = \sqrt{|g|} \Pi dx_i, \qquad \ldots \quad \text{(A-2)}$$

(Sokolnikoff, 1964, p. 204), where $|g|$ is the determinant of g.

Suppose that $x_1, \ldots, x_n$ are the Cartesian coordinates of a point in a Euclidean n-space and that the subspace of dimension $n-k$ is expressed in the implicit form

$$T_i = T_i \, (x_1, \ldots, x_n) = \text{const} \quad i = 1, \ldots, k \qquad \ldots \quad \text{(A-3)}$$

Theorem 1 : *Let S_T be the subspace of dimension $n-k$ defined by* (A-3). *Let J be the $k \times n$ Jacobian matrix of rank k*

$$J = (\partial T_i / \partial x_j)$$

and let

$$J = (J_1, J_2)$$

where J_1 and J_2 are $k \times k$ and $k \times (n-k)$ matrices respectively and J_1 is non-singular. Then the volume element in S_T is

$$dV_T = \sqrt{|J_2'(J_1 J_1')^{-1} J_2 + I|} \prod_{i=k+1}^{n} dx_i. \qquad \ldots \quad \text{(A-4)}$$

Proof : Since from (A-3) the T_i are constant in the subspace S_T, the differentials satisfy

$$\sum_{j=1}^{n} (\partial T_i / \partial x_j) dx_j = 0, \quad i = 1, \ldots, k$$

$$dx_{k+i} = dx_{k+i}, \qquad i = 1, \ldots, n-k.$$

These can be written in matrix form

$$\begin{pmatrix} 0 \\ dX_2 \end{pmatrix} = \begin{pmatrix} J_1 & J_2 \\ 0 & I \end{pmatrix} dX$$

where I is the $(n-k) \times (n-k)$ identity matrix and $dX_2' = (dx_{k+1}, \ldots, dx_n)$. Thus in the subspace S_T,

$$dX = \begin{pmatrix} J_1^{-1} & -J_1^{-1} J_2 \\ 0 & I \end{pmatrix} \begin{pmatrix} 0 \\ dX_2 \end{pmatrix} = \begin{pmatrix} B \\ I \end{pmatrix} dX_2$$

where $B = -J_1^{-1}J_2$. The differential of arc in S_T is given by

$$ds^2 = dX'\,dX = dX_2'(B', I)\begin{pmatrix} B \\ I \end{pmatrix} dX_2' = dX_2'(B'B+I)dX_2.$$

Using (A-1) and (A-2)

$$dV_T = \sqrt{|B'B+I|}\ \prod_{i=k+1}^{n} dx_i$$

which is the required result (A-4).

Formula (A-4) is not suitable for numerical computations, particularly since some searching may be necessary to find a non-singular submatrix J_1. However, it can be put in a form leading to (2) which is easier to evaluate on a computer in any particular case. This is done by the following theorems.

Theorem 2 :

$$|B'B+I| = |BB'+I|$$

where in each case I is the identity matrix of appropriate dimension.

Proof :

$$(B'B+I) = C'C$$

where

$$C = \begin{pmatrix} B \\ I \end{pmatrix}$$

and I is $(n-k)\times(n-k)$. Thus $|B'B+I| = |C'C|$ is the sum of squares of all minors of C of order $(n-k)$ (cf. Rao, 1965). It can easily be verified that the sum of squares of all minors of order $(n-k)$ of C is equal to the sum of squares of all minors of B. Consider $(BB'+I)$, which can be written as DD', where $D = (B, I)$ and I is the $(k\times k)$ identity matrix. Now $|DD'|$ is the sum of squares of all minors of order k of D. As before, it is easy to verify that this is equal to the sum of squares of all minors of B. Thus

$$\begin{aligned} |B'B+I| &= \text{sum of squares of minors of } B \\ &= |BB'+I|. \end{aligned}$$

Corollary 1 :

$$dV_T = |J_1|^{-1}\sqrt{|JJ'|}\ \prod_{i=k+1}^{n} dx_i.$$

Proof : From (A-4) and the definition of B,

$$dV_T = \sqrt{|B'B+I|}\ \prod_{k+1}^{n} dx_i = \sqrt{|BB'+I|}\ \prod_{k+1}^{n} dx_i$$

$$= \sqrt{|J_1^{-1}J_2J_2'(J_1')^{-1}+I|}\ \prod_{k+i}^{n} dx_i.$$

Thus

$$dV_T = \sqrt{|J_1^{-1}(J_2J_2'+J_1J_1')(J_1')^{-1}|}\prod_{k+1}^{n} dx_t$$

$$= |J_1|^{-1}\sqrt{|JJ'|}\prod_{k+1}^{n} dx_t.$$

Corollary 2 :

$$\Pi dx_t \quad \Pi dT_t(\theta) = dV_T\sqrt{|JJ'|}.$$

Proof : Change variables from $x_1, \ldots, x_n$ to $T_1, \ldots, T_k, x_{k+1}, \ldots, x_n$.

Then

$$\Pi dx_t = \left|\frac{\partial(x_1, \ldots, x_n)}{\partial(T_1, \ldots, T_k, x_{k+1}, \ldots, x_n)}\right| \prod_{1}^{k} dT_i \prod_{k+1}^{n} dx_t$$

$$= |J_1|^{-1}\prod_{1}^{k} dT_t dV_T/|J_1|^{-1}\sqrt{|JJ'|}.$$

from Corolary 1.

Corollary 3 : *The conditional likelihood is proportional to*

$$f(x_1, \ldots, x_n;\ \theta\,|\,T_1, \ldots, T_k)/\sqrt{|JJ'|}.$$

Proof : The conditional distribution given the sufficient statistics $T_1, \ldots, T_k$ is

$$f(x_1, \ldots, x_n;\ \theta\,|\,T_1, \ldots, T_k)\prod^{n} dx_i/\prod_{t}^{k} dT_i(\theta) = f(x_1, \ldots, x_n;\ \theta\,|\,T_1, \ldots, T_k)dV_T/\sqrt{|JJ'|}.$$

from Corollary 2. Since dV_T is kept constant in forming the conditional likelihood the result follows.

References

Alling, D. W. (1971) : Estimation of hit number. *Biometrics*, **27**, 605-613.

Andersen, E. B. (1970) : Asymptotic properties of conditional maximum-likelihood estimators. *J. Roy. Stat. Soc.*, B, **32**, 283-302.

Bacon, D. W. and Watts, D. G. (1971) : Estimating the transition between intersecting straight lines. *Biometrika*, **58**, 525-534.

Barnard G. A. (1963) : Some logical aspects of the fiducial argument. *J. Roy. Stat. Soc.*, B, **25**, 111-114.

——— (1966) : Summary remarks. *New Developments in Survey Sampling*, 696-711, edited by N. L. Johnson and Harry Smith, Jr., John Wiley and Sons, N. Y.

Bliss, C. I. (1967) : *Statistics in Biology*, **I**, McGraw Hill Co., New York.

Box, G. E. P. and Cox, D. R. (1964) : An analysis of transformations (with discussion). *J. Roy. Stat. Soc.*, B, **26**, 211-252.

COX, D. R. (1958) : Some problems connected with statistical inference. *Ann. Math. Stat.*, **29**, 357-372.

——— (1970) : *Analysis of Binary Data*, Methuen, London

COX, D. R. and LEWIS, P. A. W. (1966) : *Statistical Analysis of Series of Events*, Methuen, London.

FISHER, R. A. (1941) : The negative binomial distribution. *Ann. of Eugenics*, **11**, 182-187.

FRASER, D. A. S. (1967) : Data transormations and the linear model. *Ann. Math. Stat.*, **38**, 1456-1466.

——— (1968) : *The Structure of Inference*, John Wiley and Sons, New York.

KALBFLEISCH, J. D. and PRENTICE, R. L. (1972) : Marginal likelihoods based on Cox's regression and life model. Submitted for publication.

KALBFLEISCH, J. D. and SPROTT, D. A. (1970) : Applications of likelihood methods to models involving large numbers of parameters (with discussion). *J. Roy. Stat. Soc.*, B, **32**, 175-208.

NEYMAN, J. and SCOTT, E. L. (1948) : Consistent estimates based on partially consistent observations. *Econometrica*, **16**, 1-32.

PATIL, G. P. and JOSHI, S. W. (1968) : *A Dictionary and Bibliography of Discrete Distributions*, Oliver and Boyd, Edinburgh.

PLAIT, A. (1962) : The Weibull distribution with tables. *Industrial Quality Control*, **19**, No. 5, 17-26.

RAO, C. R. (1965) : *Linear Statistical Inference*, John Wiley & Sons, New York (excercise 2.6).

SOKOLNIKOFF, I. S. (1964) : *Tensor Analysis Theory and Applications to Geometry and Mechanics of Continua*, 2nd Edition.

SPROTT, D. A. (1970) : Some exact methods of inference applied to discrete distributions. *Random Counts in Scientific Work*, **1**, 207-232, edited by G. P. Patil, The Penn. State Univ. Press.

WILLIAMS, E. J. (1962) : Exact fiducial limits in non-linear estimation. *J. Roy. Stat. Soc.*, B, **24**, 125-139.

ZELEN, M. (1969) : Play the winner rule and the controlled clinical trial. *Jour. Am. Stat. Assoc.*, **64**, 131-146.

Paper received : June, 1972.
Revised : November, 1972.

Note added in proof : The calibration problem, about which there has been some recent controversy (cf. G. K. Shukla, Technometrics 14, 1972, 547-553) is a special case of Example 4.2, non-linear regression. In the calibration problem, there are n independent observations y_α from a normal distribution $N(\beta_1+\beta_2 x_\alpha, \sigma^2)$, $\alpha = 1, 2, \ldots, n$, and m further observations $y_{n+1}, y_{n+2}, \ldots, y_{n+m}$ corresponding to a fixed unknown $x = \xi$. It is required to estimate ξ. This is a special case of Example 4.2 with $f_{1\alpha} = 1$, $f_{2\alpha} = x_\alpha$ $(\alpha = 1, 2, \ldots, n)$, $f_{2\alpha} = \xi (\alpha = n+1, \ldots, n+m)$, $k = 2$ and $n+m$ replacing n.