Bernoulli **5**(4), 1999, 677–682

Self-similarity of mathematical likelihood

STEPHEN RAYMOND CHAMBERLIN

Department of Mathematics and Statistics, York University, 4700 Keele Street, Toronto, Ontario, Canada M3J 1P3. E-mail:chamber@mathstat.yorku.ca

We study sufficiency in terms of the extent to which the Taylor series expansion of the normed log-likelihood function generated by the sampling distribution of a statistic conforms to that of the sample normed log-likelihood function, and give an associated self-similar property of mathematical likelihood (not self-similarity in the sense of self-similar stochastic processes).

Keywords: fidelity; likelihood; sufficiency

1. Introduction

This paper is a study of the process of information recovery concerning a full parameter vector $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$, where $\theta \in \Theta \subset \mathbb{R}^p$ in parametric inference. Fisher (1934) identified the two complementary aspects of this process. The first is associated with the subset of the universe of p-dimensional parametric models called prime (p, p) exponential families (Barndorff–Nielsen and Cox 1994, p. 63) for which the maximum likelihood estimator $\hat{\theta}$ is sufficient for θ. Its sampling distribution provides a basis for constructing unconditional inference procedures for θ. The second is associated with the subset of the universe referred to as transformation models for which complete recovery of the information is achieved by conditioning the maximum likelihood estimator $\hat{\theta}$ on the maximal invariant of the group action (Barndorff–Nielsen *et al.* 1989). The conditional sampling distribution of $\hat{\theta}$ is sufficient for θ and allows conditional inference procedures for θ to be constructed.

Let $l(\theta; x)$ be the log-likelihood function,

$$l_{/i_1 i_2 \ldots i_k}(\theta; x) = \partial^k l(\theta; x)/\partial\theta_{i_1} \ldots \partial\theta_{i_k}$$

and $\hat{l}_{/i_1 i_2 \ldots i_k} = l_{/i_1 i_2 \ldots i_k}(\hat{\theta}; x)$. Fisher (1925) also suggested that an initial step towards a process of information recovery more generally could be achieved by reducing the sample information to the maximum likelihood estimate $\hat{\theta}$ and the log-likelihood derivative arrays or likelihood shape statistics $\{\hat{l}_{/i_1 i_2}\}, \ldots, \{\hat{l}_{/i_1 i_2 \ldots i_m}\}$ which specify the Taylor series expansion of $l(\theta; x)$ as a function of θ about $\theta = \hat{\theta}$ up to some finite order m.

As discussed by Barndorff–Nielsen and Cox (1994, p. 226), it is typically argued that, since the normed log-likelihood function $\bar{l}(\cdot) = l(\cdot; \mathrm{x}) - l(\hat{\theta}; x)$ is a minimal sufficient statistic and $\bar{l}$ can be approximated to any degree of accuracy by its associated Taylor series expansion, $\hat{\theta}$ and the likelihood shape statistics to order m are approximately sufficient for θ. The most rigorous formulation is attributed to Michel (1978). Michel (1978) showed that for certain model sequences associated with ordinary repeated sampling one could construct a new sequence of models for which these statistics are exactly sufficient and such that the

1350–7265 © 1999 ISI/BS

variational distance between the constructed model sequence and the target model sequence is $O(n^{-m/2})$ uniformly for θ in compact subsets of Θ.

The purpose of the present paper is to show that mathematical likelihood has in fact an elegant property, under projection onto the σ-algebra generated by $\hat{\theta}$ and the likelihood shape statistics, which establishes their joint approximate sufficiency in terms of the reproduction of the Taylor series expansion of $\bar{l}(\theta)$ to a given order for fixed sample size and without the restriction to the ordinary repeated sampling framework.

To this end, we consider a parametric model $\mathscr{P} = \{P_\theta\}$ where $\theta \in \Theta \subset \mathbb{R}^p$ and $X \in \mathscr{X} \subset \mathbb{R}^n$, for which the probability measures in $\mathscr{P}$ have common support. Then for any specified θ_0,

$$L(\theta;\, x) = \frac{\mathrm{d}P_\theta}{\mathrm{d}P_{\theta_0}}(x) \tag{1.1}$$

is a version of the likelihood function of θ generated by the sampling distribution of X and

$$\tilde{L}(\theta;\, A(x)) = \mathrm{E}_{\theta_0}[L(\theta;\, X) | A(X) = A(x)] \tag{1.2}$$

is a version of the likelihood function of θ generated by the sampling distribution of the statistic $A(X)$ for the observation x.

The extent to which the likelihood function (1.2) based on the marginal distribution of $A(X)$ reproduces the sample likelihood function (1.1) for the observed x will be called its *fidelity*. This terminology was introduced by Chamberlin and Sprott (1991). For example, in the case of prime exponential family models the likelihood function generated by the sampling distribution of the maximum likelihood estimator $\hat{\theta}(X)$ has perfect fidelity for every observed x.

One way to assess the extent of likelihood reproduction under the data reduction to $A(X)$ is to compare the Taylor series expansion of the normed log-likelihood $\tilde{l}(\theta;\, A(x)) - \tilde{l}(\tilde{\theta}(A(x));\, A(x))$ based on the marginal distribution of $A(X)$ with the corresponding expansion of the original normed log-likelihood $l(\theta;\, x) - l(\hat{\theta}(x);\, x)$ based on the distribution of X. Here, $l(\theta;\, x)$ is the original log-likelihood, $\tilde{l}(\theta;\, A(x))$ is the log-likelihood based on the marginal distribution of $A(X)$, $\hat{\theta}(x)$ is a maximum likelihood estimate of θ from the original model, and $\tilde{\theta}(A(x))$ maximizes $\tilde{L}(\theta;\, A(x))$. Then fidelity to first order means that the estimate $\tilde{\theta}(A(x))$ based on the marginal distribution of $A(X)$ satisfies $\tilde{\theta}(A(x)) = \hat{\theta}(x)$. Fidelity to order m requires this as well as

$$\tilde{l}_{/i_1 i_2 \ldots i_k}(\tilde{\theta}(A(x));\, A(x)) = l_{/i_1 i_2 \ldots i_k}(\hat{\theta}(x);\, x), \qquad k \leqslant m. \tag{1.3}$$

Thus, fidelity to order m implies that the two Taylor series expansions agree to order m.

Fidelity to first order implies that $\hat{\theta}(x)$ is a function of x through $A(x)$. Therefore, any statistic $A(x)$ that achieves fidelity to first order is of the form $A(x) = (\hat{\theta}(x),\, T_1(x))$. Fidelity to second order implies furthermore that $\hat{j}(x)$, the observed Fisher information matrix evaluated at $\hat{\theta}$, is a function of x through $A(x)$. Thus in this case $A(x)$ is of the form $A(x) = (\hat{\theta}(x),\, \hat{j}(x),\, T_2(x))$. More generally, fidelity to order m implies that the maximum likelihood estimator and the likelihood shape statistics up to that order can be expressed as a function of x through $A(x)$. We show below that this necessary condition on the form of $A(x)$ is in fact sufficient to achieve fidelity to order m.

A range value θ of the maximum likelihood estimator $\hat{\theta}(X)$ is said to be *stationary* if, for almost all $x \in \{x\colon \hat{\theta}(x) = \theta\}$, the log-likelihood $l(\theta; x)$ admits a Taylor series expansion as a function of θ about $\hat{\theta}(x)$, where $\hat{\theta}(x)$ is a root of the score equations and a relative maximum of $l(\theta; x)$.

In what follows, $L_{/i_1 i_2 \ldots i_k}(\theta; x) = \partial^k L(\theta; x)/\partial\theta_{i_1} \ldots \partial\theta_{i_k}$, and similarly for derivatives of $l(\theta; x)$, $\tilde{L}(\theta; A(x))$ and $\tilde{l}(\theta; A(x))$. The observed Fisher information matrix generated by $\tilde{L}(\theta; A(x))$ and evaluated at $\theta = \tilde{\theta}(A(x))$ is denoted by $\tilde{j}(x)$.

In the following we often interchange the order in which we apply the operations of differentiation with respect to θ and expectation with respect to P_{θ_0}. The regularity conditions required to do this are of the standard kind required for likelihood calculations.

2. Self-similarity of the method of maximum likelihood

In this section, we show that the method of maximum likelihood applied to the likelihood function generated by the sampling distribution of any statistic $A(X) = (\hat{\theta}(X), T_1(X))$ yields $\tilde{\theta}(A(x)) = \hat{\theta}(x)$ for all x. Also, if $\hat{\theta}(x)$ is a stationary value of $\hat{\theta}(X)$ the estimate $\tilde{\theta}(A(x))$ is a root of the corresponding score equation.

If $\hat{\theta}(x)$ is observed then $\tilde{L}(\hat{\theta}; A(x))$ is a weighted average of possible sample likelihood functions $L(\theta; x)$ each of which achieves its supremum at $\hat{\theta}$, and hence, by (1.2),

$$\begin{aligned}
\tilde{L}(\hat{\theta}(x); A(x)) &= \tilde{L}(\hat{\theta}(x); [\hat{\theta}(x), T_1(x)]) \\
&= \mathrm{E}_{\theta_0}[L(\hat{\theta}(x); X) | (\hat{\theta}(X), T_1(X)) = (\hat{\theta}(x), T_1(x))] \\
&= \mathrm{E}_{\theta_0}[L(\hat{\theta}(X); X) | (\hat{\theta}(X), T_1(X)) = (\hat{\theta}(x), T_1(x))] \\
&\geqslant \mathrm{E}_{\theta_0}[L(\theta; \mathrm{X}) | (\hat{\theta}(X), T_1(X)) = (\hat{\theta}(x), T_1(x))] \\
&= \tilde{L}(\theta; [\hat{\theta}(x), T_1(x)]) \\
&= \tilde{L}(\theta; A(x)),
\end{aligned}$$

irrespective of the value of $T_1(x)$. Hence, for each x,

$$\tilde{\theta}(A(x)) = \hat{\theta}(x) \tag{2.1}$$

is a point which maximizes $\tilde{L}(\theta; A(x))$ and $\tilde{\theta}(A(X))$ is a maximum likelihood estimator of θ based on the sampling distribution of $A(X)$ alone. This observation simply says that the principle of maximizing a likelihood function is self-similar with respect to any information reduction of the form $X \to (\hat{\theta}(X), T_1(X))$. That is, this likelihood characteristic is exactly preserved under the projection (1.2).

If $\tilde{L}(\theta; A(x))$, given by (1.2), is differentiated with respect to θ_i one obtains, for each $i = 1, 2, \ldots, p$,

$$\tilde{L}_{/i}(\theta; A(x)) = \mathrm{E}_{\theta_0}[L_{/i}(\theta; \mathrm{X}) | (\hat{\theta}(X), T_1(X)) = (\hat{\theta}(x), T_1(x))]. \tag{2.2}$$

If $\hat{\theta}(x)$ is a stationary value, evaluating (2.2) at $\theta = \hat{\theta}(x)$ yields

$$\begin{aligned}\tilde{L}_{/i}(\hat{\theta}(x);\,[\hat{\theta}(x),\,T_1(x)]) &= \mathrm{E}_{\theta_0}[L_{/i}(\hat{\theta}(x);\,X)|(\hat{\theta}(X),\,T_1(X)) = (\hat{\theta}(x),\,T_1(x))]\\ &= \mathrm{E}_{\theta_0}[L_{/i}(\hat{\theta}(X);\,X)|(\hat{\theta}(X),\,T_1(X)) = (\hat{\theta}(x),\,T_1(x))]\\ &= \mathrm{E}_{\theta_0}[0|(\hat{\theta}(X),\,T_1(X)) = (\hat{\theta}(x),\,T_1(x))]\\ &= 0, \end{aligned} \tag{2.3}$$

irrespective of the value of $T_1(x)$. Hence, $\tilde{\theta}(A(x)) = \hat{\theta}(x)$ is also a root of the score equations obtained from $\tilde{l}(\theta;\,A(x))$ whenever $\hat{\theta}(x)$ is a stationary value. Moreover, the Hessian matrix

$$\tilde{L}_{/ij}(\hat{\theta}(x);\,A(x)) = \mathrm{E}_{\theta_0}[L_{/ij}(\hat{\theta}(X);\,X)|(\hat{\theta}(X),\,T_1(X)) = (\hat{\theta}(x),\,T_1(x))] \tag{2.4}$$

is negative definite. It is a weighted average of negative definite matrices since $\hat{\theta}$ is a relative maximum for almost all X giving rise to $\hat{\theta}$. Hence, solving the score equations associated with the likelihood function generated by the sampling distribution of $A(X) = (\hat{\theta}(X),\,T_1(X))$ yields, irrespective of the choice of $T_1(X)$, the same value of the optimum as that obtained when this method is applied to the sampling distribution of X. This self-similarity of the method of maximum likelihood was evidently first stated by Fisher (1922). More generally, note that if one chooses a relative maximum of the sample likelihood function which is not necessarily a global maximum in repeated sampling, the sampling distribution of the resulting estimator will satisfy the self-similar property (2.3).

Of course, if $A(X)$ is a sufficient statistic for θ the self-similar properties (2.1) and (2.3) are trivially true since then $\tilde{L}(\theta;\,A(x)) = L(\theta;\,x)$ for every x. Note further that the minimal choice of $A(X)$ to recover in our sense the point that maximizes the sample likelihood function is to take $A(X) = \hat{\theta}(X)$. The sampling distribution of $\hat{\theta}$ satisfies the self-similar properties (2.1) and (2.3) and therefore has fidelity to first order. The data reduction $X \to \hat{\theta}(X)$ is the first iteration in Fisher's process of iterative information recovery. If $\hat{\theta}(X)$ were sufficient one would stop. Otherwise, we can proceed to the next stage.

3. Higher-order aspects of self-similarity

To achieve fidelity to second order we can refine $T_1(X)$ as $T_1(X) = (\hat{j}(X),\,T_2(X))$, where $T_2(X)$ is arbitrary and $\hat{j}(x)$ is the observed Fisher information matrix evaluated at $\hat{\theta}(x)$, with $\hat{j}_{ij}(x) = -l_{/ij}(\hat{\theta}(x);\,x)$. Then $A(X) = (\hat{\theta}(X),\,\hat{j}(X),\,T_2(X))$ and we can generate a likelihood function based on the distribution of $A(X)$ using (1.2). Then $\tilde{L}(\theta;\,A(X))$ satisfies (2.1) and (2.3), and $A(X)$ has fidelity to first order. Differentiating $\tilde{L}(\theta;\,[\hat{\theta}(x),\,\hat{j}(x),\,T_2(x)])$ with respect to θ_i and θ_j using (1.2) and evaluating the resulting expression at $\theta = \tilde{\theta}(A(x)) = \hat{\theta}(x)$, where $\hat{\theta}(x)$ is a stationary value, yields

$$
\begin{aligned}
&\tilde{L}_{/ij}(\tilde{\theta}(A(x));\ [\hat{\theta}(x),\ \hat{j}(x),\ T_2(x)]) \\
&= \tilde{L}_{/ij}(\hat{\theta}(x);\ [\hat{\theta}(x),\ \hat{j}(x),\ T_2(x)]) \\
&= \mathrm{E}_{\theta_0}[L_{/ij}(\hat{\theta}(x);\ X)|(\hat{\theta}(X),\ \hat{j}(X),\ T_2(X)) = (\hat{\theta}(x),\ \hat{j}(x),\ T_2(x))] \\
&= \mathrm{E}_{\theta_0}[L_{/ij}(\hat{\theta}(X);\ X)|(\hat{\theta}(X),\ \hat{j}(X),\ T_2(X)) = (\hat{\theta}(x),\ \hat{j}(x),\ T_2(x))] \\
&= \mathrm{E}_{\theta_0}[-\hat{j}_{ij}(X)L(\hat{\theta}(X);\ X)|(\hat{\theta}(X),\ \hat{j}(X),\ T_2(X)) = (\hat{\theta}(x),\ \hat{j}(x),\ T_2(x))] \\
&= -\hat{j}_{ij}(x)\mathrm{E}_{\theta_0}[L(\hat{\theta}(X);\ X)|(\hat{\theta}(X),\ \hat{j}(X),\ T_2(X)) = (\hat{\theta}(x),\ \hat{j}(x),\ T_2(x))] \\
&= -\hat{j}_{ij}(x)\tilde{L}(\hat{\theta}(x);\ [\hat{\theta}(x),\ \hat{j}(X),\ T_2(x)]) \\
&= -\hat{j}_{ij}(x)\tilde{L}(\hat{\theta}(A(x));\ [\hat{\theta}(x),\ \hat{j}(x),\ T_2(X)]), \qquad (3.1)
\end{aligned}
$$

since $\hat{j}_{ij}(x) = -L_{/ij}(\hat{\theta}(x);\ x)/L(\hat{\theta}(x);\ x)$. Now $\tilde{j}_{ij}(x) = -\tilde{l}_{/ij}(\tilde{\theta}(A(x));\ A(x))$, which can be written as $\tilde{j}_{ij}(x) = -\tilde{L}_{/ij}(\tilde{\theta}(A(x));\ A(x))/\tilde{L}(\tilde{\theta}(A(x));\ A(x))$. Hence, we obtain, regardless of the choice of $T_2(x)$, in addition to (2.1) and (2.3) the identity

$$\tilde{j}(x) = \hat{j}(x) \qquad (3.2)$$

for x, yielding a stationary value of $\hat{\theta}(X)$. In particular, the sampling distribution of $(\hat{\theta}(X),\ \hat{j}(X))$ satisfies the self-similar properties (2.1), (2.3) and (3.2) and therefore has fidelity to second order. The data reduction $X \rightarrow (\hat{\theta}(X),\ \hat{j}(X))$ is the second iteration in Fisher's process of iterative information recovery. If $(\hat{\theta}(X),\ \hat{j}(X))$ were sufficient one would stop. Otherwise, we can proceed to the next stage.

This process of iterative information recovery can be extended to any finite order m. Let $A(x)$ be any statistic for which the maximum likelihood estimator and the likelihood shape statistics to order m can be expressed as a function of x through $A(x)$. One can then easily show that the statistic $L_{/i_1 i_2 \ldots i_k}(\hat{\theta};\ x)/L(\hat{\theta};\ x)$ depends on x through $A(x)$. To see this, note that for every k one can express $l_{/i_1 i_2 \ldots i_k}(\theta)$ algebraically in the form

$$l_{/i_1 i_2 \ldots i_k}(\theta) = \frac{L_{/i_1 i_2 \ldots i_k}(\theta)}{L(\theta)} + h(\{l_{/j_1}\},\ \{l_{/j_1 j_2}\},\ \ldots,\ \{1_{/j_1 \ldots j_{k-1}}\}). \qquad (3.3)$$

Following exactly our previous pattern given by (3.1), one can obtain that

$$\tilde{L}_{/i_1 i_2 \ldots i_k}(\tilde{\theta}(A(x));\ A(x))/\tilde{L}(\tilde{\theta}(A(x));\ A(x)) = L_{/i_1 i_2 \ldots i_k}(\hat{\theta}(x);\ x)/L(\hat{\theta}(x);\ x)$$

for arbitrary $i_1,\ i_2,\ \ldots,\ i_k$ and each $k = 1, 2, \ldots, m$. Using these identities, one can sequentially construct the identities (1.3) since both $\tilde{l}(\theta;\ A(x))$ and $l(\theta;\ x)$ follow identical patterns of the form (3.3) under differentiation. Hence, the sampling distribution of $A(X)$ has fidelity to at least order m.

In particular, the joint sampling distribution of $\hat{\theta}(X)$ and the likelihood shape statistics up to order m has fidelity to at least order m. In this sense we can say that $\hat{\theta}(X)$ and the likelihood shape statistics up to order m are necessary and sufficient to order m. That is, these statistics characterize fidelity to order m.

Acknowledgements

The author would like to thank the Editors and a referee for several comments which greatly improved the presentation of this paper. The author would also like to thank J.L. Jensen for working through an earlier version of the manuscript. This research was partially supported through a grant from the Natural Sciences and Engineering Research Council of Canada.

References

Barndorff–Nielsen, O.E., Blæsild, P. and Eriksen, P.S. (1989) *Decomposition and Invariance of Measures, and Statistical Transformation Models*, Lecture Notes in Statist. 58. New York: Springer-Verlag.

Barndorff–Nielsen, O.E. and Cox, D.R. (1994) *Inference and Asymptotics*, Monographs Statist. Appl. Probab. 52. London: Chapman & Hall.

Chamberlin, S.R. and Sprott, D.A. (1991) Inferential estimation, likelihood, and maximum likelihood linear estimating functions, In V.P. Godambe (ed.), *Estimating Functions*, pp. 255–266. Oxford: Oxford University Press.

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

Fisher, R.A. (1925) Theory of statistical estimation. *Proc. Cambridge Philos. Soc.*, **22**, 700–725.

Fisher, R.A. (1934) Two new properties of mathematical likelihood. *Proc. Roy. Soc. London Ser. A*, **144**, 285–307.

Michel, R. (1978) Higher order asymptotic sufficiency. *Sankhyā A*, **40**, 76–84.

Received August 1996 and revised March 1998.