*Bernoulli* **10**(3), 2004, 421–446

# Likelihood functions based on parameter-dependent functions

THOMAS A. SEVERINI

*Department of Statistics, Northwestern University, Evanston IL 60208-4070, USA.*
*E-mail: severini@northwestern.edu*

Consider likelihood inference about a scalar function $\psi$ of a parameter $\theta$. Two methods of constructing a likelihood function for $\psi$ are conditioning and marginalizing. If, in the model with $\psi$ held fixed, $T$ is ancillary, then a marginal likelihood may be based on the distribution of $T$, which depends only on $\psi$; alternatively, if a statistic $S$ is sufficient when $\psi$ is fixed, then a conditional likelihood function may be based on the conditional distribution of the data given $S$. The statistics $T$ and $S$ are generally required to be the same for each value of $\psi$. In this paper, we consider the case in which either $T$ or $S$ is allowed to depend on $\psi$. Hence, we might consider the marginal likelihood function based on a function $T_\psi$ or the conditional likelihood given a function $S_\psi$. The properties and construction of marginal and conditional likelihood functions based on parameter-dependent functions are studied. In particular, the case in which $T_\psi$ and $S_\psi$ may be taken to be functions of the maximum likelihood estimators is considered and approximations to the resulting likelihood functions are presented. The results are illustrated on several examples.

*Keywords:* ancillary statistics; conditional likelihood; likelihood inference; marginal likelihood; maximum likelihood estimators; modified profile likelihood; nuisance parameters

## 1. Introduction

Consider a parametric model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ for an observed random variable $Y$. Let $g$ denote a real-valued function on $\Theta$ such that $\psi = g(\theta)$ represents the parameter of interest of the model and consider likelihood-based inference for $\psi$. For models that are parametrized by $\psi$ alone, inference may be based directly on $L(\psi)$, the likelihood function for $\psi$. For models with a nuisance parameter, likelihood inference is often based on a *pseudo-likelihood*, a function of the data and $\psi$ that has properties similar to those of a likelihood function. One commonly used pseudo-likelihood is the profile likelihood, in which $\theta$ is replaced by $\hat{\theta}_\psi$, the maximum likelihood estimator of $\theta$ for fixed $\psi$, in $L(\theta)$, leading to $L_p(\psi) = L(\hat{\theta}_\psi)$. Inference for $\psi$ then proceeds by treating $L_p(\psi)$ as a likelihood function for $\psi$. Although this approach leads to optimal procedures in terms of first-order asymptotic theory, in small or moderate samples inferences based on $L_p(\psi)$ may be misleading.

Two approaches to forming a likelihood function for $\psi$ are conditioning and marginalizing. If there exists a statistic $T$ such that the distribution of $T$ depends only on $\psi$, then a marginal likelihood may be based on this distribution. Let $\mathcal{P}_\psi$ denote the model with $\psi$ held fixed. Then $T$ is an ancillary statistic in this model; here, as throughout the paper, an ancillary statistic in a model will mean a statistic that has the same distribution under each distribution in the model. Alternatively, if there exists a statistic $S$ which is sufficient in $\mathcal{P}_\psi$, and not sufficient in $\mathcal{P}$, then a conditional likelihood can be based on the conditional distribution of $Y$ given $S$. See, for example, Kalbfleisch and Sprott (1970; 1973) for a discussion of marginal and conditional likelihoods.

Hence, a marginal or conditional likelihood exists only in certain special cases – specifically when, in $\mathcal{P}_\psi$, there exists either an ancillary statistic $T$ or a sufficient statistic $S$ which is not sufficient in the full model. A key part of this assumption is that either the ancillary statistic or the sufficient statistic is the same for all values of $\psi$. One way to weaken these assumptions is to allow either $T$ or $S$ to depend on the parameter value $\psi$. Hence, we might consider the marginal likelihood function based on a function $T_\psi$ or the conditional likelihood given a function $S_\psi$.

In this paper, the construction of marginal and conditional likelihood functions based on parameter-dependent functions is considered. The important messages and results of the paper are as follows. First, if a likelihood function is based on a parameter-dependent function then it is important to include a 'volume element' in the specification of the likelihood. Second, there are a number of important considerations to keep in mind when specifying this volume element. Third, if likelihoods based on parameter-dependent functions are considered, there is a close connection between marginal and conditional likelihoods (at least from an approximation-based point of view). Finally, using likelihoods based on parameter-dependent functions, it is generally possible to construct an approximate marginal or conditional likelihood (this is in contrast to the modified profile likelihood, which requires a specific model structure for its justification).

In Section 2, the general problem of constructing such likelihood functions is discussed. In Section 3, the cases in which $T_\psi$ and $S_\psi$ may be taken to be functions of the maximum likelihood estimators $\hat{\theta}$ and $\hat{\theta}_\psi$ are considered and approximations to the resulting marginal and conditional likelihoods are presented. In this section, it is also shown that, using likelihoods based on parameter-dependent functions, it is generally possible to construct an approximate marginal or conditional likelihood. These approximations are related to the modified profile likelihood proposed by Barndorff-Nielsen (1980; 1983), and this connection is discussed in Section 4. Section 5 contains several examples.

The main discussion of parameter-dependent likelihood functions given in Section 2 is based on the assumption that the underlying data have an absolutely continuous distribution and the pseudo-likelihood functions derived in Section 3 are obtained under this assumption. In Section 6, it is shown that the pseudo-likelihood functions derived in Section 3 are also valid when the underlying data have a lattice distribution.

The problem of constructing a likelihood function for a parameter of interest has been considered from many different points of view. Here we follow the same general approach used by Kalbfleisch and Sprott (1970; 1973). The purpose of this paper is to expand on this approach and to apply it to the cases in which $T_\psi$ and $S_\psi$ may be taken to be functions of the maximum likelihood estimators $\hat{\theta}$ and $\hat{\theta}_\psi$.

Conditional and marginal likelihood functions based on parameter-dependent functions are considered by Fraser (1967; 1968; 1972; 1979) for the case of a linear model. Related results for more general models are discussed by Fraser and Reid (1989). As noted above, the approximations obtained here are closely related to the modified profile likelihood function proposed by Barndorff-Nielsen (1980; 1983); see also Cox and Reid (1987), McCullagh and Tibshirani (1990), Barndorff-Nielsen and Cox (1994), Fraser and Reid (1995) and Severini (1998; 2000a).

## 2. Conditional and marginal likelihood functions

### 2.1. Marginal likelihoods

Suppose that the distribution of $T$ depends only on $\psi$ so that $T$ is ancillary in the model with $\psi$ held fixed. If $T$ itself does not depend on $\psi$, then a marginal likelihood for $\psi$ based on $T$ is given by $L(\psi; T) = p(t; \psi)$, where $t$ is the observed value of $T$; here, and throughout the paper, the symbol $p$ is used to denote a density function with the argument of $p$ indicating the specific density under consideration. Unless explicitly stated otherwise, all density functions will be assumed to be with respect to Lebesgue measure.

However, when the statistic $T$ depends on $\psi$, $T \equiv T_\psi$, there are difficulties in constructing a marginal likelihood function for $\psi$ based on $T$. These are illustrated by the following example, which is also discussed, for example, in McCullagh and Nelder (1989, Exercise 7.6).

***Example 1 Ratio of normal means.*** Consider two independent samples of size $n$ from normal distributions with means $\mu_1$ and $\mu_2$, respectively, and each with standard deviation 1. Let $X$ and $Y$ denote the sample means and take $\psi = \mu_1/\mu_2$ as the parameter of interest, with $\lambda = \mu_2$ as a nuisance parameter. Thus, in this example, the data will be denoted by $(x, y)$. The log-likelihood function for the model is given by

$$\ell(\theta) = n\lambda(\psi x + y) - \frac{n}{2}(\psi^2 + 1)\lambda^2.$$

The function $X - \psi Y$ is normally distributed with mean 0 and variance $(1 + \psi^2)/n$; the marginal likelihood based on this function is given by

$$-\frac{n}{2}\frac{(x - \psi y)^2}{1 + \psi^2} - \frac{1}{2}\log(1 + \psi^2).$$

Alternatively, a marginal likelihood could be based on $\sqrt{n}(X - \psi Y)/\sqrt{1 + \psi^2}$ which has a standard normal distribution; the corresponding marginal likelihood is given by

$$-\frac{n}{2}\frac{(x - \psi y)^2}{1 + \psi^2}.$$

Hence, equivalent statistics lead to different forms for the marginal likelihood.

The problem with defining the marginal likelihood based on $T_\psi$ by $p(t_\psi; \psi)$ becomes apparent when we interpret $p(t_\psi; \psi)$ as the probability that $T_\psi$ lies in a small set containing

$t_\psi$. That is, it is perhaps more accurate to write the marginal likelihood function based on $T_\psi$ as $p(t_\psi;\,\psi)\mathrm{d}t_\psi$ where $\mathrm{d}t_\psi$ denotes the volume of a infinitesimally small set containing $t_\psi$. If $T_\psi$ does not depend on $\psi$ then $\mathrm{d}t_\psi$ does not depend on $\psi$ and, hence, it may be ignored when forming the likelihood function. However, whenever $t_\psi$ does depend on $\psi$, $\mathrm{d}t_\psi$ also depends on $\psi$ and, hence, this dependence must be taken into account. Alternatively, the problem may be viewed as arising from the fact that the densities used to construct the likelihood function are with respect to a dominating measure that depends on $\psi$.

One approach to dealing with this issue is to express all volume elements in terms of a variable that does not depend on $\psi$; see, for example, Kalbfleisch and Sprott (1970). Let $z$ denote a function of the data $y$, of the same dimension as $t_\psi$, such that $t_\psi$ is a one-to-one function of $z$ for each fixed $\psi$. Note that $z$ does not depend on $\psi$. Then, using the usual Euclidean volume for $\mathrm{d}z$, the volume element $\mathrm{d}t_\psi$ may be written

$$\mathrm{d}t_\psi = \left|\frac{\partial t_\psi}{\partial z}\right|\mathrm{d}z$$

and the marginal likelihood based on $T$ may be written

$$p(t_\psi;\,\psi)\left|\frac{\partial t_\psi}{\partial z}\right|\mathrm{d}z;$$

in calculating the likelihood, $\mathrm{d}z$, which does not depend on $\psi$, can be omitted.

In many cases, $t_\psi$ cannot be written as a function of a variable $z$ with the same dimension as $t_\psi$. Suppose that we wish to express $\mathrm{d}t_\psi$ in terms of $\mathrm{d}z$, where $\dim(t_\psi) \leqslant \dim(z)$. Then

$$\mathrm{d}t_\psi = \left|\frac{\partial t_\psi}{\partial z}\left(\frac{\partial t_\psi}{\partial z}\right)^{\mathrm{T}}\right|^{1/2}\mathrm{d}z.$$

This result may be derived using the following argument; for further details, see Tjur (1980, Chapter 3) or Hoffmann-Jørgensen (1994, Chapter 8).

Let $n$ denote the dimension of $z$ and let $m$ denote the dimension of $t \equiv t(z)$, where $m \leqslant n$. The basic idea is that, locally, $t$ depends on $z$ only through a linear function of $z$ of dimension $m$, which we denote by $v$. The correct Jacobian term to include in the specification of the volume element is $|\partial t/\partial v|$, which may be expressed in terms of $z$. To carry out this approach, consider a fixed value of $z$, $z_0$; the argument is a local one, near $z = z_0$, which is all that is needed. Let $v_1, \ldots, v_m$ denote vectors of length $n$ that form an orthonormal basis for the space

$$\left\{c \in \mathbb{R}^n : c = \left(\frac{\partial t}{\partial z}(z_0)\right)^{\mathrm{T}} x,\, x \in \mathbb{R}^m\right\}$$

and let $D$ denote the matrix with $j$th row $v_j^{\mathrm{T}}$. It is straightforward to show that

$$\frac{\partial t}{\partial z}(z_0)D^{\mathrm{T}}D = \frac{\partial t}{\partial z}(z_0).$$

Define $v = D(z - z_0)$ and $v_0 = 0$. Then, for $z$ near $z_0$,

$$t(z) = t(z_0) + \frac{\partial t}{\partial z}(z_0)(z - z_0) = t(z_0) + \frac{\partial t}{\partial z}(z_0)D^{\mathrm{T}}v.$$

Hence,

$$\left|\frac{\partial t}{\partial v}(v_0)\right| = \left|\frac{\partial t}{\partial v}(v_0)\left(\frac{\partial t}{\partial v}(v_0)\right)^{\mathrm{T}}\right|^{1/2} = \left|\frac{\partial t}{\partial z}(z_0)D^{\mathrm{T}}D\left(\frac{\partial t}{\partial z}(z_0)\right)^{\mathrm{T}}\right|^{1/2} = \left|\frac{\partial t}{\partial z}(z_0)\left(\frac{\partial t}{\partial z}(z_0)\right)^{\mathrm{T}}\right|^{1/2}.$$

It is worth noting that the vectors $v_1, \ldots, v_m$, in general, depend on $\psi$ and, hence, $D$ and $v$ depend on $\psi$. However, since $\mathrm{d}v = |DD^{\mathrm{T}}|^{1/2}\,\mathrm{d}z$ and, by construction, $DD^{\mathrm{T}}$ is an identity matrix, the volume element associated with $\mathrm{d}v$ does not depend on $\psi$.

Using this approach, the marginal likelihood function based on $T_\psi$ is given by

$$p(t_\psi;\,\psi)\left|\frac{\partial t_\psi}{\partial z}\left(\frac{\partial t_\psi}{\partial z}\right)^{\mathrm{T}}\right|^{1/2}. \tag{1}$$

The marginal likelihood function (1) was given by Kablfleisch and Sprott (1970; 1973). Note that this marginal likelihood is invariant under one-to-one differentiable transformations of $t_\psi$. To see this, let $w_\psi = f(t_\psi)$ for some one-to-one differentiable function $f$. Then

$$p(w_\psi;\,\psi) = p(t_\psi;\,\psi)\left|\frac{\partial t_\psi}{\partial w_\psi}\right|, \qquad t_\psi \equiv t_\psi(w_\psi).$$

The result now follows from the fact that

$$\left|\frac{\partial w_\psi}{\partial z}\left(\frac{\partial w_\psi}{\partial z}\right)^{\mathrm{T}}\right|^{1/2} = \left|\frac{\partial t_\psi}{\partial z}\left(\frac{\partial t_\psi}{\partial z}\right)^{\mathrm{T}}\right|^{1/2}\left|\frac{\partial w_\psi}{\partial t_\psi}\right|.$$

***Example 1 Ratio of normal means (continued).*** Consider the function $t_\psi = x - \psi y$. Taking $z = (z_1, z_2) = (x, y)$ yields

$$\left|\frac{\partial t_\psi}{\partial z}\left(\frac{\partial t_\psi}{\partial z}\right)^{\mathrm{T}}\right| = 1 + \psi^2;$$

the resulting marginal log-likelihood function is given by

$$-\frac{n}{2}\frac{(x - \psi y)^2}{1 + \psi^2}. \tag{2}$$

Now consider $t_\psi = \sqrt{n}(x - \psi y)/\sqrt{1 + \psi^2}$. Then $|(\partial t_\psi/\partial z)(\partial t_\psi/\partial z)^{\mathrm{T}}| = n$ and the resulting marginal log-likelihood is also given by (2).

In this example, $D = v_1^{\mathrm{T}} = (1, \psi)/\sqrt{(1 + \psi^2)}$ and, taking $z_0 = 0$, $v = (z_1 + \psi z_2)/\sqrt{1 + \psi^2}$ so that $v$ depends on $\psi$. However, $\partial v/\partial z = (1, \psi)/\sqrt{1 + \psi^2}$ so that $(\partial v/\partial z)(\partial v/\partial z)^{\mathrm{T}} = 1$, in agreement with the general result given above.

Suppose that $\tilde{z}$ is given and that $z$ is a one-to-one function of $\tilde{z}$; then

$$\left|\frac{\partial x}{\partial \tilde{z}}\left(\frac{\partial x}{\partial \tilde{z}}\right)^{\mathrm{T}}\right| = \left|\frac{\partial x}{\partial z}\frac{\partial z}{\partial \tilde{z}}\left(\frac{\partial z}{\partial \tilde{z}}\right)^{\mathrm{T}}\left(\frac{\partial x}{\partial z}\right)^{\mathrm{T}}\right|$$

so that the corresponding volume element in terms of $z$ is

$$\left|\frac{\partial x}{\partial z}H(z)\left(\frac{\partial x}{\partial z}\right)^{\mathrm{T}}\right|, \qquad H(z) = \frac{\partial z}{\partial \tilde{z}}\left(\frac{\partial z}{\partial \tilde{z}}\right)^{\mathrm{T}}\bigg|_{\tilde{z}=\tilde{z}(z)}.$$

Hence, the problem of specifying the volume element may be addressed by making a convenient choice for $z$ and then choosing a non-negative definite symmetric matrix $H$. Thus, we are using the metric $|(\mathrm{d}z)^{\mathrm{T}}H(z)^{-1}\,\mathrm{d}z|^{1/2}$ for $\mathrm{d}z$ or, equivalently, using a Euclidean metric for $\mathrm{d}w = H(z)^{-1/2}\,\mathrm{d}z$; a similar approach is used by Fraser and Reid (1989). Using this approach, the marginal likelihood function based on a particular choice of $z$ and $H$ is given by

$$p(t_\psi;\,\theta)\left|\frac{\partial t_\psi}{\partial z}H(z)\left(\frac{\partial t_\psi}{\partial z}\right)^{\mathrm{T}}\right|^{1/2}.$$

## 2.2. Conditional likelihood

Suppose that $S_\psi$, the sufficient statistic for $\mathcal{P}_\psi$, depends on the value of $\psi$. Then the conditional distribution of the data given $S_\psi = s_\psi$ depends only on $\psi$ and, hence, this conditional distribution can, in principle, be used to form a conditional likelihood function. Let $X$ denote a function of the data $Y$ such that $(X,\, S_\psi)$ is a one-to-one differentiable function of $Y$; a conditional likelihood may be based on the conditional density of $X$ given $S_\psi$,

$$p(x|s_\psi;\,\psi) = \frac{p(x,\, s_\psi;\,\theta)}{p(s_\psi;\,\theta)}.$$

As with the marginal likelihood, the conditional likelihood given $S_\psi$ is not well defined unless the densities are expressed with respect to a volume element that is independent of $\psi$.

***Example 1 Ratio of normal means (continued).*** Consider the model with $\psi$ held fixed; the maximum likelihood estimator of $\lambda$ is $\hat{\lambda}_\psi = (\psi X + Y)/(\psi^2+1)$, which is sufficient. The conditional log-likelihood based on the conditional distribution of $X$ given $\hat{\lambda}_\psi$ is given by

$$\frac{n}{2}\frac{(\psi x + y)^2}{\psi^2+1} + \frac{1}{2}\log(\psi^2+1).$$

Alternatively, the conditional likelihood may be based on the conditional distribution of $Y$ given $\hat{\lambda}_\psi$, leading to the conditional log-likelihood

$$\frac{n}{2}\frac{(\psi x + y)^2}{\psi^2+1} + \frac{1}{2}\log(\psi^2+1) - \log\psi.$$

Hence, the conditional log-likelihood is not well defined.

To specify a conditional likelihood given $S_\psi$ we need to specify the variable $z$ along with the matrix $H(z)$. When $s_\psi$ depends on $\psi$, different choices of $z$ and $H$ yield different conditional likelihoods. For particular $z$ and $H$, the conditional likelihood function given $S_\psi$ is given by

$$\frac{p(x, s_\psi; \theta)}{p(s_\psi; \theta)} \frac{|(\partial(x, s_\psi)/\partial z) H(z) (\partial(x, s_\psi)/\partial z)^{\mathrm{T}}|^{1/2}}{|(\partial s_\psi/\partial z) H(z) (\partial s_\psi/\partial z)^{\mathrm{T}}|^{1/2}}.$$

Conditional likelihoods of this form are considered by Kalbfleisch and Sprott (1970; 1973) for the case in which $H$ is an identity matrix; see also Fraser and Reid (1989).

The resulting conditional likelihood does not depend on the choice of the complementary variable $x$. In particular, the conditional likelihood may be given by

$$\frac{p(y; \theta)}{p(s_\psi; \theta)} \left| \frac{\partial s_\psi}{\partial z} H(z) \left( \frac{\partial s_\psi}{\partial z} \right)^{\mathrm{T}} \right|^{-1/2}.$$

To see this, note that, for any complementary variable $x$,

$$p(x, s_\psi; \theta) = p(y; \theta) \left| \frac{\partial y}{\partial(x, s_\psi)} \right|, \qquad y \equiv y(x, s_\psi).$$

Hence,

$$\begin{aligned} p(x, s_\psi; \theta) \left| \frac{\partial(x, s_\psi)}{\partial z} H(z) \left( \frac{\partial(x, s_\psi)}{\partial z} \right)^{\mathrm{T}} \right|^{1/2} &= p(y; \theta) \left| \frac{\partial y}{\partial(x, s_\psi)} \right| \left| \frac{\partial(x, s_\psi)}{\partial z} H(z) \left( \frac{\partial(x, s_\psi)}{\partial z} \right)^{\mathrm{T}} \right|^{1/2} \\ &= p(y; \theta) \left| \frac{\partial y}{\partial z} H(z) \left( \frac{\partial y}{\partial z} \right)^{\mathrm{T}} \right|^{1/2} \end{aligned}$$

and the result follows from the fact that $\partial y/\partial z$ does not depend on $\psi$.

## 2.3. Choice of volume element

Specification of the volume element requires specification of the variable $z$, a function of the data $y$, as well as the matrix $H$. Although there are relatively few mathematical requirements on these choices, essentially only that $H$ is non-negative definite symmetric, there are at least three desirable properties for the volume element to possess; see Kalbfleisch and Sprott (1970; 1973) for further discussion.

The first property is concerned with invariance of the model under transformations of the data. Suppose that the model for $\tilde{z} = g(z)$ is the same as the model for $z$, for some one-to-one function $g$. Then, since

$$\frac{\partial x}{\partial z} H(z) \left( \frac{\partial x}{\partial z} \right)^{\mathrm{T}} = \frac{\partial x}{\partial \tilde{z}} \frac{\partial \tilde{z}}{\partial z} H(z) \left( \frac{\partial \tilde{z}}{\partial z} \right)^{\mathrm{T}} \left( \frac{\partial x}{\partial \tilde{z}} \right)^{\mathrm{T}},$$

we should have

$$H(\tilde{z}) = \frac{\partial \tilde{z}}{\partial z} H(z) \left( \frac{\partial \tilde{z}}{\partial z} \right)^{\mathrm{T}}.$$

In particular, suppose that, for some class of $n \times n$ matrices $B$, the model for $Bz$ is the same as the model for $z$. Then $H$ should satisfy $H(Bz) = BH(z)B^{\mathrm{T}}$. This holds, for example, if $H(z)$ is taken to be the covariance matrix of $z$.

A second issue is ancillarity. Suppose that we may write $z = (z_0, a)$, where $a$ is an ancillary statistic. Then, holding $a$ fixed, the volume element is expressed in terms of $z_0$ alone. More generally, $H(z)$ corresponds to holding $a$ fixed provided that

$$H(z) \left( \frac{\partial a}{\partial z} \right)^{\mathrm{T}} = 0.$$

A third consideration is the properties of the resulting likelihood. Note that neither the marginal likelihood based on a parameter-dependent function nor the conditional likelihood given a parameter-dependent function is guaranteed to satisfy the Bartlett identities. Hence, unlike marginal and conditional likelihoods based on statistics $T$ and $S$ that do not depend on $\psi$, these likelihoods are not ‘true’ likelihoods. Consider the case of a marginal likelihood based on $T_\psi$. A likelihood function $\bar{L}(\psi)$ for $\psi$ satisfies the Bartlett identities provided that for any parameter value $\psi_0$,

$$\mathrm{E}\left\{ \frac{\bar{L}(\psi)}{\bar{L}(\psi_0)}; \theta \right\} = 1$$

for all $\theta$ such that $g(\theta) = \psi_0$. For the marginal likelihood function based on $t_\psi$, this becomes

$$\mathrm{E}\int \frac{p(t_\psi; \psi)}{p(t_0; \psi_0)} \frac{|(\partial t_\psi/\partial z) H(z) (\partial t_\psi/\partial z)^{\mathrm{T}}|^{1/2}}{|(\partial t_0/\partial z) H(z) (\partial t_0/\partial z)^{\mathrm{T}}|^{1/2}} p(y; \theta)\, \mathrm{d}y = 1, \quad z \equiv z(y), \quad t_\psi \equiv t_\psi(y), \quad t_0 \equiv t_0(y),$$

where $t_0 = t_{\psi_0}$. It is easy to see that this does not hold in general; similar considerations apply to conditional likelihoods. Hence, likelihood functions based on parameter-dependent functions do not satisfy the Bartlett identities exactly, although they may satisfy them approximately.

The choice of $z$ and $H(z)$ is an assumption about the data and would not depend, for example, on the parameter of interest. However, given that the Bartlett identities are central to likelihood-based statistical inference, it may be desirable in some cases to choose $z$ and $H$ so that these identities are at least approximately satisfied by the marginal or conditional likelihood under consideration.

***Example 2 Exponential regression.*** Let $Y_1, \ldots, Y_n$ denote independent exponential random variables such that $Y_j$ has mean $\lambda \exp(\psi x_j)^{-1}$, where $x_1, \ldots, x_n$ are fixed scalar constants and $\psi$ and $\lambda$ are unknown scalar parameters. Here the data are denoted by $y = (y_1, \ldots, y_n)$. The log-likelihood function for this model is

$$\ell(\psi, \lambda) = -n \log \lambda - \psi \sum x_j - \frac{1}{\lambda} \sum \exp(-\psi x_j) y_j.$$

For fixed $\psi$, $s_\psi = \sum \exp(-\psi x_j) y_j$ is sufficient and, aside from an additive constant,

$$\log p(y;\, \theta) - \log p(s_\psi;\, \theta) = -\psi \sum x_j - (n-1) \log \sum \exp(-\psi x_j) y_j.$$

In order to complete specification of the conditional likelihood, we need to choose $z$ and $H(z)$. Three choices are considered here, leading to three different conditional likelihoods.

First, consider $z = (y_1, \ldots, y_n)$ with $H$ taken to be the identity matrix. Then the conditional log-likelihood is given by

$$-\psi \sum x_j - (n-1) \log \sum \exp(-\psi x_j) y_j - \frac{1}{2} \log \sum \exp(-2\psi x_j). \tag{3}$$

Note that this model is unchanged if each $y_j$ is transformed to $\alpha y_j$, $\alpha > 0$. Hence, $H$ should satisfy $H(\alpha z) = \alpha^2 H(z)$. The choice of $H$ as an identity matrix, which was used to derive (3), does not satisfy this condition.

Perhaps the simplest matrix $H$ satisfying $H(\alpha z) = \alpha^2 H(z)$ is a diagonal matrix with $j$th diagonal element equal to $z_j^2$, which leads to a second choice for the conditional likelihood. It is easy to see that this choice is equivalent to using the volume element

$$\left| \frac{\partial s_\psi}{\partial w} \left( \frac{\partial s_\psi}{\partial w} \right) \right|^{1/2}$$

where $w = (w_1, \ldots, w_n)$, $w_j = \log y_j$. The resulting conditional log-likelihood function is

$$-\psi \sum x_j - (n-1) \log \sum \exp(-\psi x_j) y_j - \frac{1}{2} \log \sum \exp(-2\psi x_j) y_j^2. \tag{4}$$

Ancillarity can be used to suggest a third choice for $H$. Note that the model for $y_1, \ldots, y_n$ corresponds to a linear model for $w_1, \ldots, w_n$. Let $\hat{\alpha}$ and $\hat{\beta}$ denote the least-squares estimators of $\alpha$ and $\beta$, respectively, in the model $w_j = \alpha + \beta x_j + e_j$; then the residuals $a_j = w_j - \hat{\alpha} - \hat{\beta} x_j$, $j = 1, \ldots, n$, are ancillary in the full model.

It is straightforward to show that

$$\frac{\partial s_\psi}{\partial \hat{\alpha}} = \sum \exp(-\psi x_j) y_j \quad \text{and} \quad \frac{\partial s_\psi}{\partial \hat{\beta}} = \sum \exp(-\psi x_j) x_j y_j,$$

and that the covariance matrix of $(\hat{\alpha}, \hat{\beta})$ is proportional to

$$\Sigma = \begin{pmatrix} \sum x_j^2 / n & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}.$$

The volume element based on $z = (\hat{\alpha}, \hat{\beta})$ and $H(z) = \Sigma$ leads to the conditional log-likelihood

$$-\psi \sum x_j - (n-1)\log \sum \exp(-\psi x_j) y_j$$

$$-\frac{1}{2}\log\left\{\left[\sum \exp(-\psi x_j) y_j\right]^2 + \frac{\left[\sum \exp(-\psi x_j)(x_j - \bar{x}) y_j\right]^2}{\sum (x_j - \bar{x})^2/n}\right\}. \tag{5}$$

These three conditional log-likelihoods may be compared based on the expected value of the corresponding conditional score function, that is, based on how close they are to satisfying the first Bartlett identity; the expected value of the score function based on a given pseudo-likelihood function is sometimes referred to as the *score bias* of the pseudo-likelihood. For (3), the score bias is

$$\frac{\sum \exp(-2\psi x_j)(x_j - \bar{x})}{\sum \exp(-2\psi x_j)} = O(1).$$

For (4) it is $O(n^{-2})$, while for (5) it is

$$\frac{\sum x_j (x_j - \bar{x})^2}{\sum (x_j - \bar{x})^2}\frac{1}{n} + O(n^{-2}).$$

Pseudo-likelihood (3), which is based on the naive choice of $z = y$ with $H(z)$ taken to be an identity matrix, has little to recommend it. Pseudo-likelihood (4), which is based on the transformation structure of the model and satisfies the first Bartlett identity to a high degree of approximation, would be a reasonable choice for likelihood inference in this problem. However, if conditioning on an ancillary statistic is thought to be an important consideration, (5) may be a more appropriate choice.

## 2.4. Equivalence of marginal and conditional likelihoods in exponential family models

For some models, both a marginal and a conditional likelihood function are available. This is particularly true if likelihoods based on parameter-dependent functions are considered since then the ancillary statistic or sufficient statistic in the model with $\psi$ fixed may depend on the value of $\psi$ under consideration. Suppose that the sufficient statistic for the full model is equivalent to $(T_\psi, S_\psi)$ where, in the model with $\psi$ held fixed, $S_\psi$ is sufficient and $T_\psi$ is ancillary. Furthermore, suppose that $S_\psi$ is complete for fixed $\psi$; this holds, in particular, if the model with fixed $\psi$ is a full-rank exponential family model. Then, by Basu's theorem (Basu 1955; 1958), $T_\psi$ and $S_\psi$ are independent.

Using the volume based on the statistic $z$ and matrix $H(z)$, the conditional likelihood given $S_\psi$ is given by

$$p(t_\psi;\,\psi)\frac{|(\partial(t_\psi,\,s_\psi)/\partial z)H(z)(\partial(t_\psi,\,s_\psi)/\partial z)^{\mathrm{T}}|^{1/2}}{|(\partial s_\psi/\partial z)H(z)(\partial s_\psi/\partial z)^{\mathrm{T}}|^{1/2}}$$

and the marginal likelihood function based on $T_\psi$ is given by

$$p(t_\psi;\,\psi)\left|\frac{\partial t_\psi}{\partial z}H(z)\left(\frac{\partial t_\psi}{\partial z}\right)^{\mathrm{T}}\right|^{1/2}.$$

It is straightforward to show a sufficient condition for the marginal and conditional likelihoods to agree is that

$$\frac{\partial t_\psi}{\partial z}H(z)\left(\frac{\partial s_\psi}{\partial z}\right)^{\mathrm{T}}=0. \tag{6}$$

***Example 1 Ratio of normal means (continued).*** Taking $z=(x,\,y)$ and $H(z)$ equal to the identity matrix, we have seen that the marginal likelihood based on $t_\psi=x-\psi y$ and the conditional likelihood given $s_\psi=\psi x+y$ are identical. Note that condition (6) is satisfied.

***Example 2 Exponential regression (continued).*** The distribution of $t=(y_2/y_1,\,y_3/y_1,\,\ldots,\,y_n/y_1)$ depends only on $\psi$; the marginal likelihood based on $t$ is given by

$$-\psi\sum x_j-n\log\sum\exp(-\psi x_j)y_j.$$

Note that, since $t$ does not depend on $\psi$, the volume element does not need to be specified.

This marginal likelihood differs from the conditional likelihoods derived for this example. Suppose that $z$ is given by $(\log y_1,\,\ldots,\,\log y_n)$ and $H$ is taken to be a diagonal matrix with $j$th diagonal element $d_j$. The resulting conditional likelihood given $s_\psi$ is identical to the marginal likelihood based on $t$ provided that

$$\sum d_j\exp(-2\psi x_j)y_j^2=\sum\exp(-\psi x_j)y_j,\qquad -\infty<\psi<\infty;$$

clearly, this cannot occur unless $d_j$ depends on $\psi$.

Now suppose that $z=(\hat\alpha,\,\hat\beta)$ so that the volume element holds $a_1,\,\ldots,\,a_n$ fixed. The resulting conditional likelihood given $s_\psi$ is identical to the marginal likelihood based on $t$ provided that $H$ is taken to be of the form

$$H(z)=\begin{pmatrix}c&0\\0&0\end{pmatrix},$$

where $c$ does not depend on $\psi$. This volume element depends only on variation in $\hat\alpha$, while $t$

is a function of $a_1, \dots, a_n$ and $\hat{\beta}$. These results suggest that, for this model, the marginal and conditional approaches to likelihood inference for $\psi$ are fundamentally different.

# 3. Likelihood functions based on maximum likelihood estimators

## 3.1. A general framework for marginal and conditional likelihoods

We now consider the construction of likelihood functions for a parameter of interest for models satisfying the following conditions. In the full model, we assume that the sufficient statistic may be written $(\hat{\theta}, a)$, where $a$ is ancillary; similarly, in the model with $\psi$ held fixed, we assume that the sufficient statistic may be written $(\hat{\theta}_\psi, b_\psi, a)$, where $b_\psi$ is ancillary. These assumptions are satisfied if the models are exponential family models or transformation models, and they are satisfied to a sufficient degree of approximation for a wide range of models; see, for example, Severini (2000a, Chapter 6) for a detailed discussion of these issues. We assume that these conditions hold for the remainder of the paper.

Hence, the distribution of $b_\psi$ may depend on $\psi$ but not on $\lambda$. For simplicity, we assume that the parameter of the model may be written $\theta = (\psi, \lambda)$ for some nuisance parameter $\lambda$; in Section 3.4, the likelihood functions derived in this section are expressed in a form that does not require an explicit nuisance parameter. Throughout this discussion we condition on the ancillary $a$, although, for simplicity, this conditioning is not always explicitly stated.

Furthermore, we assume that, for both models, the likelihood ratio approximation to the conditional density of the maximum likelihood estimator given an ancillary statistic, also called the $p^*$ formula, holds. For a model with parameter $\theta$ this approximation is given by

$$p^*(\hat{\theta}|a; \theta) = c(\theta; a)|\hat{\jmath}|^{1/2} \exp\{\ell(\theta) - \ell(\hat{\theta})\}$$

(Barndorff-Nielsen 1980; 1983). The constant $c$ is of the form $c(\theta; a) = (2\pi)^{-q/2}\{1 + O(n^{-3/2})\}$, where $q$ denotes the dimension of $\theta$. Hence, for $\theta = \hat{\theta} + O(n^{-1/2})$, $c$ depends only on the data, neglecting terms of order $O(n^{-3/2})$.

We first show that, in this setting, a conditional or marginal likelihood function based on a parameter-dependent function always exists. Consider the model with $\psi$ held fixed. In this model, either $(\hat{\lambda}_\psi, a)$ is sufficient or it is not sufficient. If it is sufficient, the conditional distribution of the data given $\hat{\lambda}_\psi$ depends only on $\psi$ and this conditional distribution may be used to form a conditional likelihood. This approach is considered in Section 3.2.

If, in the model with $\psi$ held fixed, $(\hat{\lambda}_\psi, a)$ is not sufficient, then a sufficient statistic is given by $(\hat{\lambda}_\psi, b_\psi, a)$ and $b_\psi$ is ancillary in the model with $\psi$ fixed; that is, the distribution of $b_\psi$ depends only on $\psi$. We now consider the construction of a marginal likelihood function on the distribution of $b_\psi$.

Since $b_\psi$ is ancillary in the model with $\psi$ fixed, the conditional density of $\hat{\lambda}_\psi$ given $b_\psi$ may be approximated by

$$p^*(\hat{\lambda}_\psi | a, b_\psi; \hat{\theta}_\psi) = c_1 |j_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2} \exp\{\ell(\theta) - \ell(\hat{\theta}_\psi)\},$$

where $j_{\lambda\lambda}(\hat{\theta}_\psi)$ denotes the observed information for fixed $\psi$, evaluated at $\hat{\theta}_\psi$. The density of $(\hat{\lambda}_\psi, b_\psi)$ given $a$ may be approximated by transforming $p^*(\hat{\theta}|a; \theta)$:

$$p^*(\hat{\lambda}_\psi, b_\psi | a; \theta) = c_2 |\hat{\jmath}|^{1/2} \exp\{\ell(\theta) - \ell(\hat{\theta})\} \left| \frac{\partial \hat{\theta}}{\partial(\hat{\lambda}_\psi, b_\psi)} \right|.$$

Here $\hat{\jmath}$ denotes the observed information. Since the marginal density of $b_\psi$ is given by

$$\frac{p(\hat{\lambda}_\psi, b_\psi | a; \theta)}{p(\hat{\lambda}_\psi | b_\psi, a; \theta)},$$

it may be approximated by

$$p^*(b_\psi | a; \theta) = c \frac{|\hat{\jmath}|^{1/2} |\partial\hat{\theta}/\partial(\hat{\lambda}_\psi, b_\psi)|}{|j_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2}} \exp\{\ell(\hat{\theta}_\psi) - \ell(\hat{\theta})\}.$$

In order to complete specification of the marginal likelihood function, we need to specify the variable $z$ and the matrix $H(z)$. Since $(\hat{\theta}, a)$ is sufficient and $a$ is ancillary, we may take $z = \hat{\theta}$ and then specify the volume element by choosing the matrix $H(\hat{\theta})$.

The matrix $H(\hat{\theta})$ should be chosen so that the assumption that a Euclidean metric is appropriate for $H(\hat{\theta})^{-1/2}\,\mathrm{d}\hat{\theta}$ is reasonable. Note that $\hat{\jmath}^{1/2}(\hat{\theta} - \theta)$ is approximately distributed according to a multivariate normal distribution with mean vector 0 and covariance matrix equal to the identity matrix. Furthermore, this result holds conditionally on $a$. This suggests that a Euclidean metric is appropriate for $\hat{\jmath}^{1/2}\,\mathrm{d}\hat{\theta}$; that is, this suggests that a reasonable choice for $H$ is $H(\hat{\theta}) = \hat{\jmath}^{-1}$. Note that the resulting volume is locally invariant under transformations of $\theta$. This metric is discussed further by Fraser and Reid (1989), who refer to it as the *constant information metric*.

The marginal likelihood based on $b_\psi$ using this approximation and the constant information metric is therefore given by

$$L_b(\psi) = |j_{\lambda\lambda}(\hat{\theta}_\psi)|^{-1/2} L_p(\psi) \left| \frac{\partial\hat{\theta}}{\partial(\hat{\lambda}_\psi, b_\psi)} \right| \left| \frac{\partial b_\psi}{\partial\hat{\theta}} \hat{\jmath}^{-1} \left(\frac{\partial b_\psi}{\partial\hat{\theta}}\right)^{\mathrm{T}} \right|^{1/2}.$$

Here $L_p$ is the profile likelihood function. It is straightforward to show that, by rewriting the differential terms,

$$L_b(\psi) = |j_{\lambda\lambda}(\hat{\theta}_\psi)|^{-1/2} \cdot \left| \frac{\partial\hat{\lambda}_\psi}{\partial\hat{\theta}} \hat{\jmath}^{-1} \left(\frac{\partial\hat{\lambda}_\psi}{\partial\hat{\theta}}\right)^{\mathrm{T}} - \frac{\partial\hat{\lambda}_\psi}{\partial\hat{\theta}} \hat{\jmath}^{-1} \left(\frac{\partial b_\psi}{\partial\hat{\theta}}\right)^{\mathrm{T}} \left[ \frac{\partial b_\psi}{\partial\hat{\theta}} \hat{\jmath}^{-1} \left(\frac{\partial b_\psi}{\partial\hat{\theta}}\right)^{\mathrm{T}} \right]^{-1} \frac{\partial b_\psi}{\partial\hat{\theta}} \hat{\jmath}^{-1} \left(\frac{\partial\hat{\lambda}_\psi}{\partial\hat{\theta}}\right)^{\mathrm{T}} \right|^{-1/2} L_p(\psi). \tag{7}$$

This expression yields a family of marginal likelihoods for $\psi$ based on the specific choice used for the statistic $b_\psi$. The case in which $b_\psi$ is taken to be null, so that $\partial b_\psi/\partial\hat{\theta} = 0$, corresponds to the case in which $\hat{\lambda}_\psi$ itself is sufficient for fixed $\psi$. Hence, in a certain sense, all marginal and conditional likelihood functions for $\psi$ based on the $p^*$ approximation and the constant information metric may be generated by (7) using different choices for $b_\psi$. In Section 3.3, we take $b_\psi$ to be $r_\psi$, the signed likelihood ratio ancillary in the model with $\psi$ fixed.

## 3.2. Conditional modified profile likelihood

Suppose that, in the model with $\psi$ fixed, $(\hat{\lambda}_\psi, a)$ is a sufficient statistic. A conditional likelihood function for $\psi$ may then be based on an approximation to the conditional distribution of the data given $\hat{\lambda}_\psi$; as noted above, this corresponds to taking $b_\psi$ to be null in (7). Then

$$\frac{\partial b_\psi}{\partial \hat{\theta}} = 0$$

and the denominator in (7) becomes

$$\left| \frac{\partial \hat{\lambda}_\psi}{\partial \hat{\theta}} \, \hat{\jmath}^{-1} \left( \frac{\partial \hat{\lambda}_\psi}{\partial \hat{\theta}} \right)^{\mathrm{T}} \right|^{1/2};$$

this leads to the conditional likelihood function

$$L_C(\psi) = |j_{\lambda\lambda}(\hat{\theta}_\psi)|^{-1/2} \left| \frac{\partial \hat{\lambda}_\psi}{\partial \hat{\theta}} \, \hat{\jmath}^{-1} \left( \frac{\partial \hat{\lambda}_\psi}{\partial \hat{\theta}} \right)^{\mathrm{T}} \right|^{-1/2} L_p(\psi).$$

An alternative expression for $L_C$ is sometimes useful. Since $\hat{\lambda}_\psi$ satisfies $\ell_\lambda(\psi, \hat{\lambda}_\psi; \hat{\theta}) = 0$, it follows that

$$\ell_{\lambda\lambda}(\psi, \hat{\lambda}_\psi) \frac{\partial \hat{\lambda}_\psi}{\partial \hat{\theta}} + \ell_{\lambda;\hat{\theta}}(\psi, \hat{\lambda}_\psi) = 0$$

where

$$\ell_{\lambda;\hat{\theta}}(\theta) = \frac{\partial}{\partial \hat{\theta}} \ell_\lambda(\theta; \hat{\theta}).$$

Hence,

$$\frac{\partial \hat{\lambda}_\psi}{\partial \hat{\theta}} = \hat{\jmath}_{\lambda\lambda}(\hat{\theta}_\psi)^{-1} \ell_{\lambda;\hat{\theta}}(\hat{\lambda}_\psi).$$

Substituting this in the expression for $L_C$ above shows that

$$L_C(\psi) = \frac{| \, \hat{\jmath}_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2}}{|\ell_{\lambda;\hat{\theta}}(\hat{\theta}_\psi) \, \hat{\jmath}^{-1} \ell_{\lambda;\hat{\theta}}(\hat{\theta}_\psi)^{\mathrm{T}}|^{1/2}} L_p(\psi).$$

The pseudo-likelihood $L_C$ will be called the *conditional modified profile likelihood*. This approximation to a conditional likelihood was also derived by Fraser and Reid (1989) using a closely related, but slightly different approach; its connection to the modified profile likelihood of Barndorff-Nielsen (1983) will be discussed in Section 4.

It is straightforward to show that, when $(\hat{\lambda}_\psi, a)$ is sufficient, $L_C$ approximates the true conditional likelihood function given $\hat{\lambda}_\psi$ to $O(n^{-3/2})$ for $\psi$ of the form $\hat{\psi} + O(n^{-1/2})$.

## 3.3. Likelihood ratio modified profile likelihood

We now consider an approximation to the marginal likelihood function based on the distribution of the signed likelihood ratio statistic

$$R \equiv R_\psi \equiv \operatorname{sgn}(\hat{\psi} - \psi)\{2[\ell(\hat{\theta}) - \ell(\hat{\theta}_\psi)]\}^{1/2}.$$

Using (7), along with the fact that

$$\frac{\partial r}{\partial \hat{\theta}} = \frac{1}{r}\{\ell_{;\hat{\theta}}(\hat{\theta}) - \ell_{;\hat{\theta}}(\hat{\theta}_\psi)\}, \qquad \ell_{;\hat{\theta}}(\theta) = \frac{\partial}{\partial \hat{\theta}}\ell(\theta; \hat{\theta}),$$

leads to the likelihood function

$$L_R(\psi) = \frac{|\,\hat{\jmath}_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2}}{|\ell_{\lambda;\hat{\theta}}(\hat{\theta}_\psi)\,\hat{\jmath}^{-1}\ell_{\lambda;\hat{\theta}}(\hat{\theta}_\psi)^{\mathrm{T}} - D|^{1/2}} L_p(\psi),$$

where the matrix $D$ is given by

$$D = \frac{\ell_{\lambda;\hat{\theta}}(\hat{\theta}_\psi)\,\hat{\jmath}^{-1}\{\ell_{;\hat{\theta}}(\hat{\theta}) - \ell_{;\hat{\theta}}(\hat{\theta}_\psi)\}^{\mathrm{T}}[\ell_{\lambda;\hat{\theta}}(\hat{\theta}_\psi)\,\hat{\jmath}^{-1}\{\ell_{;\hat{\theta}}(\hat{\theta}) - \ell_{;\hat{\theta}}(\hat{\theta}_\psi)\}^{\mathrm{T}}]^{\mathrm{T}}}{\{\ell_{;\hat{\theta}}(\hat{\theta}) - \ell_{;\hat{\theta}}(\hat{\theta}_\psi)\}\,\hat{\jmath}^{-1}\{\ell_{;\hat{\theta}}(\hat{\theta}) - \ell_{;\hat{\theta}}(\hat{\theta}_\psi)\}^{\mathrm{T}}}.$$

An alternative expression for $L_R$ is

$$L_R(\psi) = |\,\hat{\jmath}_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2}$$
$$\cdot \begin{vmatrix} \ell_{;\hat{\theta}}(\hat{\theta}) - \ell_{;\hat{\theta}}(\hat{\theta}_\psi) \\ \ell_{\lambda;\hat{\theta}}(\hat{\theta}_\psi) \end{vmatrix} |\{\ell_{;\hat{\theta}}(\hat{\theta}) - \ell_{;\theta}(\hat{\theta}_\psi)\}\,\hat{\jmath}^{-1}\{\ell_{;\hat{\theta}}(\hat{\theta}) - \ell_{;\theta}(\hat{\theta}_\psi)\}^{\mathrm{T}}|^{1/2} L_p(\psi).$$

We will call $L_R$ the *likelihood ratio modified profile likelihood*; its connection to the modified profile likelihood of Barndorff-Nielsen (1983) will be discussed in Section 4.

If the distribution of $R$ depends only on $\psi$, then $L_R(\psi)$ approximates the true marginal likelihood function based on $R$ to $O(n^{-3/2})$ for $\psi = \hat{\psi} + O(n^{-3/2})$. If, for fixed $\psi$, $R$ is a second-order ancillary statistic, so that the marginal likelihood function based on $R$ depends only on $\psi$ to $O(n^{-3/2})$, then $L_R(\psi)$ approximates this marginal likelihood function with error $O(n^{-3/2})$. In the general case, $R$ is a first-order ancillary statistic so that the marginal likelihood function based on $R$ depends only on $\psi$ to $O(n^{-1})$. In this case, $L_R(\psi)$ approximates the marginal likelihood function based on $R$ with error $O(n^{-1})$.

### 3.4. Expressions for $L_C$ and $L_R$ that do not require an explicit nuisance parameter

The expressions for $L_C$ and $L_R$ given earlier are all based on a parametrization of the model of the form $\theta = (\psi, \lambda)$ where $\psi$ is the parameter of interest and $\lambda$ is a nuisance parameter. We now show that it is possible to calculate these likelihoods without choosing an explicit form for the nuisance parameter. That is, consider a model with parameter $\theta$ and let $\psi \equiv \psi(\theta)$ denote the real-valued parameter of interest. We may calculate $L_C$ and $L_R$ based on this structure alone, without explicitly selecting a form for the nuisance parameter $\lambda$. Hence, these results also serve to verify that these likelihoods are invariant under interest-respecting reparametrizations.

First note that an explicit form for $\lambda$ was only used in calculation of the derivatives $\ell_{\lambda;\hat{\lambda}}(\theta)$, $\ell_{\lambda;\hat{\theta}}(\theta)$ and $\ell_{\lambda\lambda}(\theta)$. Let $f(\theta)$ be an arbitrary function of $\theta$ and consider calculation of $f_\lambda(\theta) = \partial f(\theta)/\partial\lambda$. By the chain rule,

$$f_\lambda(\theta) = f_\theta(\theta)\frac{\partial\theta}{\partial\lambda};$$

hence, it is sufficient to obtain an expression for $\partial\theta/\partial\lambda$. Note that

$$\begin{pmatrix} \dfrac{\partial\theta}{\partial\psi} & \dfrac{\partial\theta}{\partial\lambda} \end{pmatrix} = \begin{pmatrix} \dfrac{\partial\psi}{\partial\theta} \\ \dfrac{\partial\lambda}{\partial\theta} \end{pmatrix}^{-1}$$

and, hence, $\partial\theta/\partial\lambda$ satisfies

$$\frac{\partial\lambda}{\partial\theta}\frac{\partial\theta}{\partial\lambda} = I \quad \text{and} \quad \frac{\partial\psi}{\partial\theta}\frac{\partial\theta}{\partial\lambda} = 0;$$

here $I$ represents the identity matrix with rank $q-1$, where $q = \dim(\theta)$.

Let $B \equiv B(\theta)$ denote a $q \times (q-1)$ matrix such that for all $\theta$,

$$\frac{\partial\psi}{\partial\theta}B = 0 \quad \text{and} \quad \left|\frac{\partial\lambda}{\partial\theta}B\right| \neq 0.$$

Then

$$\frac{\partial\theta}{\partial\lambda} = B\left(\frac{\partial\lambda}{\partial\theta}B\right)^{-1}.$$

Hence,

$$\ell_{\lambda;\hat{\theta}}(\hat{\theta}_\psi) = \left[\left\{\frac{\partial\lambda}{\partial\theta}(\hat{\theta}_\psi)B(\hat{\theta}_\psi)\right\}^{-1}\right]^{\mathrm{T}} B(\hat{\theta}_\psi)^{\mathrm{T}}\ell_{\theta;\hat{\theta}}(\hat{\theta}_\psi),$$

$$\ell_{\lambda;\hat{\lambda}}(\hat{\theta}_\psi) = \left[\left\{\frac{\partial\lambda}{\partial\theta}(\hat{\theta}_\psi)B(\hat{\theta}_\psi)\right\}^{-1}\right]^{\mathrm{T}} B(\hat{\theta}_\psi)^{\mathrm{T}}\ell_{\theta;\hat{\theta}}(\hat{\theta}_\psi)B(\hat{\theta})\left[\left\{\frac{\partial\lambda}{\partial\theta}(\hat{\theta})B(\hat{\theta})\right\}^{-1}\right]^{\mathrm{T}},$$

and

$$\hat{\jmath}_{\lambda\lambda}(\hat{\theta}_\psi) = \left[\left\{\frac{\partial\lambda}{\partial\theta}(\hat{\theta}_\psi)B(\hat{\theta}_\psi)\right\}^{-1}\right]^{\mathrm{T}} B(\hat{\theta}_\psi)^{\mathrm{T}}\,\hat{\jmath}(\hat{\theta}_\psi)B(\hat{\theta}_\psi)\left[\left\{\frac{\partial\lambda}{\partial\theta}(\hat{\theta}_\psi)\mathrm{B}(\hat{\theta}_\psi)\right\}^{-1}\right]^{\mathrm{T}}.$$

Define the matrix $P_\psi$ by

$$P_\psi = I - \frac{\psi'(\hat{\theta}_\psi)^{\mathrm{T}}\psi'(\hat{\theta}_\psi)}{\psi'(\hat{\theta}_\psi)\psi'(\hat{\theta}_\psi)^{\mathrm{T}}}$$

where

$$\psi'(\theta) = \frac{\mathrm{d}}{\mathrm{d}\psi}\psi(\theta).$$

For a given $q\times q$ non-singular matrix $M_0$, let $v_1, \ldots, v_{q-1}$ denote the eigenvectors of $(I-P_\psi)M_0$ corresponding to the non-zero eigenvalues $\eta_1, \ldots, \eta_{q-1}$. Then the matrix $B \equiv B(\hat{\theta}_\psi)$ needed for calculation of the sample space derivatives is of the form $VM_1$, where $V$ is the matrix with $j$th column $v_j$ and $M_1$ is an arbitrary $(q-1)\times(q-1)$ non-singular matrix. It follows that

$$|B^{\mathrm{T}}M_0B|^{1/2} = |M_1|(\eta_1\cdots\eta_{q-1})^{1/2} \quad \text{and} \quad |B^{\mathrm{T}}M_0| = |M_1|(\eta_1\cdots\eta_{q-1}).$$

For a $q\times q$ non-singular matrix $M$ define $|M|_\psi$ to be the product of the non-zero eigenvalues of $(I-P_\psi)M$. Then we may write

$$L_C(\psi) = \frac{|j(\hat{\theta}_\psi)|_\psi^{1/2}}{|\ell_{\theta;\hat{\theta}}(\hat{\theta}_\psi)\,\hat{\jmath}^{-1}\ell_{\theta;\hat{\theta}}(\hat{\theta}_\psi)^{\mathrm{T}}|_\psi^{1/2}}L_p(\psi)$$

and

$$L_R(\psi) = \frac{|j(\hat{\theta}_\psi)|_\psi^{1/2}}{|\ell_{\theta;\hat{\theta}}(\hat{\theta}_\psi)\,\hat{\jmath}^{-1}\ell_{\theta;\hat{\theta}}(\hat{\theta}_\psi)^{\mathrm{T}} - D_0|_\psi^{1/2}}L_p(\psi),$$

where

$$D_0 = \frac{\ell_{\theta;\hat{\theta}}(\hat{\theta}_\psi)\,\hat{\jmath}^{-1}\{\ell_{;\hat{\theta}}(\hat{\theta}) - \ell_{;\hat{\theta}}(\hat{\theta}_\psi)\}^{\mathrm{T}}[\ell_{\theta;\hat{\theta}}(\hat{\theta}_\psi)\,\hat{\jmath}^{-1}\{\ell_{;\hat{\theta}}(\hat{\theta}) - \ell_{;\hat{\theta}}(\hat{\theta}_\psi)\}^{\mathrm{T}}]^{\mathrm{T}}}{\{\ell_{;\hat{\theta}}(\hat{\theta}) - \ell_{;\hat{\theta}}(\hat{\theta}_\psi)\}\,\hat{\jmath}^{-1}\{\ell_{;\hat{\theta}}(\hat{\theta}) - \ell_{;\hat{\theta}}(\hat{\theta}_\psi)\}^{\mathrm{T}}}.$$

## 4. Relationships with the modified profile likelihood

### 4.1. Modified profile likelihood

The modified profile likelihood function is given by

$$L_M(\psi) = \left|\frac{\partial\hat{\lambda}_\psi}{\partial\hat{\lambda}}\right|^{-1} \left|\hat{\jmath}_{\lambda\lambda}(\hat{\theta}_\psi)\right|^{-1/2} L_p(\psi) = \frac{|\hat{\jmath}_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2}}{|\ell_{\lambda;\hat{\lambda}}(\hat{\theta}_\psi)|} L_p(\psi).$$

Here we give a brief overview of the properties of $L_M$; see Barndorff-Nielsen and Cox (1994, Chapter 8) for a more detailed discussion.

The modified profile likelihood is derived as an approximation to either a conditional or a marginal likelihood. Let $\hat{\psi}$ denote the maximum likelihood estimator of $\psi$ and let $\hat{\lambda}$ denote the maximum likelihood estimator of $\lambda$. First, suppose that $\hat{\lambda}$ is sufficient in the model with $\psi$ held fixed. A conditional likelihood function may be based on the density of $\hat{\psi}$ given $\hat{\lambda}$, which is given by

$$p(\hat{\psi}|\hat{\lambda}, a; \psi) = \frac{p(\hat{\psi}, \hat{\lambda}|a; \psi, \lambda)}{p(\hat{\lambda}|a; \psi, \lambda)}. \tag{8}$$

The density $p(\hat{\psi}, \hat{\lambda}|a; \psi, \lambda)$ may be approximated using the $p^*$ approximation. In order to approximate $p(\hat{\lambda}|a; \psi, \lambda)$, we first approximate $p(\hat{\lambda}_\psi|a; \psi, \lambda)$ using the $p^*$ approximation in the model with $\psi$ fixed; since $\hat{\lambda}$ is sufficient for fixed $\psi$, $\hat{\lambda}_\psi$ must be a function of $\hat{\lambda}$. Hence, an approximation to $p(\hat{\lambda}|a; \psi, \lambda)$ may be obtained using the usual change-of-variable formula for density functions. Substituting these approximations in (8) yields $L_M$.

Alternatively, suppose that the distribution of $\hat{\psi}$ does not depend on $\lambda$. Note that

$$p(\hat{\psi}|a; \psi) = \frac{p(\hat{\psi}, \hat{\lambda}|a; \psi, \lambda)}{p(\hat{\lambda}|\hat{\psi}, a; \psi, \lambda)} \tag{9}$$

and a marginal likelihood function for $\psi$ can be based on (9). The density $p(\hat{\psi}, \hat{\lambda}|a; \psi, \lambda)$ can be approximated using the $p^*$ approximation. The density $p(\hat{\lambda}|\hat{\psi}, a; \psi, \lambda)$ can also be approximated using the $p^*$ approximation by noting that, in the model with $\psi$ held fixed, $\hat{\psi}$ is ancillary. Substituting these approximations in (9) yields $L_M$.

Whenever either of these assumptions holds, $L_M(\psi)$ approximates the conditional or marginal likelihood to order $O(n^{-3/2})$ for $\psi$ of the form $\hat{\psi} + O(n^{-1/2})$. In general – when these assumptions do not necessarily hold – the marginal likelihood based on $\hat{\psi}$ is independent of $\lambda$ to $O(n^{-1/2})$, that is to say, it does not depend on $\psi$ excluding terms of order $O(n^{-1/2})$ and $L_M(\psi)$ approximates this marginal likelihood function to $O(n^{-1/2})$. Since, for general models, $\hat{\lambda}$ is not even sufficient to a first-order approximation (Severini 1994) whenever $(\hat{\lambda}, a)$ is not sufficient for fixed $\psi$, $L_M(\psi)$ cannot interpreted as an approximation to a conditional likelihood.

It is worth noting that when $\hat{\lambda}_\psi$ is sufficient for fixed $\psi$, $L_M(\psi)$ may also be derived by conditioning on $\hat{\lambda}_\psi$ by using

$$p^*(\hat{\psi}|\hat{\lambda}_\psi, a; \psi) = \frac{|\hat{\jmath}|^{1/2}\exp\{\ell(\theta) - \ell(\hat{\theta})\}|\partial\hat{\theta}/\partial(\hat{\psi}, \hat{\lambda}_\psi)|}{|\hat{\jmath}_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2}\exp\{\ell(\theta) - \ell(\hat{\theta}_\psi)\}}.$$

Simplifying this expression and ignoring terms not depending on the data, this approximation may be seen to be exactly $L_M$. Equivalently, we may derive $L_M$ by using the same approximation used in the derivation of $L_C$ with the differential term based on $z = \hat{\theta}$ and

$$H(z) = \begin{pmatrix} 0 & 0 \\ 0 & H_{\lambda\lambda} \end{pmatrix},$$

where $H_{\lambda\lambda}$ is any full rank matrix not depending on $\psi$.

Barndorff-Nielsen (1994, Section 6.2) shows that the modified profile likelihood function $L_M(\psi)$ may be derived as an approximation to the marginal likelihood function based on the modified signed likelihood ratio statistic $R^*_\psi$. The modified signed likelihood ratio statistic is given by

$$R^* = R + \frac{1}{R}\log\left(\frac{U}{R}\right),$$

where $R$ is the signed likelihood ratio statistic and

$$U = \begin{vmatrix} \ell_{;\hat{\theta}}(\hat{\theta}) - \ell_{;\hat{\theta}}(\hat{\theta}_\psi) \\ \ell_{\lambda;\hat{\theta}}(\hat{\theta}_\psi) \end{vmatrix} |\,\hat{\jmath}_{\lambda\lambda}(\hat{\theta}_\psi)|^{-1/2}|\,\hat{\jmath}|^{-1/2}.$$

Using the fact that $r^* = r + O(n^{-1/2})$,

$$\exp\left\{-\frac{(r^*)^2}{2}\right\} = \frac{u}{r}\exp\left\{-\frac{r^2}{2}\right\}[1 + O(n^{-1})].$$

Hence, for the portion of the marginal likelihood based on the $p(r^*;\,\psi)$, Barndorff-Nielsen uses

$$\frac{u}{r}\exp\left\{-\frac{r^2}{2}\right\} \propto \frac{u}{r}L_p(\psi).$$

The volume element used is based on

$$\left|\frac{\partial r^*}{\partial\hat{\psi}}\right| = \left|\frac{\partial r^*(r,\,\hat{\lambda}_\psi)}{\partial r}\,\frac{\partial r(\hat{\psi},\,\hat{\lambda}_\psi)}{\partial\hat{\psi}}\right|. \tag{10}$$

According to Barndorff-Nielsen (1994, Section 6.2), $\partial r^*/\partial r$ may be approximated by 1 and

$$\frac{\partial r(\hat{\psi},\,\hat{\lambda}_\psi)}{\partial\hat{\psi}} = \frac{|\,\hat{\jmath}_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2}|\,\hat{\jmath}|^{1/2}}{|\ell_{\lambda;\hat{\lambda}}(\hat{\theta}_\psi)|}\frac{u}{r}$$

so that the resulting marginal likelihood is $L_M$.

In order to compare this approach to the one used in deriving $L_R$, we may write the differential term corresponding to (10) in the notation

$$\left|\frac{\partial r}{\partial\hat{\theta}}H(\hat{\theta})\left(\frac{\partial r}{\partial\hat{\theta}}\right)^{\mathrm{T}}\right|^{1/2}.$$

Hence, the differential term used by Barndorff-Nielsen corresponds to taking $H = H_M$, where

$$H_M = \begin{pmatrix} 1 & -\ell_{\lambda;\hat{\psi}}(\hat{\theta}_\psi)^{\mathrm{T}}\{\ell_{\lambda;\hat{\lambda}}(\hat{\theta}_\psi)^{-1}\}^{\mathrm{T}} \\ -\ell_{\lambda;\hat{\lambda}}(\hat{\theta}_\psi)^{-1}\ell_{\lambda;\hat{\psi}}(\hat{\theta}_\psi) & \ell_{\lambda;\hat{\lambda}}(\hat{\theta}_\psi)^{-1}\ell_{\lambda;\hat{\psi}}(\hat{\theta}_\psi)\ell_{\lambda;\hat{\psi}}(\hat{\theta}_\psi)^{\mathrm{T}}\{\ell_{\lambda;\hat{\lambda}}(\hat{\theta}_\psi)^{-1}\}^{\mathrm{T}} \end{pmatrix}.$$

Note that the matrix $H_M$, in addition to depending on $\psi$, is not of full rank. Therefore, $L_M(\psi)$ may be viewed as an approximation to the marginal likelihood based on $R$, using the differential term based on $H_M$; from this point of view $L_M$ has the same properties as an approximation to the marginal likelihood as does $L_R$, the only difference being in the choice of metric used.

## 4.2. Large-sample comparison

The likelihood functions $L_M$, $L_C$, and $L_R$ have very similar behaviour in large samples. In particular, in this subsection it is shown that $L_M(\psi)$, $L_C(\psi)$ and $L_R(\psi)$ all agree to order $O(n^{-1})$ for $\psi$ of the form $\psi = \hat{\psi} + O(n^{-1/2})$; for example,

$$\ell_M(\psi) - \ell_M(\hat{\psi}) = \ell_C(\psi) - \ell_C(\hat{\psi}) + O(n^{-1}),$$

where $\ell_M = \log L_M$ and $\ell_C = \log L_C$.

First, consider the term $\ell_{\lambda;\hat{\theta}}(\hat{\theta}_\psi)\,\hat{\jmath}^{-1}\ell_{\lambda;\hat{\theta}}(\hat{\theta}_\psi)^{\mathrm{T}}$. Throughout the following argument assume that $\psi$ and $\hat{\psi}$ differ by $O(n^{-1/2})$. Note that we may write

$$\ell_{\lambda;\hat{\psi}}(\hat{\theta}_\psi) = \hat{\jmath}_{\lambda\psi} + M_\psi(\hat{\psi} - \psi) + O(1) \quad \text{and} \quad \ell_{\lambda;\hat{\lambda}}(\hat{\theta}_\psi) = \hat{\jmath}_{\lambda\lambda} + M_\lambda(\hat{\psi} - \psi) + O(1),$$

where $M_\psi$ and $M_\lambda$ are $O(n)$ matrices depending only on the data. A tedious, but elementary, calculation shows that

$$\ell_{\lambda;\hat{\theta}}(\hat{\theta}_\psi)\,\hat{\jmath}^{-1}\ell_{\lambda;\hat{\theta}}(\hat{\theta}_\psi)^{\mathrm{T}} = \hat{\jmath}_{\lambda\lambda} + (M_\lambda + M_\lambda^{\mathrm{T}})(\hat{\psi} - \psi) + O(1)$$

so that

$$\frac{1}{2}\log|\ell_{\lambda;\hat{\theta}}(\hat{\theta}_\psi)\,\hat{\jmath}^{-1}\ell_{\lambda;\hat{\theta}}(\hat{\theta}_\psi)^{\mathrm{T}}| = \frac{1}{2}\log|\,\hat{\jmath}_{\lambda\lambda}| + \mathrm{tr}(\,\hat{\jmath}_{\lambda\lambda}^{-1}H_\lambda)(\hat{\psi} - \psi) + O(n^{-1}).$$

A similar calculation shows that

$$\log|\ell_{\lambda;\hat{\lambda}}(\hat{\theta}_\psi)| = \log|\,\hat{\jmath}_{\lambda\lambda}| + \mathrm{tr}(\hat{\jmath}_{\lambda\lambda}^{-1}M_\lambda)(\hat{\psi} - \psi) + O(n^{-1}),$$

which establishes the result for $L_M$ and $L_C$.

To establish this result for $L_R$ it suffices to show that the matrix $D$ is $O(1)$. Then

$$\ell_{\lambda;\hat{\theta}}(\hat{\theta}_\psi)\,\hat{\jmath}^{-1}\ell_{\lambda;\hat{\theta}}(\hat{\theta}_\psi)^{\mathrm{T}} - D = \hat{\jmath}_{\lambda\lambda} + (M_\lambda + M_\lambda^{\mathrm{T}})(\hat{\psi} - \psi) + O(1)$$

as above, which yields the result. It is straightforward to show that $\ell_{;\hat{\theta}}(\hat{\theta}) - \ell_{;\hat{\theta}}(\hat{\theta}_\psi) = O(\sqrt{n})$ so that

$$\{\ell_{;\hat{\theta}}(\hat{\theta}) - \ell_{;\hat{\theta}}(\hat{\theta}_\psi)\}\,\hat{\jmath}^{-1}\{\ell_{;\hat{\theta}}(\hat{\theta}) - \ell_{;\hat{\theta}}(\hat{\theta}_\psi)\}^{\mathrm{T}} = O(1).$$

Using Taylor's series expansions,

$$\ell_{;\hat{\psi}}(\hat{\theta}) - \ell_{;\hat{\psi}}(\hat{\theta}_\psi) = c(\hat{\psi} - \psi) + O(1),$$

where $c = -\ell''_p(\hat{\psi})$ and $\ell_{;\hat{\lambda}}(\hat{\theta}) - \ell_{;\hat{\lambda}}(\hat{\theta}_\psi) = O(1)$. Using the facts that

$$\ell_{\lambda;\hat{\psi}}(\hat{\theta}_\psi) = \hat{\jmath}_{\lambda\psi} + \mathrm{O}(\sqrt{n}) \quad \text{and} \quad \ell_{\lambda;\hat{\lambda}}(\hat{\theta}_\psi) = \hat{\jmath}_{\lambda\lambda} + \mathrm{O}(\sqrt{n}),$$

it follows that

$$\ell_{\lambda;\hat{\theta}}(\hat{\theta}_\psi)\,\hat{\jmath}^{-1}\{\ell_{;\hat{\theta}}(\hat{\theta}) - \ell_{;\hat{\theta}}(\hat{\theta}_\psi)\}^{\mathrm{T}} = O(1)$$

and, hence, that $D = O(1)$.

The results of this subsection may be used to consider the extent to which $L_C$ and $L_R$ satisfy the first Barlett identity. Ferguson *et al*. (1991) show that $\mathrm{E}[\ell'_M(\psi); \theta] = O(n^{-1})$. According to the analysis above, both $\ell_C(\psi) - \ell_C(\hat{\psi})$ and $\ell_R(\psi) - \ell_R(\hat{\psi})$ are of the form $\ell_M(\psi) - \ell_M(\hat{\psi}) + O(n^{-1})$, where the $O(n^{-1})$ term has a leading term of the form $Q(\hat{\psi} - \psi)^2$ and $Q = O(1)$. Hence, $\ell'_C(\psi) = \ell'_M(\psi) + Q_1(\hat{\psi} - \psi) + O(n^{-1})$ for some $O(1)$ term $Q_1$. It follows that

$$\mathrm{E}[\ell'_C(\psi); \theta] = \mathrm{E}[\ell'_M(\psi); \theta] + O(n^{-1}) = O(n^{-1});$$

similarly, $\mathrm{E}[\ell'_R(\psi); \theta] = O(n^{-1})$. Thus, both $L_C$ and $L_R$ satisfy the first Bartlett identity to a high degree of approximation.

## 4.3. Exact agreement between $L_C$, $L_R$, and $L_M$

In some cases, either $L_C$ or $L_R$ agrees exactly with $L_M$. For instance, if $\hat{\lambda}$ is sufficient for fixed $\psi$, then $L_M$ is valid to $O(n^{-3/2})$ and $L_M = L_C$, which follows from the fact that $\ell_\lambda(\psi, \lambda)$ depends on the data only through $(\hat{\lambda}, a)$ so that $\ell_{\lambda;\hat{\psi}}(\hat{\theta}_\psi) = 0$.

More generally, $L_M = L_C$ whenever there exists a statistic $s$, with dimension equal to that of $\hat{\lambda}$, such that $(s, a)$ is sufficient for fixed $\psi$. To show this, let $\tilde{\ell}(\psi, \lambda) \equiv \tilde{\ell}(\psi, \lambda; s, a)$ denote the log-likelihood function for $\lambda$ with $\psi$ fixed, written as a function of $(s, a)$. Then, for given values of the data,

$$\ell_{\lambda;\hat{\theta}}(\psi, \lambda) = \tilde{\ell}_{\lambda;s}(\psi, \lambda)\frac{\partial s}{\partial \hat{\theta}} \quad \text{and} \quad \ell_{\lambda;\hat{\lambda}}(\psi, \lambda) = \tilde{\ell}_{\lambda;s}(\psi, \lambda)\frac{\partial s}{\partial \hat{\lambda}}.$$

It follows that

$$|\ell_{\lambda;\hat{\theta}}(\hat{\theta}_\psi)\,\hat{\jmath}^{-1}\ell_{\lambda;\hat{\theta}}(\hat{\theta}_\psi)^{\mathrm{T}}|^{1/2} = |\tilde{\ell}_{\lambda;s}(\hat{\theta}_\psi)|\left|\frac{\partial s}{\partial \hat{\theta}}\,\hat{\jmath}^{-1}\left(\frac{\partial s}{\partial \hat{\theta}}\right)^{\mathrm{T}}\right|^{1/2}$$

and that

$$|\ell_{\lambda;\hat{\lambda}}(\hat{\theta}_\psi)| = |\tilde{\ell}_{\lambda;s}(\hat{\theta}_\psi)|\left|\frac{\partial s}{\partial \hat{\lambda}}\right|$$

so that $L_M$ and $L_C$ differ by a factor depending only on the data.

Also, if $\hat{\lambda}_\psi$ does not depend on $\psi$, then $L_M = L_C$. This follows immediately from the fact that in this case $\ell_{\lambda;\hat{\psi}}(\hat{\theta}_\psi) = 0$. Since $L_M$ and $L_C$ are invariant under interest-respecting parametrizations, $L_M = L_C$ holds provided that $\hat{\lambda}_\psi$ is a function of $(\psi, T)$, where $T$ is a statistic with the same dimension as $\lambda$.

If the signed likelihood ratio statistic $R$ depends on the data only through $\hat{\psi}$ then $\ell_{;\hat{\lambda}}(\hat{\theta}) - \ell_{;\hat{\lambda}}(\hat{\theta}_\psi) = 0$. Hence,

$$\left| \begin{matrix} \ell_{;\hat{\theta}}(\hat{\theta}) - \ell_{;\hat{\theta}}(\hat{\theta}_\psi) \\ \ell_{\lambda;\hat{\theta}}(\hat{\theta}_\psi) \end{matrix} \right| = |\ell_{\lambda;\hat{\lambda}}(\hat{\theta}_\psi)| \, |\ell_{;\hat{\psi}}(\hat{\theta}) - \ell_{;\hat{\psi}}(\hat{\theta}_\psi)|$$

and, neglecting terms depending only on the data,

$$|\{\ell_{;\hat{\theta}}(\hat{\theta}) - \ell_{;\theta}(\hat{\theta}_\psi)\} \, \hat{\jmath}^{-1} \{\ell_{;\hat{\theta}}(\hat{\theta}) - \ell_{;\theta}(\hat{\theta}_\psi)\}^{\mathrm{T}}|^{1/2} = |\ell_{;\hat{\psi}}(\hat{\theta}) - \ell_{;\hat{\psi}}(\hat{\theta}_\psi)|$$

so that $L_R = L_M$.

## 5. Examples

***Example 1 Ratio of normal means (continued).*** Recall that the log-likelihood function for the model is given by

$$\ell(\theta) = n\lambda(\psi x + y) - \frac{n}{2}(\psi^2 + 1)\lambda^2,$$

where $x$ and $y$ are normal random variables with means $\psi\lambda$ and $\lambda$, respectively, and each with variance $n^{-1}$. Since $x = \hat{\psi}\hat{\lambda}$ and $y = \hat{\lambda}$, we may write

$$\ell(\theta) = n\lambda(\psi\hat{\psi} + 1)\hat{\lambda} - \frac{n}{2}(\psi^2 + 1)\lambda^2.$$

The profile log-likelihood function is given by

$$\ell_p(\psi) = \frac{n}{2}\frac{(\psi x + y)^2}{\psi^2 + 1}.$$

It is straightforward to show that $\ell_C(\psi) = \ell_R(\psi) = \ell_p(\psi)$. The modified profile log-likelihood function is given by

$$\ell_M(\psi) = \frac{n}{2}\frac{(\psi x + y)^2}{\psi^2 + 1} + \frac{1}{2}\log(\psi^2 + 1) - \log|\psi x + y|.$$

At least in some cases, inferences based on $\ell_C$ are preferable to those based on $\ell_M$. Note that $\hat{\psi}$ maximizes $\ell_C(\psi)$. If $x = 0$ and $y \neq 0$ then $\ell_M$ is maximized at $\psi = \infty$ while $\hat{\psi} = 0$; if $y = 0$ and $x \neq 0$, then $\ell_M$ is maximized at $\psi = 0$ while $\hat{\psi}$ is either $+\infty$ or $-\infty$.

***Example 3 Linear regression model.*** Let $Y_1, \ldots, Y_n$ denote independent random variables of the form $Y_j = \psi + x_j\beta + \sigma\epsilon_j$ where $x_1, \ldots, x_n$ are known covariate vectors of length $p$ satisfying $\sum x_j = 0$, $\psi$ and $\sigma > 0$ are unknown scalar parameters, $\beta$ is a unknown parameter vector, and the $\epsilon_j$ are independent unobservable standard normal random variables. Assume that the matrix $x$ with columns $x_1, \ldots, x_n$ is of full rank.

The log-likelihood function for the model is given by

$$\ell(\psi, \beta, \sigma) = -n \log \sigma - \frac{1}{2\sigma^2}[n\hat{\sigma}^2 + n(\hat{\psi} - \psi)^2 + (\hat{\beta} - \beta)^{\mathrm{T}} x^{\mathrm{T}} x(\hat{\beta} - \beta)]$$

and $\ell_p(\psi) = -n \log \hat{\sigma}_\psi$, where $\hat{\sigma}_\psi^2 = \hat{\sigma}^2 + (\hat{\psi} - \psi)^2$. It is straightforward to show that

$$L_C(\psi) = \hat{\sigma}_\psi^{-(n-p-2)} \left\{ 1 + 2\frac{(\hat{\psi} - \psi)^2}{\hat{\sigma}^2} \right\}^{-1/2}$$

and that

$$L_R(\psi) = \hat{\sigma}_\psi^{-(n-p)} \left\{ 1 + \frac{(\hat{\psi} - \psi)^2}{2\hat{\sigma}^2} \right\}^{1/2}.$$

The profile likelihood function is given by $L_p(\psi) = \hat{\sigma}_\psi^{-n}$; it is well known that for this model $L_M(\psi) = \hat{\sigma}_\psi^{-(n-p-2)}$ (Barndorff-Nielsen and Cox, 1994, Chapter 8).

Thus, there are at least four possible choices of pseudo-likelihood function for $\psi$. Each of $L_C(\psi)$, $L_R(\psi)$ and $L_M(\psi)$ provides a degrees-of-freedom adjustment to $L_p(\psi)$. Note, however, that when $n = p + 2$, the single-degree-of-freedom case, $L_M(\psi)$ is constant, suggesting that $L_C$ and $L_R$ may be preferable to $L_M$ when the degrees of freedom are small.

For instance, each of these pseudo-likelihood functions is maximized at $\psi = \hat{\psi}$, the maximum likelihood estimate. The inverse of the negative second derivative of the log of the pseudo-likelihood function evaluated at $\hat{\psi}$ may be used as an estimate of the variance of $\hat{\psi}$. It is straightforward to show that

$$-\ell_p''(\hat{\psi})^{-1} = \frac{\hat{\sigma}^2}{n}, \qquad -\ell_M''(\hat{\psi})^{-1} = \frac{\hat{\sigma}^2}{n - p - 2},$$

$$-\ell_C''(\hat{\psi})^{-1} = \frac{\hat{\sigma}^2}{n - p}, \qquad -\ell_R''(\hat{\psi})^{-1} = \frac{\hat{\sigma}^2}{n - p - 1/2}.$$

Recall that the usual unbiased estimate of the variance of $\hat{\psi}$ is given by $\hat{\sigma}^2/(n - p - 1)$. Hence, $L_M$, $L_C$ and $L_R$ are each a great improvement over $L_p$ in terms of estimating the variance of $\hat{\psi}$, with the estimate based on $L_R$ being closest to the unbiased estimate.

***Example 4 Normal distributions with common mean.*** Let $Y_{jk}$, $k = 1, \ldots, n_j$, $j = 1, \ldots, m$, denote independent normal random variables such that $Y_{jk}$ has mean $\mu$ and standard deviation $\sigma_j$. Take the common mean $\mu$ as the parameter of interest and take $(\sigma_1, \ldots, \sigma_m)$ as the nuisance parameter. Let

$$Y_j = \frac{1}{n_j}\sum Y_{jk} \quad \text{and} \quad S_j = \sum (Y_{jk} - Y_j)^2;$$

by sufficiency, the analysis may be based on $(Y_1, S_1), \ldots, (Y_m, S_m)$. The log-likelihood function is given by

$$\ell(\theta) = -\sum n_j \log \sigma_j - \frac{1}{2}\sum \frac{n_j}{\sigma_j^2}(y_j - \mu)^2 - \frac{1}{2}\sum \frac{s_j}{\sigma_j^2}.$$

For this model, exact calculation of the sample space derivatives needed to determine $L_C$,

$L_R$ and $L_M$ is not possible. Hence, we use approximations to these functions based on the approach of of Skovgaard (1996) and Severini (1998) in which sample space derivatives are approximated by covariances of the log-likelihood function and its derivatives. These approximations have error $O(n^{-1})$ for $\psi = \hat{\psi} + O(n^{-1/2})$.

Let $\bar{\ell}_M$ denote the approximation to $\ell_M$ and so on. It is straightforward to show that

$$\bar{\ell}_M(\mu) = -\sum (n_j - 2)\log \hat{\sigma}_{j\mu}, \qquad \bar{\ell}_C(\mu) = -\sum (n_j - 2)\log \hat{\sigma}_{j\mu} - \frac{1}{2}\log|F|,$$

where $F$ is an $m \times m$ matrix with $(i, j)$th element given by

$$F_{ij} = \begin{cases} \dfrac{4n_i n_j}{c}[(\hat{\mu} - \mu)^2 + \{(y_i - \hat{\mu}) + (y_j - \hat{\mu})\}(\hat{\mu} - \mu)] & \text{if } i \neq j, \\ \dfrac{4n_i^2}{c}\{(\hat{\mu} - \mu)^2 + 2(y_i - \hat{\mu})(\hat{\mu} - \mu)\} + 2n_i\hat{\sigma}_i^4 & \text{if } i = j, \end{cases}$$

$c = \sum n_i/\hat{\sigma}_i^2$ and

$$\bar{\ell}_R(\mu) = -\sum (n_j - 2)\log \hat{\sigma}_{j\mu} + \frac{1}{2}\log G,$$

where

$$G = 2\sum \frac{n_i}{\hat{\sigma}_{i\mu}^4}\left(y_i - \frac{\mu + \hat{\mu}}{2}\right)^2 + \frac{d^2}{c} - 4\frac{d}{c}\sum \frac{n_i}{\hat{\sigma}_{i\mu}^2\hat{\sigma}_i^2}(y_i - \hat{\mu})\left(y_i - \frac{\mu + \hat{\mu}}{2}\right)$$

and $d = \sum n_i/\hat{\sigma}_{i\mu}^2$. Also, note that $\ell_p(\mu) = -\sum n_j \log \hat{\sigma}_{j\mu}$.

For small sample sizes, $\bar{\ell}_C$ and $\bar{\ell}_R$ may be very different than $\bar{\ell}_M$. In particular, $\bar{\ell}_M$ ignores any observations $(Y_j, S_j)$ with $n_j = 2$. If there are several observations of this type or if those observations have a relatively small standard deviation, this can be a serious waste of information.

Likelihood inference for this model has been considered by Kalbfleisch and Sprott (1970), Chamberlin and Sprott (1989) and Barndorff-Nielsen (1983), among others. Barndorff-Nielsen (1983) obtains the expression for $L_M$ obtained here by ignoring the factor $\partial\hat{\lambda}/\partial\hat{\lambda}_\psi$.

## 6. Discrete data

The results in the preceding section were based on the assumption that the underlying data have a continuous distribution. The main difference in the discrete case is that, if the relevant densities are with respect to counting measure, it is not obvious that the differential terms used in deriving $L_C$ and $L_R$ are still appropriate.

Suppose that the underlying data have a lattice distribution. Then Severini (2000b) shows that the density of $\hat{\theta}$ may still be approximated using $p^*$; however, the density $p^*$ is with respect to a measure that depends on the structure of the sample space of $\hat{\theta}$, rather than with respect to counting measure. Let $\mu^*(\cdot; \hat{\Theta})$ denote this underlying measure, where $\hat{\Theta}$

denote the sample space of $\hat{\theta}$; see Severini (2000b) for further discussion of $\mu^*$. Then we may approximate the density of $\hat{\theta}$ by

$$p^*(\hat{\theta}|a;\, \theta) = c|\,\hat{\jmath}|^{1/2} \exp\{\ell(\theta) - \ell(\hat{\theta})\} \mathrm{d}\mu^*(\hat{\theta};\, \hat{\Theta}).$$

Now consider the density function of $X = g(\hat{\theta})$ where $g$ is a one-to-one continuously differentiable function. It may be shown that an approximation to the density of $X$ is given by

$$p^*(\hat{\theta}|a;\, \theta)\left|\frac{\partial\hat{\theta}}{\partial x}\right| \mathrm{d}\mu^*(x;\, \mathcal{X}), \qquad \hat{\theta} = \hat{\theta}(x);$$

here $\mathcal{X}$ denotes the sample space of $X$. That is, the density of $X$ with respect to $\mu^*(\cdot;\, \mathcal{X})$ may be obtained from $p^*(\hat{\theta}|a;\, \theta)$ using the change-of-variable formula commonly used for densities with respect to Lebesgue measure. It follows that the methods used to derive $L_C$ and $L_R$ hold when the underlying data have a lattice distribution as well.

## Acknowledgement

This work was supported by the National Science Foundation.

## References

Barndorff-Nielsen, O.E. (1980) Conditionality resolutions. *Biometrika*, **67**, 293–310.

Barndorff-Nielsen, O.E. (1983) On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, **70**, 343–365.

Barndorff-Nielsen, O.E. (1994) Adjusted versions of profile likelihood and directed likelihood, and extended likelihood. *J. Roy. Statist. Soc. Ser. B*, **56**, 125–140.

Barndorff-Nielsen, O.E. and Cox, D.R. (1994) *Inference and Asymptotics.* London: Chapman & Hall.

Basu, D. (1955) On statistics independent of a complete sufficient statistic. *Sankhyā*, **15**, 377–380.

Basu, D. (1958) On statistics independent of a sufficient statistic. *Sankhyā*, **20**, 223–226.

Chamberlin, S.R. and Sprott, D.A. (1989) Linear systems of pivotals and associated pivotal likelihoods with applications. *Biometrika*, **76**, 685–691.

Cox, D.R. and Reid, N. (1987) Parameter orthogonality and approximate conditional inference. *J. Roy. Statist. Soc. Ser. B*, **49**, 1–39.

Ferguson, H., Reid, N. and Cox, D.R. (1991) Estimating functions from modified profile likelihood. In V.P. Godambe (ed.), *Estimating Functions*. Oxford: Oxford University Press.

Fraser, D.A.S. (1967) Data transformations and the linear model. *Ann. Math. Statist.*, **38**, 1456–1465.

Fraser, D.A.S. (1968) *The Structure of Inference.* New York: Wiley.

Fraser, D.A.S. (1972) The determination of likelihood and the transformed regression model. *Ann. Math. Statist.*, **43**, 898–916.

Fraser, D.A.S. (1979) *Inference and Linear Models.* Toronto: McGraw-Hill.

Fraser, D.A.S. and Reid, N. (1989) Adjustments to profile likelihood. *Biometrika*, **76**, 477–488.

Fraser, D.A.S. and Reid, N. (1995) Bayes posteriors for scalar interest parameters. In J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith (eds), *Bayesian Statistics 5*. Oxford: Oxford University Press.

Hoffmann-Jørgensen, J. (1994) *Probability with a View towards Statistics, Volume II.* New York: Chapman & Hall.

Kalbfleisch, J.D. and Sprott, D.A. (1970) Application of likelihood methods to models involving large numbers of parameters (with discussion). *J. Roy. Statist. Soc. Ser. B*, **32**, 175–208.

Kalbfleisch, J.D. and Sprott, D.A. (1973) Marginal and conditional likelihoods. *Sankhyā Ser. A*, **35**, 311–328.

McCullagh, P. and Nelder, J. (1989) *Generalized Linear Models, 2nd edition*. London: Chapman & Hall.

McCullagh, P. and Tibshirani, R. (1990) A simple method for the adjustment of profile likelihoods. *J. Roy. Statist. Soc. Ser. B*, **52**, 325–344.

Severini, T.A. (1994) On the approximate elimination of nuisance parameters by conditioning. *Biometrika*, **81**, 649–661.

Severini, T.A. (1998) An approximation to the modified profile likelihood function. *Biometrika*, **85**, 403–411.

Severini, T.A. (2000a) *Likelihood Methods in Statistics.* Oxford: Oxford University Press.

Severini, T.A. (2000b) The likelihood ratio approximation to the conditional distribution of the maximum likelihood estimate in the lattice case. *Biometrika*, **87**, 939–945.

Skovgaard, I.M. (1996) An explicit large-deviation approximation to one-parameter tests. *Bernoulli*, **2**, 145–165.

Tjur, T. (1980) *Probability Based on Radon Measures.* New York: Wiley.

Received June 2002 and revised January 2004