Parameter Orthogonality and Approximate Conditional Inference

D. R. Cox; N. Reid

Journal of the Royal Statistical Society. Series B (Methodological), Vol. 49, No. 1 (1987), 1-39.

Stable URL:
http://links.jstor.org/sici?sici=0035-9246%281987%2949%3A1%3C1%3APOAACI%3E2.0.CO%3B2-E

Journal of the Royal Statistical Society. Series B (Methodological) is currently published by Royal Statistical Society.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/rss.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.

http://www.jstor.org/
Sat Oct 22 21:06:53 2005

J. R. Statist. Soc. B (1987)
49, *No.* 1, *pp.* 1–39

Parameter Orthogonality and Approximate Conditional Inference

D. R. COX[†] and N. REID

Imperial College, London | *University of British Columbia, Vancouver*

[*Read before the* Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, 8th October, 1986,* Professor A. F. M. Smith *in the Chair*]

SUMMARY

We consider inference for a scalar parameter ψ in the presence of one or more nuisance parameters. The nuisance parameters are required to be orthogonal to the parameter of interest, and the construction and interpretation of orthogonalized parameters is discussed in some detail. For purposes of inference we propose a likelihood ratio statistic constructed from the conditional distribution of the observations, given maximum likelihood estimates for the nuisance parameters. We consider to what extent this is preferable to the profile likelihood ratio statistic in which the likelihood function is maximized over the nuisance parameters. There are close connections to the modified profile likelihood of Barndorff-Nielsen (1983). The normal transformation model of Box and Cox (1964) is discussed as an illustration.

Keywords: ASYMPTOTIC THEORY; CONDITIONAL INFERENCE; LIKELIHOOD RATIO TEST; NORMAL TRANSFORMATION MODEL; NUISANCE PARAMETERS; ORTHOGONAL PARAMETERS

1. INTRODUCTION

The primary objective of this paper is to explore the connection between orthogonality of parameters and the asymptotic theory of conditional inference. Orthogonality is defined with respect to the expected Fisher information matrix as described in Section 2. In general it is not possible to have total parameter orthogonality at all parameter values but it is possible to obtain orthogonality of a scalar parameter of interest ψ to a set of nuisance parameters. The concept of orthogonal parameters seems to have fairly broad implications and is discussed in some detail in Section 2 and illustrated with several examples in Section 3.

A widely used procedure for inference about a parameter in the presence of nuisance parameters is to replace the nuisance parameters in the likelihood function by their maximum likelihood estimates and examine the resulting profile likelihood as a function of the parameter of interest. This procedure is known to give inconsistent or inefficient estimates for problems with large numbers of nuisance parameters, which suggests that it may not be close to optimal for a small number of nuisance parameters, even though the likelihood ratio statistic with no nuisance parameters is in some sense optimal. We consider an approach to inference based on the conditional likelihood given maximum likelihood estimates of the orthogonalized parameters. To the extent that the maximum likelihood estimates of the nuisance parameters are complete sufficient statistics for the nuisance parameters, this conditional likelihood procedure generalizes the usual procedure for obtaining similar tests, described for example in Cox and Hinkley (1974, p. 134). There are close connections to the modified profile likelihood of Barndorff-Nielsen (1983, 1985b).

The conditional profile likelihood function is discussed and illustrated in Section 4, and a possible justification for preferring it to the usual profile likelihood function is presented in Section 4.3. Inference for the normal transformation model is discussed separately in Section 5. In Section 6 some further points and open questions are discussed.

[†] *Address for correspondence*: Department of Mathematics, Imperial College, London SW7 2BZ, UK.

© 1987 Royal Statistical Society

0035–9246/87/49001

2. ORTHOGONAL PARAMETERS

2.1. *Introduction*

We deal throughout with parametric problems for which the vector of observations is represented by an $n \times 1$ vector Y of random variables having density $f_Y(y; \theta)$ depending on a $1 \times p$ vector θ of unknown parameters. We write $l(\theta)$ for the log-likelihood; depending on the context this will be either log $f_Y(y; \theta)$ for given observations y, or the random variable log $f_Y(Y; \theta)$. Occasionally we write $l_Y(\theta)$ to emphasize that the log-likelihood is derived from the density of Y. Our arguments will be informal without explicit attention to regularity conditions, these being essentially those required for the expansions needed for maximum likelihood theory in regular estimation problems.

If θ is partitioned into two vectors θ_1 and θ_2 of length p_1 and p_2 respectively, $p_1 + p_2 = p$, we define θ_1 to be orthogonal to θ_2 if the elements of the information matrix satisfy

$$i_{\theta_s \theta_t} = \frac{1}{n} E\left(\frac{\partial l}{\partial \theta_s} \frac{\partial l}{\partial \theta_t}; \theta \right) = \frac{1}{n} E\left(- \frac{\partial^2 l}{\partial \theta_s \partial \theta_t}; \theta \right) = 0 \qquad (1)$$

for $s = 1, \ldots, p_1$, $t = p_1 + 1, \ldots, p_1 + p_2$; this is to hold for all θ in the parameter space, and is sometimes called global orthogonality. Note that i refers to information per observation, which will be assumed to be $O(1)$ as $n \to \infty$. If (1) holds at only one parameter value θ^0, then the vectors θ_1 and θ_2 are said to be locally orthogonal, at θ^0. The most direct statistical interpretation of (1) is that the relevant components of the score statistic are uncorrelated.

The definition of orthogonality can be extended to more than two sets of parameters, and in particular θ is totally orthogonal if the information matrix is diagonal. While orthogonality can always be achieved locally, global orthogonality is possible only in special cases (Jeffreys, 1961, p. 208; Huzurbazar, 1950; Mitchell, 1962; Amari, 1985).

2.2 *Consequences of Orthogonality*

There are a number of statistical consequences of orthogonality which we now outline. For simplicity, suppose $\theta = (\psi, \lambda)$ has just two components. Then orthogonality of ψ and λ implies that

(i) the maximum likelihood estimates $\hat{\psi}$ and $\hat{\lambda}$ are asymptotically independent;

(ii) the asymptotic standard error for estimating ψ is the same whether λ is treated as known or unknown;

(iii) there may be simplifications in the numerical determination of $(\hat{\psi}, \hat{\lambda})$; see Ross (1970) in the context of nonlinear regression.

A further property related to (iii) and of particular relevance for the present paper is

(iv) $\hat{\psi}_\lambda = \hat{\psi}(\lambda)$, the maximum likelihood estimate of ψ when λ is given, varies only slowly with λ.

To study (iv), we write the log-likelihood function near the maximum $(\hat{\psi}, \hat{\lambda})$ as

$$l(\hat{\psi}, \hat{\lambda}) + \tfrac{1}{2}\{-n\hat{j}_{\psi\psi}(\psi - \hat{\psi})^2 - 2n\hat{j}_{\psi\lambda}(\psi - \hat{\psi})(\lambda - \hat{\lambda}) - n\hat{j}_{\lambda\lambda}(\lambda - \hat{\lambda})^2\} + O_p(\| \theta - \hat{\theta} \|^3), \qquad (2)$$

where, for example, $n\hat{j}_{\psi\psi} = [-\partial^2 l(\psi, \lambda)/\partial\psi^2]_{\theta = \hat{\theta}}$. Write $\hat{j}_{\psi\psi} = i_{\psi\psi} + Z_{\psi\psi}/\sqrt{n}$, etc., where $Z_{\psi\psi}, \ldots$ are random variables of zero mean and $O_{\rm p}(1)$ as $n \to \infty$. The dependence of i, Z on θ is suppressed. We rewrite (2) in terms of i and Z, differentiate (2) with respect to ψ so that $\hat{\psi}_\lambda$ satisfies

$$n i_{\psi\psi}(\hat{\psi}_\lambda - \hat{\psi}) + \sqrt{n} Z_{\psi\psi}(\hat{\psi}_\lambda - \hat{\psi}) + \tfrac{1}{2}(\hat{\psi}_\lambda - \hat{\psi})^2 \sqrt{n} \frac{\partial Z_{\psi\psi}}{\partial \psi} + \tfrac{1}{2}(\hat{\psi}_\lambda - \hat{\psi})^2 n \frac{\partial i_{\psi\psi}}{\partial \psi} + \ldots = 0, \qquad (3)$$

where derivatives are evaluated at (ψ, λ). Provided that random variables such as $\partial Z_{\psi\psi}/\partial_\psi$ are $O_p(1)$ and quantities such as $\partial i_{\psi\psi}/\partial\psi$ are $O(1)$, and noting that $\hat{\psi}_\lambda - \hat{\psi} = O_p(1/\sqrt{n})$, then if and only if $i_{\psi\lambda} = 0$, the first term of (3) is $O_p(\sqrt{n})$, whereas the remaining terms are $O_p(1)$ as λ varies by an amount that is $O(1/\sqrt{n})$. It follows that the first term is in fact $O_p(1)$, requiring that $\hat{\psi}_\lambda - \hat{\psi}$ is $O_p(1/n)$. A similar proof holds if the parameters are not scalars. The argument is of course symmetric in (ψ, λ), and we will use also the result $\hat{\lambda}_\psi - \hat{\lambda} = O_p(1/n)$ in later sections.

It is easy to see that if $\hat{\psi}_\lambda = \hat{\psi}$ for all λ, then λ and ψ are orthogonal parameters. Examples of families for which this holds are discussed in Barndorff-Nielsen (1978), an important class being regular exponential models with ψ as part of the canonical parameter and λ as the complementary part of the expectation parameter; see Example 3.2. It would, of course, be possible to have $\hat{\psi}_\lambda$ functionally independent of λ and at the same time for the distribution and, in particular the standard error, of $\hat{\psi}_\lambda$ to depend strongly on λ.

Property (iv) is discussed also by Sweeting (1984b) in the context of location-scale models. A numerical illustration is provided in Section 3.5.

Note that from a pair (ψ, λ) of orthogonal parameters other pairs could be obtained by suitable transformation. However, in this paper we shall regard ψ as a preassigned parameter of particular relevance.

2.3. *Construction of Orthogonal Parameters*

As noted above, it is not in general possible to find a totally orthogonal parametrization. We now discuss the special case in which a scalar parameter ψ is orthogonal to the remaining parameters $\lambda_1, \ldots, \lambda_q$. Typically ψ will be the parameter of interest and $\lambda_1, \ldots, \lambda_q$ will be nuisance parameters, although it is possible that ψ is the nuisance parameter and one or all components of λ are the parameters of interest; see Example 3.5 below. In the notation of equation (1), $\theta_1 = \psi$, $\theta_2 = (\lambda_1, \ldots, \lambda_q)$.

The following argument generalizes Huzurbazar (1950); see also Jeffreys (1961, p. 208) and Amari (85, p. 254). Suppose that initially the likelihood is specified in terms of $(\psi, \phi_1, \ldots, \phi_q)$. We then write $\phi_1 = \phi_1(\psi, \lambda)$, $\phi_2 = \phi_2(\psi, \lambda), \ldots, \phi_q = \phi_q(\psi, \lambda)$, where $\lambda = (\lambda_1, \ldots, \lambda_q)$, and

$$l(\psi, \lambda) = l^*\{\psi, \phi_1(\psi, \lambda), \ldots, \phi_q(\psi, \lambda)\},$$

regarding l^* as a function of $(\psi, \phi_1, \ldots, \phi_q)$. Then

$$\frac{\partial l}{\partial \psi} = \frac{\partial l^*}{\partial \psi} + \sum \frac{\partial l^*}{\partial \phi_r} \frac{\partial \phi_r}{\partial \psi},$$

$$\frac{\partial^2 l}{\partial \psi \partial \lambda_t} = \sum \frac{\partial^2 l^*}{\partial \psi \partial \phi_s} \frac{\partial \phi_s}{\partial \lambda_t} + \sum \frac{\partial^2 l^*}{\partial \phi_r \partial \phi_s} \frac{\partial \phi_s}{\partial \lambda_t} \frac{\partial \phi_r}{\partial \psi} + \sum \frac{\partial l^*}{\partial \phi_r} \frac{\partial^2 \phi_r}{\partial \psi \partial \lambda_t}.$$

On taking expectations the last term in the second derivative vanishes, so that the orthogonality equations are

$$\sum \frac{\partial \phi_s}{\partial \lambda_t} \left(i^*_{\psi\phi_s} + \sum i^*_{\phi_r\phi_s} \frac{\partial \phi_r}{\partial \psi} \right) = 0, \quad t = 1, \ldots, q,$$

where the i^* are the information measures calculated in the (ψ, ϕ) parametrization. We require that the transformation from (ψ, ϕ) to (ψ, λ) have nonzero Jacobian; hence

$$\sum i^*_{\phi_r\phi_s} \frac{\partial \phi_r}{\partial \psi} = -i^*_{\psi\phi_s}, \quad s = 1, \ldots, q. \tag{4}$$

These partial differential equations determine the dependence of ϕ on ψ, but there is considerable arbitrariness in the dependence of ϕ on λ; see the examples. It is often convenient to take ϕ_1 to depend only on (ψ, λ_1), ϕ_2 to depend on $(\psi, \lambda_1, \lambda_2)$, etc., and to aim to give the λ meaningful interpretation in the context of the particular problem.

From (4) it is clear why in general we cannot obtain global orthogonality when ψ is not a scalar. If $\psi = (\psi_1, \psi_2)$, we can use (4) independently to calculate $\partial \phi_s / \partial \psi_1$ and $\partial \phi_s / \partial \psi_2$, and there is no guarantee that in general the compatability condition $\partial^2\phi_s/\partial\psi_1\partial\psi_2 = \partial^2\phi_s/\partial\psi_2\partial\psi_1$ is satisfied.

3. EXAMPLES

3.1. *Exponential Distribution*

Let Y_1 and Y_2 be exponential random variables with means ϕ and $\psi\phi$ respectively; the parameter of interest is the ratio of the means. The differential equation corresponding to (4) is

$$\frac{2}{\phi^2}\frac{\partial\phi}{\partial\psi} = -\frac{1}{\psi\phi},$$

with solution $\phi\psi^{1/2} = a(\lambda)$, where $a(\lambda)$ is an arbitrary function of λ. A convenient choice is $a(\lambda) = \lambda$; in the new parametrization Y_1 and Y_2 have means $\lambda\psi^{-1/2}$ and $\lambda\psi^{1/2}$, respectively. Note that for n independent replications of (Y_1, Y_2),

$$\hat{\lambda}_\psi = \tfrac{1}{2}\psi^{-1/2}(\psi\bar{y}_1 + \bar{y}_2), \quad \hat{\lambda} = (\bar{y}_1\bar{y}_2)^{1/2}, \quad \hat{\lambda}_\psi - \hat{\lambda} = O_p(1/n),$$

and $\hat{\lambda}_\psi$ has a distribution depending only on λ.

The extension to exponential regression has $Y_1, \ldots, Y_n$ independent exponential random variables with $EY_i = \lambda \exp(-\psi z_i)$, where z_i are given constants. Requiring $\Sigma z_i = 0$ ensures that λ and ψ are orthogonal. If we add on another explanatory variable to give $EY_i = \lambda \exp(-\psi z_i - \beta x_i)$ and also require $\Sigma x_i = 0$, then λ and ψ are still orthogonal, as are λ and β. Assuming ψ is still the parameter of interest, we need the orthogonal expression of the nuisance parameter β with respect to ψ. This is obtained by subtracting from z_i its regression on x_i giving, in the new parametrization,

$$EY_i = \lambda \exp[-\psi\{z_i - x_i(S_{xz}/S_{xx})\} - \eta x_i],$$

where $S_{xz} = \Sigma x_i z_i$ and $S_{xx} = \Sigma x_i^2$.

A different version of the two-sample problem concerns inference about the difference between two exponential means. Let Y_1, Y_2 be independent exponential random variables with means ϕ and $(\phi + \psi)$ respectively. The differential equation (4) gives

$$\left\{\frac{1}{(\phi+\psi)^2} + \frac{1}{\phi^2}\right\}\frac{\partial\phi}{\partial\psi} = -\frac{1}{(\phi+\psi)^2};$$

this can be solved by separation of variables, leading to $a(\lambda) = \phi(\psi + \phi)/(\psi + 2\phi)$, where again $a(\lambda)$ is an arbitrary function of λ. In most of our examples we choose $a(\lambda) = \lambda$; in this example $a(\lambda) = e^\lambda$ might be more suitable.

3.2. *Regular Exponential Families*

Write $f(y;\ \theta) = \exp\{\theta_1 t_1 + \theta_2 t_2 - c(\theta) - d(y)\}$, where (θ_1, θ_2) are components of the canonical parameter and $\{t_1(y), t_2(y)\}$ are the corresponding components of the sufficient statistic. Let $\eta = (\eta_1, \eta_2) = (Et_1, Et_2)$ be the expectation parameter. It is easy to verify directly, and is implicit in Amari (1982) and Barndorff-Nielsen (1983), that θ_1 is orthogonal to η_2 and θ_2 is orthogonal to η_1.

As a simple example the normal distribution with mean μ and variance τ has canonical parameter $(\mu/\tau,\ -1/(2\tau))$ and expectation parameter $(\mu,\ \mu^2 + \tau)$. Thus μ is orthogonal to $-1/(2\tau)$, hence to τ, and μ/τ is orthogonal to $\mu^2 + \tau$. The normal distribution will be studied separately as Example 3.3.

Another example is the gamma distribution with shape parameter ψ and scale parameter ϕ

$$f(y;\ \psi,\ \phi) = \phi^{-\psi}y^{\psi-1}\exp(-y/\phi)/\Gamma(\psi).$$

The canonical parameter is $(-1/\phi,\ \psi)$ corresponding to $(y,\ \log y)$ and we have immediately that $EY = \psi\phi$ is orthogonal to ψ. The new parametrization is

$$f(y;\, \psi, \lambda) = (\lambda\psi^{-1})^{-\psi} y^{\psi-1} \exp\{-(\psi y/\lambda)\}/\Gamma(\psi).$$

In this example $\hat{\lambda}_\psi = \bar{y}$ does not depend on ψ, although its distribution does.

These results are discussed in Barndorff-Nielsen (1978, p. 184), where he shows also that the dispersion parameter of a generalized linear model is orthogonal to the expectation parameter.

3.3. *Normal Distribution*

As noted above, the (μ, τ) parametrization of the normal distribution is orthogonal. Note that $\hat{\mu}_\tau = \bar{y}$ does not depend on the nuisance parameter τ, whereas $\hat{\tau}_\mu = n^{-1}\{\Sigma(y_i - \bar{y})^2 + n(\bar{y} - \mu)^2\}$ differs from $\hat{\tau}$ by $O_p(n^{-1})$. In the regression setting, the variance τ is orthogonal to the regression coefficients β; if the components of β are to be orthogonal to each other the design matrix must be orthogonalized, as in the exponential regression example.

More generally, when Y has a multivariate normal distribution with mean vector $X\beta$ and covariance matrix $V(\psi)$, then β and ψ are orthogonal, so long as they are functionally unrelated. This generalisation includes, in particular, components of variance models (Patterson and Thompson, 1971).

As an example of nonorthogonal parameters take τ and $\xi = (\mu - a)/\tau^{1/2}$, the latter determining the probability of an observation falling below the fixed tolerance level a. Then $\hat{\xi}_\tau = (\bar{y} - a)/\tau^{1/2}$ differs from $\hat{\xi} = \hat{\xi}_{\hat{\tau}}$ by $O_p(1/\sqrt{n})$. The parameter that is orthogonal to ξ is an arbitrary function of $(\xi^2 + 2)\tau$.

3.4. *Weibull Distribution*

We take the index of the Weibull distribution as the parameter ψ, writing

$$f(y;\, \psi, \phi) = \left(\frac{\psi}{\phi}\right)\left(\frac{y}{\phi}\right)^{\psi-1} \exp\left\{-\left(\frac{y}{\phi}\right)^{\psi}\right\}.$$

Then $i_{\phi\phi} = (\psi/\phi)^2$, $i_{\phi\psi} = -\Gamma'(2)/\psi$, and the orthogonal nuisance parameter $\lambda = \phi \exp(\Gamma'(2)/\psi)$. The survivor function in the new parametrization is

$$1 - F(y) = \exp\{-(y/\lambda)^{\psi} \exp(\Gamma'(2))\}.$$

The value of $\Gamma'(2)$ is $1 - \gamma$, where $\gamma = .577215\ldots$ is Euler's constant, so $1 - F(\lambda) \simeq 0.22$.

In practice it may be of more interest to estimate the rate parameter, treating ψ as a nuisance parameter. A statistical interpretation of the above parametrization is that maximum likelihood estimation of the 80th percentile of the distribution depends very little on ψ; in particular $\hat{\lambda}$ will be nearly the same whether we assume an exponential distribution ($\psi = 1$), or estimate both parameters by maximum likelihood, provided that the true value is not very different from 1. Thus the maximum likelihood estimate of this percentile is in a rather special sense robust. This interpretation of orthogonality is discussed in more detail in the context of the normal transformation model.

3.5. *Normal Transformation Model*

We assume that for some non-zero ψ, Y_i^ψ has a normal distribution with mean μ, variance τ. (The case $\psi = 0$ will be taken to correspond to $\log Y_i$.) The usual formulation of this model involves $\psi^{-1}(Y_i^\psi - 1)$ (Box and Cox, 1964) but the argument for that family is essentially the same. Although in practice interest will usually focus on the mean, and possibly the variance, for the present we look for a reparametrization of μ and τ to make them orthogonal to the transformation parameter ψ. In this model it is necessary that Y be non-negative; this could be achieved by truncation but we will assume that the variance is sufficiently small relative to the mean that nonpositive observations have negligible probability.

To extend the argument more easily to the regression setting we change the notation for μ and τ to ϕ_1 and ϕ_0, respectively. The ϕ part of the information matrix, $i_{\phi\phi}$, is orthogonal, as in

Example 3.4, but $i_{\phi_1\psi}$ and $i_{\phi_0\psi}$ can only be evaluated approximately, using

$$E(Y^{\psi} \log Y) = \phi_1 \log \phi_1/\psi + O(\phi_0),$$

$$E\{Y^{\psi} \log Y(Y^{\psi} - \phi_1)\} = \phi_0(1 + \log \phi_1)/\psi + O(\phi_0^2).$$

The pair of differential equations to be solved is, approximately,

$$\frac{1}{\phi_0}\frac{\partial\phi_1}{\partial\psi} = \frac{\phi_1 \log \phi_1}{\phi_0 \psi},$$

$$\frac{1}{2\phi_0^2}\frac{\partial\phi_0}{\partial\psi} = \frac{\phi_0(1 + \log \phi_1)}{\phi_0^2 \psi}.$$

From the first equation $\phi_1 = \exp\{a(\lambda_1, \lambda_0)\psi\}$, and from the second equation $\phi_0^{1/2} = \psi\lambda_1^{\psi}b(\lambda_1, \lambda_0)$, where a and b are arbitrary functions of (λ_1, λ_0). We choose $a(\lambda_1, \lambda_0) = \log \lambda_1$ and $b(\lambda_1, \lambda_0) = \lambda_0^{1/2}/\lambda_1$, so the model is represented in the form

$$Y_i^{\psi} \sim N(\lambda_1^{\psi}, \lambda_1^{2\psi-2}\psi^2\lambda_0).$$

Note that if Y_i has mean λ_1 and variance λ_0, then Y_i^{ψ} has approximately the normal distribution just given; this was the motivation for the particular choice of a, b above.

We can use this for a simple numerical illustration of property (iv) of Section 2.2, the stability of maximum likelihood estimates of one parameter as another parameter varies. We have taken the set of 15 systolic blood pressures recorded by Cox and Snell (1981, Table E.1, col. 1). The mean is 176.9 mm Hg and the standard deviation is 20.56 mm Hg. As ψ varies from 2 to -2 there is a large change in the estimated means and variances; in fact the means change by a factor of 10^9. On the other hand the estimated λ_1 and λ_0 vary respectively from 178.0 to 173.6 and from 424 to 411, illustrating the considerable stability of the estimates of λ_1 and λ_0 with respect to changes in ψ.

The extension of this model to the regression setting proceeds as follows. Assume $Y_i^{\psi} \sim N(\Sigma x_{ir}\phi_r, \phi_0)$; then

$$l(\phi, \psi; y) = -\frac{n}{2}\log \phi_0 - \frac{1}{2\phi_0}\sum (y_i^{\psi} - \sum x_{ir}\phi_r)^2 + n \log \psi + (\psi - 1)\sum \log y_i,$$

the last two terms being derived from the Jacobian of the transformation from y_i^{ψ} to y_i. The computations are simplified if we assume that the matrix of explanatory variables has been standardized; then $i_{\phi\phi} = \text{diag}(1/\phi_0, \ldots, 1/\phi_0, \frac{1}{2}/\phi_0^2)$, where the variance component $i_{\phi_0\phi_0}$ is last. Approximating $E(Y^{\psi} \log Y)$ as before, we have

$$i_{\phi_0\psi} \simeq -\textstyle\sum_i \{\log(\sum x_{is}\phi_s) + 1\}/(\psi\phi_0),$$

$$i_{\phi_r\psi} \simeq -\textstyle\sum_i (\sum x_{is}\phi_s) \log(\sum x_{is}\phi_s)x_{ir}/(\psi\phi_0).$$

A further simplification is to take $x_{i1} = 1$, $\Sigma_r x_{ir} = 0$, so that ϕ_1 is an overall mean. Assuming other effects to be relatively small, we have

$$\log(\textstyle\sum x_{is}\phi_s) = \log \phi_1 + \displaystyle\sum_{i=2}^{q} x_{is}\phi_s/\phi_1 + O(\phi_1^{-2}),$$

giving approximately

$$i_{\phi_0\psi} = (1 + \log \phi_1)/(\psi\phi_0),$$

$$i_{\phi_1\psi} = \phi_1 \log \phi_1/(\psi\phi_0),$$

$$i_{\phi_r\psi} = \phi_r(1 + \log \phi_1)/(\psi\phi_0), \quad r = 2, \ldots, q.$$

One solution of the set of equations (4) gives

$$\phi_1 = \lambda_1^{\psi}, \quad \phi_0 = \lambda_1^{2\psi-2}\psi^2\lambda_0,$$
$$\phi_r = \psi\lambda_1^{\psi}\lambda_r, \quad r = 2, \ldots, q.$$

The orthogonal expression of the model is thus

$$Y_i^{\psi} \sim N\left(\lambda_1^{\psi} + \psi\lambda_1^{\psi}\sum_{s=2}^{q} x_{is}\lambda_s, \lambda_1^{2\psi-2}\psi^2\lambda_0\right). \tag{5}$$

Note that λ_1 and $\lambda_0^{1/2}$ have the dimensions of Y and $\lambda_2, \ldots, \lambda_q$ are dimensionless. Analysis of this model will be discussed in Section 5.

To discuss the statistical interpretation of these results we take a slightly broader setting. Suppose we have a model $f(y; \phi)$ involving an unknown parameter ϕ of interest and the model is enriched by a nuisance parameter ψ in order to produce a more realistic model. One possibility is that there is a second model $g(y; \phi)$ and that ψ indexes the exponential mixtures with density proportional to

$$\{f(y; \phi)\}^{\psi}\{g(y; \phi)\}^{1-\psi}.$$

We concentrate on estimating ϕ, treating ψ as essentially totally unknown. For this problem to have a clear meaning, ϕ should be defined so as to have an interpretation in some sense independent of ψ.

In some problems the components of ϕ may have a descriptive interpretation that is unaffected by the value of ψ; two examples are the components of the mean response vector and regression coefficients on some fixed scale. Then direct comparison of estimates of ϕ from different analyses is possible, even if different values of ψ are used. In general such an interpretation is not available, and then a basis for comparison can be provided by expressing $\phi = \phi(\psi, \lambda)$, choosing λ to be orthogonal to ψ. Estimates of ψ for different values of ψ can be compared *via* conversion to the corresponding estimate of λ. In particular, we might consider

(i) the overall maximum likelihood estimates $(\hat{\psi}, \hat{\phi})$;
(ii) the maximum likelihood estimate of ϕ at $\psi = \psi^0$, say $\hat{\phi}_0$;
(iii) the maximum likelihood estimate of ϕ at some other, possibly data dependent, value $\tilde{\psi}$, say $\tilde{\phi}$.

By the parameter orthogonality, we have that $\hat{\phi}$, $\hat{\phi}_0$, and $\tilde{\phi}$ are approximately equivalent in the sense that if

$$\hat{\phi} = \phi(\hat{\psi}, \hat{\lambda}), \quad \hat{\phi}_0 = \phi(\psi^0, \hat{\lambda}_0), \quad \tilde{\phi} = \phi(\tilde{\psi}, \tilde{\lambda})$$

then $\hat{\lambda}$, $\hat{\lambda}_0$ and $\tilde{\lambda}$ are exactly or nearly the same. Whatever the choice of ψ we would have reached nearly the same inference about ϕ, after re-expressing the two estimates on the same ψ scale.

For the normal transformation model the orthogonal parameters are the components of the mean vector and the variance for the untransformed observations: the above argument says that inference on two different ψ scales should be compared *via* transformation to these parameters. Hinkley and Runger (1984) make essentially the same argument from a slightly different point of view: they rescale the observations in order that the maximum likelihood estimates of the regression coefficients β do not depend strongly on the transformation parameter ψ. By property (iv) of Section 2.2 this implies that β and ψ are approximately orthogonal.

4. APPLICATION TO CONDITIONAL INFERENCE

4.1. *Introduction*

Conditioning plays at least two roles in the sampling theory of statistical inference; one to induce relevance of the probability calculations to the particular data under analysis, and the other to eliminate or reduce the effect of nuisance parameters. We concentrate here on the latter.

We deal only with problems in which the parameter ψ of interest is a scalar. This is a nontrivial restriction, although it may be argued that at each stage of interpretation attention can often profitably be focussed on a single parameter describing one aspect of the system under study.

Suppose then that the nuisance parameters $\lambda_1, \ldots, \lambda_q$ have been defined to be orthogonal to ψ, as described in Section 2.3. Confidence intervals for ψ, the usual objective, are approached *via* consideration of tests of the null hypothesis $\psi = \psi^0$, where ψ^0 is a fixed but arbitrary value of ψ. An important general procedure for testing $\psi = \psi^0$ is based on the generalized likelihood ratio statistic

$$w(\psi^0) = 2\{l(\hat{\psi}, \hat{\lambda}) - l(\psi^0, \hat{\lambda}_{\psi^0})\}, \qquad (6)$$

treated as having an asymptotic chi-squared distribution with one degree of freedom, when $\psi = \psi^0$. The approximation to the null distribution can be improved by dividing by a suitable constant, the Bartlett adjustment, (Barndorff-Nielsen and Cox, 1984) or, if equi-tailed tests are desired, by an adjustment for skewness (McCullagh, 1984; Barndorff-Nielsen, 1986). To obtain confidence intervals, it is useful to consider (6) as a function of ψ^0; the term $l(\psi^0, \hat{\lambda}_{\psi^0})$ in (6) is the log-profile likelihood function.

In simple cases the problem can be reduced to one without nuisance parameters. If for each fixed ψ^0 there is a complete sufficient statistic for λ, the likelihood ratio statistic (6) can be constructed from the conditional distribution of the observations given this statistic (Bartlett, 1937; Cox and Hinkley, 1974, p. 134). If the conditional distribution is free of λ, even when $\psi \neq \psi^0$, then the problem has been reduced to a one-parameter problem, and the optimality of (6) for such problems now holds among asymptotically equivalent procedures not depending on λ. Unfortunately, this approach typically only works in important but rather special problems in regular exponential families, with ψ a component of the canonical parameter.

We now explore the extension of the conditional approach to more general problems. We will condition on the observed value of $\hat{\lambda}_{\psi^0}$, the maximum likelihood estimate of λ given $\psi = \psi^0$. Because λ is required to be orthogonal to ψ, the dependence of $\hat{\lambda}_{\psi^0}$ on ψ^0 is reduced. The resulting likelihood is closely related to Barndorff-Nielsen's modified profile likelihood (Barndorff-Nielsen, 1983, p. 351), especially when his approximation to the distribution of the maximum likelihood estimator is used. There are also connections with a long chain of work on conditional and marginal inference (Bartlett, 1936, 1937; Kalbfleisch and Sprott, 1970; Patterson and Thompson, 1971; Godambe and Thompson, 1974; Godambe, 1976; Lindsay, 1982). Note that for those normal theory problems in which the conditioning statistics are linear, conditional and marginal inference are equivalent. In full exponential families the usual approach is to condition on the components of the sufficient statistic that correspond to the nuisance parameters. These of course are just the maximum likelihood estimates of the expectation parameters, which are orthogonal to the canonical parameters.

We wish to derive a conditional profile likelihood for ψ using $\hat{\lambda}_{\psi^0}$ as the conditioning statistic. We write $\hat{\lambda}_0$ when no possibility of confusion exists. Transform y to $(\hat{\lambda}_0, h)$, where h is any convenient function of the observations, and write $J(\hat{\lambda}_0)$ for the Jacobian of the transformation. The conditional density given $\hat{\lambda}_0$ is then

$$f_Y(y; \psi, \lambda)\, J(\hat{\lambda}_0)/f_{\Lambda^0}(\hat{\lambda}_0; \psi, \lambda),$$

where the denominator is the marginal density of $\hat{\lambda}_0$. This leads to a conditional version of the likelihood ratio statistic (6), still a function of λ, in the form

$$2\left[\sup_{\psi}\,\{l_Y(\psi, \lambda) - l_{\Lambda^0}(\psi, \lambda)\} - \{l_Y(\psi^0, \lambda) - l_{\Lambda^0}(\psi^0, \lambda)\}\right].$$

Note that the Jacobian $J(\hat{\lambda}_0)$ no longer appears, the precise choice of h is irrelevant, and the answer is invariant under one-to-one transformations of λ. Finally replace λ by $\hat{\lambda}_0$ to get the conditional profile likelihood

$$w_c^*(\psi^0) = 2\left[\sup_{\psi}\,\{l_Y(\psi, \hat{\lambda}_0) - l_{\Lambda^0}(\psi, \hat{\lambda}_0)\} - \{l_Y(\psi^0, \hat{\lambda}_0) - l_{\Lambda^0}(\psi^0, \hat{\lambda}_0)\}\right]. \tag{7}$$

To calculate this expression it is necessary to compute the marginal distribution of $\hat{\lambda}_0$ for values of ψ different from ψ^0 and this typically involves a noncentral distribution. An alternative statistic can be derived from (7) by conditioning in the first term on $\hat{\lambda}_\psi$ rather than $\hat{\lambda}_0$, leading to

$$2\left[\sup_{\psi}\,\{l_Y(\psi, \hat{\lambda}_\psi) - l_{\Lambda^\psi}(\psi, \hat{\lambda}_\psi) + \log\det J(\hat{\lambda}_\psi) - \log\det J(\hat{\lambda}_0)\}\right.$$
$$\left. - \{l_Y(\psi^0, \hat{\lambda}_0) - l_{\Lambda^0}(\psi^0, \hat{\lambda}_0)\}\right],$$

which is frequently much easier to calculate exactly or approximately. The Jacobian term is log $\det(d\hat{\lambda}_\psi/d\hat{\lambda}_0)$ and (because ψ and λ are orthogonal) is $O_p(1/n)$. We shall therefore ignore this term in what follows, defining

$$w_c(\psi^0) = 2\left[\sup_{\psi}\,\{l_Y(\psi, \hat{\lambda}_\psi) - l_{\Lambda^\psi}(\psi, \hat{\lambda}_\psi)\} - \{l_Y(\psi^0, \hat{\lambda}_0) - l_{\Lambda^0}(\psi^0, \hat{\lambda}_0)\}\right]. \tag{8}$$

A further advantage of (8) is that the first half of the formula does not depend on ψ^0. A disadvantage of (8) is its non-invariance under transformation of λ, although this non-invariance has been reduced by using the orthogonal parametrization. It is perhaps best to regard w_c as defined in some reference λ parametrization. A conceptually curious feature is that two different conditioning events are used, although again the orthogonal parametrization has reduced the difference between them. The same feature arises in the discussion of locally most powerful similar tests; see Cox and Hinkley (1974, p. 146).

Applying the formula in Barndorff-Nielsen (1983) for the marginal distribution of $\hat{\lambda}_\psi$ under ψ and of $\hat{\lambda}_0$ under ψ^0, we have the further approximation

$$\tilde{w}_c(\psi^0) = 2\left(\sup_{\psi}\,[l_Y(\psi, \hat{\lambda}_\psi) - \tfrac{1}{2}\log\det\{nj_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)\}]\right.$$
$$\left. -[l_Y(\psi^0, \hat{\lambda}_0) - \tfrac{1}{2}\log\det\{nj_{\lambda\lambda}(\psi^0, \hat{\lambda}_0)\}]\right). \tag{9}$$

In (9) $j_{\lambda\lambda}$ is the per observation observed information matrix for the λ components. Equation (9) implies that we can regard the effect of conditioning as modifying the objective function for computing the profile likelihood from $l_Y(\psi, \hat{\lambda}_\psi)$ to

$$l_Y(\psi, \hat{\lambda}_\psi) - \tfrac{1}{2}\log\det\{nj_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)\}. \tag{10}$$

The effect of the second term is to penalize values of ψ for which the information about λ is relatively large. It can be shown that the value $\hat{\psi}^c$ at which the supremum of (9) or (8) is achieved satisfies $\hat{\psi}^c - \hat{\psi} = O_p(1/n)$, so that, for some purposes, we write instead of (9),

$$2\{l_Y(\hat{\psi}, \hat{\lambda}) - l_Y(\psi^0, \hat{\lambda}_0)\} - \log\det\{nj_{\lambda\lambda}(\hat{\psi}, \hat{\lambda})\} + \log\det\{nj_{\lambda\lambda}(\psi^0, \hat{\lambda}_0)\}. \tag{11}$$

Note that the term $\det(nj_{\lambda\lambda})$ can be computed as the product of the information determinant in the (ϕ, ψ) parametrization and the square of the determinant of the transformation matrix from (ϕ, ψ) to (λ, ψ).

There is a complication in the derivation of (9) to (11) in that Barndorff-Nielsen's formula for the distribution of the maximum likelihood estimator requires in general conditioning on appropriate ancillary statistics. In the special case when no ancillary is needed for fixed ψ the above argument applies directly. The same holds true if the ancillary statistic does not depend on ψ. These two possibilities cover many common cases, and all the examples in this paper.

Otherwise there is an additional term in the approximate density and hence in (9) to (11) arising from the log-likelihood ratio of the distribution of the ancillary at ψ and ψ^0. It is possible that these ancillary statistics can be approximated by maximum likelihood estimators of constructed orthogonal parameters, possibly by embedding the model in a suitable exponential family. This would imply that the omitted terms are $O_p(1/n)$. We have not, however, explored this in detail.

The difference of (9) from Barndorff-Nielsen's modified profile likelihood is the use of orthogonal parameters which allows us to ignore the term $|\partial\hat{\lambda}_\psi/\partial\hat{\lambda}_0|$. Parameter orthogonality is also essential in the asymptotic expansion of $\tilde{w}_c$ in Section 4.3. Although the factor $|\partial\hat{\lambda}_\psi/\partial\hat{\lambda}_0|$ may be difficult to compute, its inclusion ensures that the modified profile likelihood is parametrization invariant. In the special case (for example full exponential families) where the double saddlepoint approximation of Barndorff-Nielsen and Cox (1984) can be applied to approximate the conditional density, the conditional profile likelihood and modified profile likelihood are both equal to this approximation; see Barndorff-Nielsen (1983, p. 353) and Jorgensen and Pedersen (1979, p. 309). For discussion of the modified profile likelihood derived from a marginal or conditional point of view, see Barndorff-Nielsen (1985b).

The expressions (7)–(11) are in decreasing order of preference from an intuitive point of view, although in many applications $\tilde{w}_c$ is the version most easily implemented. If $w_c = \tilde{w}_c$ this implies that the ancillary statistic discussed above does not depend on ψ, so that Barndorff-Nielsen's formula does give an approximation to the appropriate conditional density.

4.2. *Examples*

We now discuss a number of examples, to illustrate the implementation of the conditional likelihoods discussed in Section 4.1.

4.2.1. *Normal Distribution*

We first consider the parameter of interest to be the variance, τ. In this case $w_c^* = w_c$, and the conditional profile likelihood is simply proportional to the χ^2_{n-1} density of S/τ, where $S = \Sigma(y_i - \bar{y})^2$, and $\tilde{\tau}^c = S/(n-1)$. Both the approximate conditional likelihood and the modified profile likelihood are also proportional to the χ^2_{n-1} density, as the approximation formula is exact (Barndorff-Nielsen, 1983, Example 3.1). No new considerations arise in replacing the mean with a linear regression; w_c^*, w_c, and $\tilde{w}_c$ are all proportional to the log of the χ^2_{n-q} density of S/τ, where now S is the residual sum of squares after regression on q explanatory variables.

Of more interest for illustrating some of the general points of Section 4.1 is the case where the mean μ is the parameter of interest. Computation of w_c is fairly straightforward. We reduce by sufficiency to the joint density of $(\bar{y}, S)$, and transform the joint distribution to that of $(\bar{y}, \hat{\tau}_\mu)$, with Jacobian $1/n$. The marginal density of $\hat{\tau}_\mu$ is proportional to a χ^2_n density, and the required conditional density is proportional to $\hat{\tau}_\mu^{-((n/2)-1)}$. This gives $w_c(\mu^0) = (n-2)\log\{1 + n(\bar{y} - \mu^0)^2/S\}$, a monotone function of the usual t-statistic. Note that the profile likelihood for this problem is $w(\mu^0) = n\log\{1 + n(\bar{y} - \mu^0)^2/S\}$. Again w_c and $\tilde{w}_c$ are identical.

Analysis using the conditional distribution given $\hat{\tau}_0$ leads to the same result, i.e. $w_c^*(\mu^0)$ is a monotone function of the usual t-statistic, but the derivation is somewhat more difficult. The marginal density of $\hat{\tau}_0$ is noncentral χ^2 with n degrees of freedom and noncentrality parameter $n(\mu - \mu^0)^2/\tau$. The required conditional density is a function of μ and τ, although it can be shown to lead to a similar test of the null value μ^0 against values $\mu > \mu^0$ for all positive τ (Cox and Hinkley, 1974 p. 143). This approach can be extended to normal theory regression, although the details are somewhat more complicated.

A normal theory problem where the profile likelihood fails is the problem of weighted means (Neyman and Scott, 1948). Assume $\bar{y}_j$, $j = 1, \ldots, q$, are independently normally distributed with mean μ and variance τ_j/n_j. The conditional density of $\bar{y}_1, \ldots, \bar{y}_q$, given $\hat{\tau}_{\mu 1}, \ldots, \hat{\tau}_{\mu q}$ is proportional to $\Pi_j \hat{\tau}_{\mu j}^{-(1/2 n_j - 1)}$, where $n_j \hat{\tau}_{\mu j} = S_j + n_j(\bar{y}_j - \mu)^2$, and S_j is the residual sum of squares from the jth sample. This gives

$$w_c(\mu^0) = \sum_j (n_j - 2) \log[\{S_j + n_j(\bar{y}_j - \mu^0)^2\}/\{S_j + n_j(\bar{y}_j - \tilde{\mu}^c)^2\}],$$

where $\tilde{\mu}^c$ satisfies

$$\sum_j \frac{(n_j - 2) n_j (\bar{y}_j - \tilde{\mu}^c)}{S_j + n_j(\bar{y}_j - \tilde{\mu}^c)^2} = 0;$$

this is the estimate derived by Bartlett (1936). This solution can again be obtained *via* the modified profile likelihood by ignoring the term $|\partial \hat{\tau}_\mu / \partial \hat{\tau}|$ (Barndorff-Nielsen, 1983, Example 3.6). Since $w_c = \tilde{w}_c$, the approximate ancillary appearing in Barndorff-Nielsen's discussion of this example does not depend on μ. The above estimating equation is also derived in Cox and Hinkley (1974, p. 147) from a slightly different point of view; see also Lindsay (1982). Note that w_c leads directly to the "correct" answer, whereas the expression for w_c^* involves the product of q noncentral χ^2 densities and is quite complicated.

4.2.2. *Exponential Regression*

We consider here the regression model with one covariate; $EY_i = \lambda \exp(-\psi z_i)$, where $\Sigma z_i = 0$. Then

$$l(\psi, \lambda) = -n \log \lambda - \lambda^{-1} \sum y_i \exp(\psi z_i)$$

from which $\hat{\lambda}_\psi = n^{-1} \Sigma y_i \exp(\psi z_i)$ and $\hat{\psi}$ satisfies $\Sigma y_i z_i \exp(\hat{\psi} z_i) = 0$. The profile log-likelihood ratio evaluated at $\psi^0 = 0$ is

$$\begin{aligned} w(0) &= 2\{-n \log(\textstyle\sum (y_i/n) \exp(\hat{\psi} z_i)) + n \log \bar{y}\} \\ &= -2n (\log \hat{\lambda} - \log \hat{\lambda}_0), \end{aligned}$$

where $\hat{\lambda} = \hat{\lambda}_{\hat{\psi}}$ and $\hat{\lambda}_0 = \hat{\lambda}_{\psi^0}$. Both expressions (8) and (9) for the conditional profile likelihood have a one degree of freedom adjustment but lead to the same estimate of ψ:

$$\tilde{w}_c(0) = w_c(0) = -2(n - 1) \log(\hat{\lambda}/\hat{\lambda}_0). \qquad (12)$$

The modified profile likelihood, by including the term $|d\hat{\lambda}/d\hat{\lambda}_\psi|$, in this case proportional to $\hat{\lambda}_\psi^{-1}$, gives

$$-2(n - 2) \log(\hat{\lambda}/\hat{\lambda}_0). \qquad (13)$$

To compute w_c^* we need the marginal density of $\bar{y} = \hat{\lambda}_0$ for an arbitrary value of ψ; then the conditional density needs to be maximized over ψ. Since the marginal density can only be evaluated approximately, it is quite cumbersome to compare the resulting expression for w_c^* to $w(0)$, $w_c(0)$ and (13). A simpler approach is to approximate

$$2[\{l_Y(\psi, \hat{\lambda}_0) - l_{\Lambda^0}(\psi, \hat{\lambda}_0)\} - \{l_Y(\psi^0, \hat{\lambda}_0) - l_{\Lambda^0}(\psi^0, \hat{\lambda}_0)\}], \qquad (14)$$

which corresponds to the definition of w_c^* in (7), but ψ^0 is regarded as fixed at 0 and (14) is a function of ψ. The approximation to (14) is, letting $\psi - \psi^0 = \delta/\sqrt{n}$ and writing m_k for $n^{-1}\Sigma z_i^k y_i$,

$$-\sqrt{n}\,\delta\left(\frac{m_1}{\bar{y}}\right) - \frac{\delta^2}{2}\frac{m_2}{\bar{y}} - \frac{\delta^3}{6\sqrt{n}}\frac{m_3}{\bar{y}} - \frac{\delta^4}{24n}\frac{m_4}{\bar{y}} + \frac{\delta^2}{2n}\left(\frac{\Sigma z_i^2}{n}\right) + \frac{\delta^4}{8n}\left(\frac{\Sigma z_i^2}{n}\right)^2. \tag{15}$$

The corresponding expansions for the unmaximized analogues of w, w_c and $\tilde{w}_c$ are actually quite straightforward to obtain from the above expressions, substituting $\hat{\lambda}_\psi$ for $\hat{\lambda}$ and expanding in terms of $\psi - \psi^0 = \delta/\sqrt{n}$. All three expressions agree with (15) in the leading, $O_p(1)$ term, and differ in the $O_p(1/n)$ terms.

In the two-sample version, letting $z_1 = \ldots = z_{n_1} = -n_2$ and $z_{n_1+1} = \ldots = z_{n_1+n_2} = n_1$, an exact solution is available. Writing $y_{1.}$ for the first sample total and $y_{..}$ for the combined sample total gives

$$f(y_{1.} \mid \hat{\lambda}_0 = y_{..}/n;\, \psi, \lambda) = \frac{y_{1.}^{n_1-1}(y_{..} - y_{1.})^{n_2-1} e^{-\{(1-\theta^2/\theta\lambda\}y_{1.}}}{B(n_1, n_2)\, y_{..}^{n-1} c(y_{..}, \theta, \lambda)}, \quad 0 \leqslant y_{1.} \leqslant y_{..},$$

where $\theta = e^{-\psi}$ and c is a normalizing constant. For $\lambda > 0$ this distribution has monotone likelihood ratio and gives a most powerful similar test of the null hypothesis $\theta = 1$ against alternatives $\theta > 1$, for all $\lambda > 0$. The two sample version of (14) can also be obtained by approximating this density directly. Curiously, the same uniformly most powerful similar test can be obtained by conditioning on $\hat{\lambda}_\psi$ under the alternative and $\hat{\lambda}_0$ under the null; i.e. by computing the exact version of w_c rather that the exact version of w_c^*.

4.2.3. *Gamma Distribution*

The parameter of interest is taken to be the shape parameter ψ; as $\hat{\lambda}_\psi$ is independent of ψ there is no difference between w_c^* and w_c, and $\tilde{w}_c$ differs from these only in the approximation of the normalizing constant. The methods are best compared *via* the estimating equations for ψ. The profile likelihood gives the following equation for the maximum likelihood estimate:

$$\log\hat{\psi} - \Gamma'(\hat{\psi})/\Gamma(\hat{\psi}) = \log(y_{.}/n) - n^{-1}\sum \log y_i.$$

The conditional profile likelihood gives

$$\frac{\Gamma'(n\tilde{\psi}^c)}{\Gamma(n\tilde{\psi}^c)} - \frac{\Gamma'(\tilde{\psi}^c)}{\Gamma(\tilde{\psi}^c)} = \log y_{.} - n^{-1}\sum \log y_i;$$

the comparison of the two is clarified by writing $\Gamma'(n\tilde{\psi}^c)/\Gamma(n\hat{\psi}^c) \simeq \log(n\hat{\psi}^c) - 1/(2n\tilde{\psi}^c)$, which gives

$$\log\tilde{\psi}^c - \frac{1}{2n\tilde{\psi}^c} - \frac{\Gamma'(\tilde{\psi}^c)}{\Gamma(\tilde{\psi}^c)} = \log(y_{.}/n) - n^{-1}\sum \log y_i.$$

This is the same estimating equation that is obtained from the modified profile likelihood. In both cases one "degree of freedom" has been lost, in analogy with the normal variance example. This adjustment is motivated from a different point of view in McCullagh and Nelder (1983, p. 157); see also Sweeting (1981).

4.3. *Comparison of Conditional and Unconditional Profile Likelihood Functions*

We now consider how to assess whether the conditional profile likelihood statistic is preferable to the unconditional form. There are several bases for comparison, no one of which is wholly convincing in itself.

As noted in Section 4.1, in special problems of the exponential family conditioning on $\hat{\lambda}_0$ generates uniformly most powerful similar tests. We can expect this optimality to be nearly retained for distributions close to the exponential family.

Two possibilities we shall not consider in detail are to compare the approximate distributions of w and w_c^* under the null hypothesis and under an appropriate one-sided alternative.

With regard to the first, it would be of some interest as a matter of convenience rather than fundamental principle to examine whether or not the distribution of w_c^* is more nearly approximated by χ_1^2 than that of w. It is known that the χ_1^2 approximation to the distribution of w can be improved by application of a Bartlett adjustment, but we have not investigated such adjustments for w_c or w_c^*.

Note, however, that in the normal theory linear model with the variance as the parameter of interest the Bartlett adjustment essentially allows for the loss of degrees of freedom due to estimating the regression parameters; this adjustment is automatically made by the conditional construction of w_c^*. In general the need for a large adjustment factor, i.e. n^{-1} correction, would make the use of one or two terms of the asymptotic expansion suspect.

With regard to the second, we have not explored higher order approximations to power. The calculations involved are complex and unlikely to lead to a definitive answer.

Comparison with Bayesian calculations is likely to be helpful, and in this regard the results of Sweeting (1981, 1984a, 1984b) are particularly relevant. Sweeting's approximate posterior distributions for location and scale parameters lead to inferences very similar to those here, although the basis of the argument is quite different.

In the development below we examine directly the first two terms of the stochastic expansion of w and w_c.

We will in our discussion concentrate on the conditional statistic w_c, although our original motivation was in terms of w_c^*. It seems likely that $w_c - w_c^* = O_p(1/n)$, but we have not proved this.

We assume that, if λ is known, the optimality results mentioned in Section 4.1 justify the use of the ordinary likelihood ratio statistic which we denote by

$$w_k(\psi^0) = 2\{l(\hat{\psi}_\lambda, \lambda) - l(\psi^0, \lambda)\}.$$

We shall compare w_k to the profile likelihood w defined in (6) and to the approximate version of the conditional profile likelihood, $\tilde{w}_c$, defined in (10).

All three statistics have asymptotically a χ_1^2 distribution under the null hypothesis. The differences $w_k(\psi^0) - w(\psi^0)$ and $w_k(\psi^0) - \tilde{w}_c(\psi^0)$ represent the loss from not knowing λ. We want this loss to be stochastically small. A major advantage of this approach is that the adoption of a very specific measure of the loss is unnecessary, at least for the analysis here. Note that we have defined $w_k(\psi^0)$ in terms of the orthogonalized nuisance parameter λ rather than in terms of an arbitrary nuisance parameter, ϕ. This seems compelling, however, in that to regard ϕ as known would in general add appreciably to the information about ψ, whereas specification of λ affects only second-order aspects of inference about ψ.

We begin by comparing $w(\psi^0)$ and $w_k(\psi^0)$ *via* suitable Taylor series expansions, calculating the term that is $O_p(1/\sqrt{n})$. On expansion about $(\hat{\psi}, \hat{\lambda})$ and $(\psi^0, \hat{\lambda}_0)$, we have

$$\begin{aligned} w_k - w &= 2[\{l(\hat{\psi}_\lambda, \lambda) - l(\hat{\psi}, \hat{\lambda})\} - \{l(\psi^0, \lambda) - l(\psi^0, \hat{\lambda}_0)\}] \\ &= -n(\lambda - \hat{\lambda})\{j_{\lambda\lambda}(\hat{\psi}, \hat{\lambda}) - j_{\lambda\lambda}(\psi^0, \hat{\lambda}_0)\}(\lambda - \hat{\lambda})^{\mathrm{T}} \\ &\quad + 2n(\hat{\lambda} - \hat{\lambda}_0)j_{\lambda\lambda}(\psi^0, \hat{\lambda}_0)(\lambda - \hat{\lambda})^{\mathrm{T}} + O_p(1/n). \end{aligned} \tag{16}$$

The terms retained in (16) are $O_p(1/\sqrt{n})$. In deriving this we have used orthogonality repeatedly to give both $(\hat{\lambda}_0 - \hat{\lambda}) = O_p(1/n)$ and $(\hat{\psi}_\lambda - \hat{\psi}) = O_p(1/n)$, and in the expansion of $l(\psi^0, \lambda) - l(\psi^0, \hat{\lambda}_0)$ we have written $\lambda - \hat{\lambda}_0 = \lambda - \hat{\lambda} + \hat{\lambda} - \hat{\lambda}_0$. It follows from expansion of $\hat{\lambda}_\psi$ as a function of ψ that

$$\hat{\lambda} - \hat{\lambda}_0 = (\hat{\psi} - \psi^0)Z_{\psi\lambda}(\psi^0, \lambda)i_{\lambda\lambda}^{-1}(\psi^0, \lambda)/\sqrt{n} - \tfrac{1}{2}(\hat{\psi} - \psi^0)^2\partial i_{\psi\psi}(\psi^0, \lambda)/\partial\lambda + O_p(n^{-3/2}),$$

where $Z_{\psi\lambda}$ is a random vector of order 1 in probability and $\partial i_{\psi\psi}(\psi^0, \lambda)/\partial\lambda$ is a fixed vector. After some further expansion we have

$$\begin{aligned} w_k - w = & -n(\hat{\psi} - \psi^0)(\lambda - \hat{\lambda})[\partial i_{\lambda\lambda}(\psi^0, \lambda)/\partial\psi](\lambda - \hat{\lambda})^{\mathrm{T}} \\ & + n\{2(\hat{\psi} - \psi^0)(Z_{\psi\lambda}/\sqrt{n})i_{\lambda\lambda}^{-1} - (\hat{\psi} - \psi^0)^2\, \partial i_{\psi\psi}(\psi^0, \lambda)/\partial\lambda\}i_{\lambda\lambda}(\psi^0, \lambda)(\lambda - \hat{\lambda})^{\mathrm{T}} \\ & + O_p(1/n). \end{aligned} \tag{17}$$

To examine the structure of (17) we write

$$\hat{\psi} - \psi^0 = -V_\psi i_{\psi\psi}^{-1/2}/\sqrt{n}, \quad \lambda - \hat{\lambda} = V_\lambda i_{\lambda\lambda}^{-1/2}/\sqrt{n},$$

where $i_{\lambda\lambda} = i_{\lambda\lambda}^{1/2}(\mathrm{i}_{\lambda\lambda}^{1/2})^{\mathrm{T}}$, all the *i*s are evaluated at (ψ^0, λ) and (V_ψ, V_λ) is a $1 \times (q + 1)$ vector of asymptotically independent standard normal random variables. Then

$$\begin{aligned} w_k - w = & \frac{V_\psi i_{\psi\psi}^{-1/2}}{\sqrt{n}} [V_\lambda i_{\lambda\lambda}^{-1/2}\{\partial i_{\lambda\lambda}(\psi^0, \lambda)/\partial\psi\}(i_{\lambda\lambda}^{-1/2})^{\mathrm{T}} V_\lambda^{\mathrm{T}}] \\ & - \frac{V_\psi^2 i_{\psi\psi}^{-1}}{\sqrt{n}} \{\partial i_{\psi\psi}(\psi^0, \lambda)/\partial\lambda\} i_{\lambda\lambda}(i_{\lambda\lambda}^{-1/2})^{\mathrm{T}} V_\lambda^{\mathrm{T}} \\ & - 2V_\psi i_{\psi\psi}^{-1/2} \frac{Z_{\psi\lambda}}{\sqrt{n}} (i_{\lambda\lambda}^{-1/2})^{\mathrm{T}} V_\lambda^{\mathrm{T}} + O_p(1/n). \end{aligned} \tag{18}$$

The corresponding term in the expansion of $w_k - \tilde{w}_c$ has one extra term, arising from

$$\begin{aligned} & \log\det j_{\lambda\lambda}(\hat{\psi}, \hat{\lambda}) - \log\det j_{\lambda\lambda}(\psi^0, \hat{\lambda}_0) \\ & = (\hat{\psi} - \psi^0)\{\partial \log\det i_{\lambda\lambda}(\psi, \lambda)/\partial\psi\}_{\psi = \psi^0} + O_p(1/n) \\ & = (\hat{\psi} - \psi^0)\, \mathrm{trace}\, \{i_{\lambda\lambda}^{-1}(\partial i_{\lambda\lambda}/\partial\psi)_{\psi = \psi^0}\} + O_p(1/n) \\ & = \frac{-V_\psi i_{\psi\psi}^{-1/2}}{\sqrt{n}}\, \mathrm{trace}\, \{i_{\lambda\lambda}^{-1}(\partial i_{\lambda\lambda}/\partial\psi)\} + O_p(1/n). \end{aligned} \tag{19}$$

The trace in this expression is the expected value of the quadratic form in V_λ in (18).

The interpretation is probably most easily seen from the case where λ is a scalar, when we can write

$$w_k - w = (aV_\psi V_\lambda^2 + bV_\psi^2 V_\lambda + cV_\psi V_\lambda)/\sqrt{n} + O_p(1/n), \tag{20}$$

$$w_k - \tilde{w}_c = \{aV_\psi(V_\lambda^2 - 1) + bV_\psi^2 V_\lambda + cV_\psi V_\lambda\}/\sqrt{n} + O_p(1/n). \tag{21}$$

Note that to first order any of the *w* statistics is equal to V_ψ^2.

Suppose that we have collected some data and calculated one or other of w and $\tilde{w}_c$. We would like to have calculated w_k but this is not possible, essentially because V_λ is totally unknown. We therefore consider the conditional representation of w_k given w or $\tilde{w}_c$; these are respectively of the form

$$w + (a'V_\lambda^2 + b'V_\lambda)/\sqrt{n} + O_p(1/n),$$

and

$$w_c + \{a'(V_\lambda^2 - 1) + b'V_\lambda\}/\sqrt{n} + O_p(1/n).$$

On average, $\tilde{w}_c$ is closer to w_k because $EV_\lambda^2 = 1$, although there is no uniform domination. The $O_p(1/\sqrt{n})$ terms are a kind of bias, and the mean squares of these terms in (22) are respectively $(3a'^2 + b'^2)/n$ and $(2a'^2 + b'^2)/n$. Further, among all linear combinations of w and $\tilde{w}_c$, the minimum possible mean square is $(2a'^2 + b'^2)/n$ and in general, unless $|a| \ll |b|$, the probability that the unknown V_λ contributes a large discrepancy from the 'optimal' w_k is greater for the unconditional version w than for the conditional version $\tilde{w}_c$.

An incidental comment is that the addition of an $O_p(1/\sqrt{n})$ term to, say, w_k, only affects its distribution to order $1/n$, provided that the additional terms have conditional mean zero, given w_k, and mild regularity conditions are satisfied (Cox and Reid, 1987).

When λ is a vector the more complicated formulae (18) and (19) hold. In the special case that the components of λ are mutually orthogonal and all components have the same value of $\partial \log i_{\lambda_r \lambda_r}/\partial\psi$, then expressions corresponding to (22) are

$$w + \{a'(V_{\lambda_1}^2 + \ldots + V_{\lambda_q}^2) + b' \sum c_s V_{\lambda_s}\}/\sqrt{n} + O_p(1/n),$$

$$\tilde{w}_c + \{a'(V_{\lambda_1}^2 + \ldots + V_{\lambda_q}^2 - q) + b' \sum c_s V_{\lambda_s}\}/\sqrt{n} + O_p(1/n)$$

and the amount of 'bias' removed by $\tilde{w}_c$ is proportional to q.

In general a simple characterization of the amount of 'bias' does not seem possible. Note, however, that if $i_{\lambda\lambda}$ does not depend on ψ, then w and $\tilde{w}_c$ have the same expansions to $O_p(1/\sqrt{n})$.

In work unpublished at the time of writing, M. A. Aitkin and J. Hinde have proposed another method for deriving a likelihood function in the presence of nuisance parameters via a notion of canonical likelihood. It would be of interest to compare their method with the present ones *via* an expansion of the form (22).

5. TRANSFORMATIONS IN NORMAL THEORY REGRESSION

5.1. *Introduction*

We now discuss in more detail some aspects of inference in the normal transformation model introduced in Example 3.5. For some unknown ψ, the random variables $(Y_1^\psi, \ldots, Y_n^\psi)$ are assumed to be independent and normally distributed with mean ϕ_1 and variance ϕ_0 in the one sample case, and mean $\Sigma x_{is}\phi_s$ in the regression model. An approximately orthogonal parametrization of the model is given in equation (5) of Section 3.5, and a possible interpretation of the statistical implications of this parametrization is outlined. It is essentially the same parametrization developed by Hinkley and Runger (1984) from a different route.

5.2. *Bayesian and Conditional Likelihood Analysis*

The Bayesian analysis of Box and Cox (1964) used a data dependent prior for (ψ, λ), proportional to the nth root of the Jacobian of the transformation from y to y^ψ. This was necessary because the relative sizes of the regression coefficients and variance depend strongly on the value of ψ, so that in the absence of any assumptions regarding ψ it does not make sense to assign uniform improper priors for them. The logical status of data-dependent priors is unclear; see, for example, Nelder's contribution to the discussion of Box and Cox (1964). One method of avoiding them was suggested by Pericchi (1981) and modified by Sweeting (1984a) by an argument similar to that below, although expressed differently; see also Hinkley and Runger (1984).

Since the approximately orthogonal parameters are by construction weakly dependent on ψ, it seems reasonable to assign uniform improper priors for them. Since λ_1 is constrained to be positive, it is given the prior $d\lambda_1/\lambda_1$; similarly the prior for the orthogonal variance component λ_0 is $d\lambda_0/\lambda_0$. The remaining components $(\lambda_2, \ldots, \lambda_q)$ are assigned the joint prior $\Pi d\lambda_s$. For the one-sample problem the likelihood is proportional to

$$f(y; \psi, \lambda_1, \lambda_0) \propto \lambda_1^{-n(\psi-1)} \lambda_0^{-n/2} \Pi y_i^{\psi-1} \exp[-\{n(\bar{y}_\psi - \lambda_1^\psi)^2 + S_\psi\}/(2\lambda_1^{2\psi-2}\psi^2\lambda_0)],$$

where $\bar{y}_\psi$ and S_ψ are the mean and residual sum of squares calculated from y_1^ψ. Integration over the prior $d\lambda_0/\lambda_0$ gives

$$f(y; \psi, \lambda_1) \propto \Pi y_i^{\psi-1} |\psi|^n \{n(\bar{y}_\psi - \lambda_1^\psi)^2 + S_\psi\}^{-n/2} \qquad (23)$$

and to obtain the contribution of the observations to the posterior density of ψ we integrate (23) over $d\lambda_1/\lambda_1$:

$$f(y;\psi) \propto \Pi y_i^{\psi-1}|\psi|^n \int_0^\infty \frac{d\lambda_1}{\lambda_1\{n(\bar{y}_\psi-\lambda_1^\psi)^2+S_\psi\}^{n/2}}.$$

After some computation this reduces to

$$f(y;\psi) \propto \Pi y_i^{\psi-1}|\psi|^n S_\psi^{-(n-1)/2}(\bar{y}_\psi)^{-1}[1+\{S_\psi/(n\bar{y}_\psi^2)\}n^{-1}+O(n^{-2})].$$

Note that $S_\psi/(n\bar{y}_\psi^2)$ is the squared coefficient of variation of the y_i^ψ.

The computation for the linear regression model proceeds similarly, giving

$$f(y;\psi) \propto \frac{\Pi y_i^{\psi-1}|\psi|^{n-q}S_\psi^{-(n-q)/2}}{|\bar{y}_\psi|^q}\left\{1+\frac{q(q+1)}{2(n-q-2)}\frac{S_\psi/n}{\bar{y}_\psi^2}+O(n^{-2})\right\}, \tag{24}$$

the leading term agreeing with Sweeting (1984a, eq. (6)). In the corresponding expression using the data dependent prior (Box and Cox, 1964, equation 22), the term in braces in (24) is equal to 1, and the term

$$\Pi y_i^{\psi-1}|\psi|^{n-q}/(\bar{y}_\psi)^q = |\psi|^{n-q}(\Pi y_i^{\psi-1}/\bar{y}_\psi^q)$$

is replaced by

$$J(\psi;y)^{(n-q)/n} = |\psi|^{n-q}\Pi y_i^{(\psi-1)(n-q)/n}.$$

The simplest direct route to the computation of the conditional profile likelihood is to use the version corresponding to $\tilde{w}_c$ (equation (9)); i.e. the expression to be compared to (2) is $\exp\{l(\psi,\hat{\lambda}_\psi)-\frac{1}{2}\log\det(nj_{\lambda\lambda})\}$. The transformation matrix from (ϕ,ψ) to (λ,ψ) is diagonal with entries $(\psi\lambda_1^{\psi-1}, \psi\lambda_1^\psi, \ldots, \psi\lambda_1^\psi, \psi^2\lambda_1^{2\psi-2})$. The resulting expression for $\exp\{l(\psi,\hat{\lambda}_\psi)-\frac{1}{2}\log\det(nj_{\lambda\lambda})\}$, is, ignoring terms not depending on ψ,

$$\Pi y_i^{\psi-1}|\psi|^{n-q-2}S_\psi^{-(n-q-2)/2}/|\bar{y}_\psi|^{\{(q+2)\psi-3\}/\psi}. \tag{25}$$

The conditional profile likelihood defined by w_c in equation (8) gives the same expression.

Expressions (24) and (25) were evaluated as functions of ψ for the 3 × 4 × 4 factorial design discussed by Box and Cox (1964, Table 1). The Bayesian posterior density (24) has its mode at $\psi=-0.71$ and an equitailed 0.95 posterior interval is $(-1.14, -0.27)$. The conditional profile likelihood (25) is maximized at $\psi=-0.68$ and a 0.95 confidence interval obtained from the χ_1^2 approximation is $(-1.09, -0.26)$. Box and Cox obtained -0.75 for the Bayesian and likelihood estimates of ψ and corresponding intervals $(-1.18, -0.32)$ and $-1.13, -0.37)$.

One advantage of the parameter orthogonalization is the approximate result

$$\{\mathrm{var}(\hat{\psi})\}^{-1} = E(-\partial^2 l/\partial\psi^2), \tag{26}$$

so the inversion of the full information matrix is unnecessary. The value of (26) can be used to measure the transformation potential of a set of data (Box and Cox, 1982), i.e. the extent to which it is feasible to determine a suitable transformation from the data. A complicated but elementary calculation gives that (26) is equal to

$$n\left\{\frac{7}{4}CV_e^2+\frac{5}{2}CV_\Delta^2+\frac{1}{4}\frac{CV_\Delta^4}{CV_e^2}(1+c_\Delta)\right\}.$$

Here

$$CV_\Delta^2 = n^{-1}\sum\Delta_i^2 = n^{-1}\sum_i\left(\sum_{s=2}^t x_{is}\lambda_s\right)^2$$

is the squared coefficient of variation from the regression component, c_Δ is defined by $n^{-1}\Sigma\,\Delta_i^4 = CV_\Delta^4(1+c_\Delta)$, and $CV_e^2 = \Sigma_i \mathrm{var}\, Y_i/(EY_i)^2$ is the coefficient of variation of the error component, at $\psi=1$; CV_Δ^2/CV_e^2 is a kind of signal to noise ratio. In the one-sample problem $CV_\Delta = 0$.

6. DISCUSSION

The above development leaves open a number of issues some of which we raise in the form of questions.

(i) In Section 4 we concentrated on the relation between w_k, w, and w_c rather than on the null distribution. The first two statistics can be modified by a Bartlett adjustment factor to have a χ^2 distribution to $O(1/n^{3/2})$ (Barndorff-Nielsen and Cox, 1984). Is the same true of w_c and can the adjustment factor be calculated *via* that of w or w_k? Is an adjustment for skewness of w_c available, to produce nearly equitailed confidence limits for ψ (McCullagh, 1984; Barndorff-Nielsen, 1985a)?

(ii) Can stronger justification for the use of w_c or w_c^*, or some other statistic, be produced, including perhaps asymptotic calculations to higher order?

(iii) Has conditioning on exact or approximate ancillary statistics been achieved by the proposed procedure?

(iv) Do the results in Sections 2 and 4 have a useful, possibly simpler, formulation in curved exponential families?

(v) Are there special problems for discrete data?

(vi) Can the discussion be extended to nonregular problems, for example those connected with the terminal of a distribution, and to general problems with a large number of nuisance parameters?

(vii) Are there implications when the objective is the prediction of future observations, rather than estimation?

(viii) Can the discussion usefully be extended to vector parameters of interest, where in general only local orthogonality is possible?

(ix) How should the differential equations determining λ be handled when simple explicit solution is not feasible? What further conditions can usefully be imposed on λ in general?

(x) What general implications for model and parameter definition and robustness can be drawn from the notion of parameter orthogonality?

This paper was substantially completed while both authors were visiting the Department of Statistics at the University of Waterloo, and we thank the department warmly for its hospitality. We especially thank the departmental secretary Lynda Hohner. We are grateful also to referees for constructive comments.

Support of the Science and Engineering Research Council and the Natural Sciences and Engineering Research Council of Canada is gratefully acknowledged.

REFERENCES

Amari, S.-I. (1982) Differential geometry of curved exponential families—curvatures and information loss. *Ann. Statist.*, **10**, 357-85.

———(1985) *Differential geometry in statistics.* New York: Springer Verlag.

Barndorff-Nielsen, O. E. (1978) *Information and exponential families in statistical theory.* New York: Wiley.

———(1983) On a formula for the distribution of a maximum likelihood estimator. *Biometrika*, **70**, 343–65.

———(1985a) Confidence limits from $c|\hat{j}|^{1/2}\bar{L}$, in the single-parameter case. *Scand. J. Statist.*, **12**, 83–87.

———(1985b) Properties of modified profile likelihood. In *Contributions to Probability and Statistics in Honour of Gunnar Blom* (J. Lanke and G. Lindgren, eds), pp. 25–38. Lund.

———(1986) Inference on full or partial parameters, based on the standardized signed log likelihood ratio *Biometrika*, **73**, 307–22.

Barndorff-Nielsen, O. E. and Cox, D. R. (1984) Bartlett adjustments to the likelihood ratio statistic and the distribution of the maximum likelihood estimator. *J. R. Statist. Soc.* B, **46**, 483-95.

Bartlett, M. S. (1936) The information available in small samples. *Proc. Camb. Phil. Soc.*, **34**, 33-40.

———(1937) Properties of sufficiency and statistical tests. *Proc. R. Soc.* A, **160**, 268–82.

Box, G. E. P. and Cox, D. R. (1964) An analysis of transformations (with discussion). *J. R. Statist. Soc.* B, **26**, 211-52.

———(1982) An analysis of transformations revisited, rebutted. *J. Amer. Statist. Assoc.*, **77**, 209–10.

Cox, D. R. (1980) Local ancillarity. *Biometrika*, **67**, 273-8.

Cox, D. R. and Hinkley, D. V. (1974) *Theoretical Statistics.* London: Chapman and Hall.

Cox, D. R. and Reid, N. (1987) Approximations to noncentral distributions. *Can. J. Statist.*, to appear.

Cox, D. R. and Snell, E. J. (1981) *Applied Statistics.* London: Chapman and Hall.

Efron, B. and Hinkley, D. V. (1978) Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika*, **65**, 457-87.

Godambe, V. P. (1976) Conditional likelihood and unconditional optimum estimating equations. *Biometrika*, **63**, 277-84.

Godambe, V. P. and Thompson, M. E. (1974) Estimating equations in the presence of a nuisance parameter. *Ann. Statist.*, **2**, 568-71.

Hinkley, D. V. and Runger, G. (1984) The analysis of transformed data. *J. Amer. Statist. Assoc.*, **79**, 302-20.

Huzurbazar, V. S. (1950). Probability distributions and orthogonal parameters. *Proc. Camb. Phil. Soc.*, **46**, 281-4.

Jeffreys, H. (1961) *Theory of Probability*, 3rd ed. Oxford: Clarendon Press.

Jorgensen, B. and Pedersen, B. V. (1979) Contribution to discussion of paper by O. E. Barndorff-Nielsen and D. R. Cox. *J. R. Statist. Soc.* B, **41**, 305.

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

Lindsay, B. (1982) Conditional score functions: some optimality results. *Biometrika*, **69**, 503-12.

McCullagh, P. (1984) Local sufficiency. *Biometrika*, **71**, 233-44.

McCullagh, P. and Nelder, J. A. (1983) *Generalized linear models*. London: Chapman and Hall.

Mitchell, A. F. S. (1962) Sufficient statistics and orthogonal parameters. *Proc. Camb. Phil. Soc.*, **58**, 326-37.

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

Patterson, H. D. and Thompson, R. (1971) Recovery of interblock information when block sizes are unequal. *Biometrika*, **58**, 545-54.

Pericchi, L. R. (1981) A Bayesian approach to transformations to normality. *Biometrika*, **68**, 35-43.

Ross, G. J. S. (1970) The efficient use of function minimization in nonlinear maximum likelihood estimation. *Appl. Statis.*, **19**, 205-21.

Sweeting, T. J. (1981) Scale parameters: a Bayesian treatment. *J. R. Statist. Soc.* B, **43**, 333-338.

———(1984a) On the choice of prior distribution for the Box-Cox transformed linear model. *Biometrika*, **71**, 127-134.

———(1984b) Approximate inference in location-scale regression models. *J. Amer. Statist. Assoc.*, **79**, 847-852.

DISCUSSION OF THE PAPER BY PROFESSORS COX AND REID

Professor O. E. Barndorff-Nielsen (Aarhus University): The subject of inference on interest parameters in the presence of nuisance parameters is at the core of statistics, and the paper before us adds substantially to our understanding of and methodology for this subject.

The main points of the paper are the discussion of parameter orthogonality and its relevance for inference, and the definition and investigation of a new concept of conditional likelihood. Below I comment on these in turn.

Let ψ, of dimension r, denote the parameter of interest. As the authors demonstrate—and use extensively—if ψ is one-dimensional it is generally possible to find a complementary parameter $\lambda = (\lambda_1, \ldots, \lambda_q)$ such that ψ and λ are orthogonal relative to expected information metric i on the parametric model $\mathscr{M}$. It is illuminating to view this result from a general geometric vantage point.

For this purpose, suppose $\mathscr{M}$ is an arbitrary differentiable manifold of dimension $q+1$ and with metric tensor γ, and let ψ be a real valued function on $\mathscr{M}$, the level sets $\mathscr{M}_\psi$ of ψ being q-dimensional submanifolds of $\mathscr{M}$. At each point p of each submanifold $\mathscr{M}_\psi$ we may place an infinitesimal line segment which contains p and is γ-is orthogonal to $\mathscr{M}_\psi$. It is intuitively plausible, and on account of Frobenius' theorem generally true, that these infinitesimal line segments connect up to form a bundle of one-dimensional differentiable curves, each curve cutting orthogonally through the submanifolds $\mathscr{M}_\psi$. Now, let $\lambda = (\lambda_1, \ldots, \lambda_q)$ be a parameter which is complementary and γ-orthogonal to ψ, i.e. (λ, ψ) parametrizes $\mathscr{M}$ and when γ is expressed in the (λ, ψ) coordinates its mixed type elements are 0, i.e. $\gamma_{\lambda_s\psi}(\lambda, \psi) = 0$ for $s = 1, \ldots, q$. Any such parameter λ may be conceived as determining a coordinate system on a fixed, but arbitrary, of the submanifolds, $\mathscr{M}_{\psi^0}$ say, the (λ, ψ) coordinates of an arbitrary point p in $\mathscr{M}$ being obtained by finding the ψ such that p belongs to $\mathscr{M}_\psi$ and the λ such that p lies on that of the above-mentioned curves whose intersection point with $\mathscr{M}_{\psi^0}$ has coordinate λ. Thus the freedom in choice of orthogonal parameter λ consists solely in the arbitrariness with which one may define a coordinate system on $\mathscr{M}_{\psi^0}$. If $(\phi, \psi) = (\phi_1, \ldots, \phi_q, \psi)$ is any parametrization of $\mathscr{M}$ then an orthogonal complementary parameter λ can be found by solving the system of equations

$$\gamma_{\phi_s\phi_t}(\phi, \psi)\frac{\partial\phi_s}{\partial\psi} = -\gamma_{\phi_t\psi}(\phi, \psi), \quad t = 1, \ldots, q. \tag{1}$$

I have benefitted from discussing this geometrical setting with Professor Suresh Moolgavkar.

Returning to the case of $\mathscr{M}$ being a parametric statistical model it follows, in particular, that when ψ is one-dimensional we can equally find a complementary parameter λ which is orthogonal to ψ relative to the observed information metric $\not{j}$, as defined in Barndorff-Nielsen (1986a,b).

For example, suppose $x_1, \ldots, x_n$ is a sample from the location-scale model $\sigma^{-1}f((x-\mu)/\sigma)$, let a be the configuration $((x_1-\hat{\mu})/\hat{\sigma}, \ldots, (x_n-\hat{\mu})/\hat{\sigma})$ and consider μ as the parameter of interest. Then, letting $g(x) = -\log f(x)$ and solving (1) with $\gamma = j$ and

$$\not{j} = \sigma^{-2}\begin{bmatrix} \Sigma g''(a_v) & \Sigma a_v g''(a_v) \\ \Sigma a_v g''(\mathrm{a}_v) & n + \Sigma a_v^2 g''(a_v) \end{bmatrix},$$

one finds that $\lambda = \sigma + u\mu$ is $\not{j}$-orthogonal to μ, where

$$u = \{\Sigma a_v g''(a_v)\}/\{n + \Sigma a_v^2 g''(a_v)\}.$$

A general remark is that, whether one considers expected or observed information, as a point of principle the inference on ψ ought not to depend on the choice of orthogonal parameter λ.

Let ω, of dimension $d = q + r$, be any parametrization of the statistical model $\mathscr{M}$. In their discussion of conditional likelihood Professors Cox and Reid refer to the formula $p(\hat{\omega}; \omega \mid a) \doteq p^*(\hat{\omega}; \omega \mid a)$ where $p^*(\hat{\omega}; \omega \mid a) = c\surd|\hat{j}|\exp(l-\hat{l})$, which provides an expression for the conditional distribution of the maximum likelihood estimator $\hat{\omega}$ given a complementary ancillary a, and to the related concept of the modified profile likelihood for ψ, i.e.

$$\tilde{L}^0 = \left|\frac{\partial\hat{\phi}}{\partial\hat{\phi}_\psi}\right| |\hat{j}^\psi|^{-1/2}\tilde{L}(\psi) \tag{2}$$

where $\omega = (\phi, \psi)$, $\tilde{L}(\psi) = L(\hat{\phi}_\psi, \psi)$ is the profile likelihood for ψ, j^ψ is observed information given ψ, and $\hat{\phi}$ is considered as a function of $(\hat{\phi}_\psi, \hat{\psi}, a)$. Before commenting on the Cox-Reid definition of conditional likelihood I wish to make a few remarks on modified profile likelihood and parameter orthogonality.

If ϕ and ψ are orthogonal then, by (iv) of Section 2.2, we may often ignore the factor $|\partial\hat{\phi}/\partial\hat{\phi}_\psi|$, which may be a considerable simplification. However, the parametrization invariance of (2) is lost by this approximation.

The expression (2) was derived, using p^* above, by two routes in Barndorff-Nielsen (1983, 1985b): by reasoning of marginal inference on the assumptton that $p(\hat{\phi}, \hat{\psi}; \phi, \psi \mid a)$ factorizes as $p(\hat{\psi}; \psi \mid a)p(\hat{\phi}; \phi, \psi \mid \hat{\psi}, a)$ and by reasoning of conditional inference on the assumption that $p(\hat{\phi}, \hat{\psi}; \phi, \psi \mid a)$ factorizes as $p(\hat{\phi}; \phi, \psi \mid a)p(\hat{\psi}; \psi \mid \hat{\phi}, a)$. In the light of tonight's paper two other routes are now apparent. First, if $p(\phi_\psi, \hat{\psi}; \phi, \psi \mid a)$ factorizes as $p(\hat{\phi}_\psi; \phi, \psi \mid a)p(\hat{\psi}; \psi \mid \hat{\phi}_\psi, a)$, i.e. if for every fixed ψ the statistic $(\hat{\phi}_\psi, a)$, rather than $\hat{\phi}$, is sufficient for ϕ, then again (2) emerges by applying p^*, separately to the numerator and the denominator of $p(\phi_\psi, \hat{\psi}; \phi, \psi \mid a)/p(\hat{\phi}_\psi; \phi, \psi \mid a)$. (We illustrate this further below, by an example based on the Γ-distribution.) This argument is rather similar to that leading to formula (9) of the paper. Second, suppose $\pi(\phi)$ is a prior probability density function for ϕ. Then the posterior likelihood for ψ is

$$L^\pi(\psi) = \int e^{l(\phi, \psi)}\pi(\phi)d\phi.$$

In wide generality we may apply Laplace's approximation method to this integral and we thus obtain

$$L^\pi(\psi) = (2\pi)^{\frac{q}{2}}\pi(\hat{\phi}_\psi)|\hat{j}^\psi|^{-1/2}\tilde{L}(\psi).$$

Provided ϕ and ψ are orthogonal this is generally close to $\tilde{L}^0(\psi)$ on account of 2.2(iv) (and ignoring constant factors).

The idea of conditional profile likelihood as defined by (7) or (8) is certainly appealing, although the lack of parametrization invariance of (8) is somewhat disconcerting.

However, it is not clear to me to what extent (7) or (8) offer an advantage in practice over the modified profile likelihood. An inherent difficulty lies in the need to calculate the marginal distribution of $\hat{\lambda}_\psi$, either exactly or to the appropriate order of approximation. One possibility, when working conditionally on the ancillary a, is to derive an approximate expression for the distribution of $\hat{\lambda}_\psi$ by integrating $p^*(\hat{\lambda}_\psi, \hat{\psi}; \lambda, \psi \mid a) = p^*(\hat{\lambda}, \hat{\psi}; \lambda, \psi \mid a)|\partial\hat{\lambda}/\partial\hat{\lambda}_\psi|$ with respect to $\hat{\psi}$, this being feasible to sufficient approximation by means of an asymptotic expansion for p^*, given in section 3 of Barndorff-Nielsen (1986a). As another point of comparison with modified profile likelihood one may note that (7) and (8) are tied to the conditionality viewpoint whereas, as mentioned above, exactly the same expression (2) for $\tilde{L}^0$ is arrived at by an argument of marginal inference, when that is relevant, and by an argument of conditional inference, when that is relevant. To illustrate this, suppose x_1 and x_2 are independent Poisson variates with means μ_1 and μ_2, respectively, and let $\psi = \mu_1 + \mu_2$ and $\lambda = \mu_1/(\mu_1 + \mu_2)$. Then ψ and λ are orthogonal and $\hat{\psi} = x_1 + x_2$, $\hat{\lambda}_\psi = \hat{\lambda} = x_1/ (x_1 + x_2)$. Inference on ψ should clearly be performed in the marginal (Poisson) model for $\hat{\psi}$, and the modified profile likelihood for ψ is equal to the likelihood from that model. However, calculation of w_c^* or w_c would proceed via the distributon of $\hat{\lambda}$, which is quite complicated, and the result would not be equal to the marginal likelihood based on $\hat{\psi}$.

Quite a different way of defining a likelihood-like function for ψ alone would be to consider the function $\phi(r_\psi^*)$ where ϕ is the standard normal density and r_ψ^* is the standardized signed log likelihood ratio for ψ as defined in Barndorff-Nielsen (1986—see main paper's reference list).

An instructive example is provided by the Γ-distribution $\{(\lambda/\psi)^\lambda/\Gamma(\lambda)\}x^{\lambda-1} \exp\{-(\lambda/\psi)x\}$, whose mean value is ψ. Here ψ and λ are orthogonal and (for sample size $n > 1$) we have that $\hat{\lambda}_\psi$, but not $\hat{\lambda}$, is sufficient for λ given ψ. Thus, in this case, only the second of the above-mentioned two derivations of modified profile likelihood from a viewpoint of conditional inference applies. One finds

$$\tilde{L}^0(\psi) = n^{-1/2}\zeta(\hat{\lambda})^{-1}\zeta(\hat{\lambda}_\psi)^{1/2}\tilde{L}(\psi)$$

where $\zeta(\lambda) = \partial^2 \log \Gamma(\lambda)/\partial\lambda^2 - \lambda^{-1}$. In the present case the factor $|\partial\hat{\lambda}/\partial\hat{\lambda}_\psi|$ in (2) is equal to $\zeta(\hat{\lambda}_\psi)/\zeta(\hat{\lambda})$, and one would not discard this. Calculation of w_c^* or of w_c is not very tractable in the present example (although the null distribution of $\hat{\lambda}_\psi$ does not depend on ψ), but one might compare $\tilde{l}^0(\psi) = \tilde{l}(\psi) + \frac{1}{2} \log \zeta(\hat{\lambda}_\psi)$ to $\log \phi(r_\psi^*)$. It was shown in Barndorff-Nielsen (1986) that $r_\psi^* = r_\psi - (\log \mathscr{K}_\psi)/r_\psi$ where $\mathscr{K}_\psi = \{(\zeta(\hat{\lambda}_\psi)/\zeta(\hat{\lambda})\}^{1/2}n^{-1/2}\hat{\lambda}^{-1/2}(\bar{x}/\psi - 1)^{-1}r_\psi$.

It has been possible here to touch only upon the most basic aspects of the paper. No doubt the paper will generate considerable further discussion and investigations. It is a pleasure and a privilege to propose a strong vote of thanks to Professors Cox and Reid for a very stimulating and interesting paper.

Dr T. J. Sweeting (University of Surrey): I am very pleased to have been asked to second the vote of thanks for tonight's paper, which for me has been very thought provoking. In my discussion I should like to concentrate on the use of orthogonal parameters in Bayesian inference and discuss some relationships with the present work.

There are several reasons for wishing to consider an orthogonal parametrization. I can identify four main reasons, all of which are mentioned by Professors Cox and Reid; they are used as an aid to (i) computation (ii) approximation (iii) interpretation and (iv) elimination of nuisance parameters. Orthogonal parameters are also useful in Bayesian inference for the same reasons. The first part of my discussion concerns the relationship between the conditional profile likelihood (*CPL*) and an approximate Bayesian integrated likelihood, and amplifies remarks already made by Professor Barndorff-Nielson tonight. In the second part I consider the question of prior independence of orthogonal parameters.

Let ψ be a scalar parameter of interest and ϕ a vector of nuisance parameters. Let $L(\psi, \phi)$ be the likelihood function and $\pi(\phi\,|\,\psi)$ the conditional prior density of ϕ given ψ. The integrated likelihood $L(\psi)$ *of* ψ is

$$L(\psi) = \int L(\psi, \phi)\, \pi(\phi\,|\,\psi)d\phi$$

and by taking appropriate expansions of $\log L(\psi, \phi)$ and $\log \pi(\phi\,|\,\psi)$ about $\hat{\phi}_\psi$ one obtains (under suitable regularity conditions)

$$L(\psi) = \pi(\hat{\phi}_\psi\,|\,\psi)L(\psi, \hat{\phi}_\psi)\, |j_{\phi\phi}(\psi, \hat{\phi}_\psi)|^{-1/2}\, (1 + O_p(n^{-1})). \tag{1}$$

This is essentially a Laplace approximation to the above integral, and may be compared with formula (4.1) in Tierney and Kadane (1986) for a marginal posterior density. In that paper however, L is expanded about the conditional posterior mode of ϕ, rather than $\hat{\phi}_\psi$. When ϕ and ψ are orthogonal, the Cox-Reid *CPL* (10) is just formula (1) without the first term. But then we can replace $\hat{\phi}_\psi$ in $\pi(\hat{\phi}_\psi\,|\,\psi)$ by $\hat{\phi}$ to the same order of aproximation. It follows that to $O_p(n^{-1})$, the *CPL* is equal to the integrated likelihood whenever the orthogonal parameters ψ and ϕ are taken to be a priori independent. An interesting feature in this case is that, to $O_p(n^{-1})$, the posterior distribution of ψ is free from the prior adopted for the nuisance parameter ϕ.

The above analysis explains the agreement between the *CPL* in Cox and Reid and the approximate marginal posterior density for the Gamma index in Sweeting (1981). Formula (1) also explains the discrepancy between formulae (24) and (25) for the integrated likelihood and *CPL* respectively. The leading term in (24) is precisely (1), but the *CPL* cannot be the same to $O_p(n^{-1})$ here since $\phi = (\lambda_0, \lambda_1)$ is not exactly orthogonal to ψ and $\pi(\phi)$ is not constant. It is readily checked here that

$$\pi(\hat{\phi}_\psi) \propto \psi^2 S_\psi^{-1} \bar{y}_\psi^{(2\psi-3)/\psi}$$

and on multiplying (25) by this factor we recover (24).

Returning to the question of whether a priori independence is sensible for orthogonal parameters, we have seen that when this is the case things work out nicely to $O_p(n^{-1})$; the *CPL* agrees with the Bayesian likelihood for every smooth prior for ϕ. Although one cannot argue that orthogonal parameters should always be taken *a priori* independent, in certain problems it does seem very natural to take them at least approximately independent.

Consider again the normal transformation model. As the transformation index ψ varies, one can identify directions in (ψ, ϕ) space along which there is very little local change in the model. Reparametrize so that these directions correspond to $\lambda = \text{constant}$. If our prior opinion about λ given ψ is formed by considering the type of data we would expect to see, then we can argue that our beliefs about λ given $\psi = \psi_0$ should hardly be affected when ψ moves to a neighbouring value $\psi = \psi_1$. No *compensation* in λ is required for this small change in ψ to preserve the model. Such an argument is made in Sweeting (1984a), and it turns out that the resulting parametrization agrees with Cox and Reid's approximate orthogonal parametrization. This is not so surprising when one views the process of orthogonalizing to ψ as one of finding directions of least model change under the information metric. Omitting details, local distance in model space is minimized at each point of the parameter space by moving in a direction $\phi(\psi)$ satisfying the orthogonality equation (4). In model space, this amounts to moving from $M(\psi_0, \phi_0)$ in a direction orthogonal to the space $M(\psi_0 + d\psi_0, \phi)$.

A direct "compensation" argument applies quite generally to arbitrary transformations and error distributions (Sweeting, 1985), and for the reasons given above the resulting parametrization should approximate the orthogonal parametrization, which will normally be complex. It would be interesting to find other models for which a simple compensation argument can be made when an exact orthogonal parametrization is difficult. I am sure there will be many other interesting avenues of research arising from tonight's paper, and it gives me very great pleasure to second the vote of thanks.

The vote of thanks was passed by acclamation.

Professor R. L. Smith (University of Surrey): The three-parameter Weibull distribution

$$F(x;\, \theta, \phi, \alpha) = 1 - \exp\{-((x-\theta)/\phi)^\alpha\} \qquad (x \geqslant \theta),$$

with θ the parameter of interest is an inference problem harder than those in the paper, although within the domain of the theory, the problem being regular for $\alpha > 2$.

J. Naylor and I, in a paper as yet unpublished, have compared sampling-theory and Bayesian analyses, a difficulty with the former being that the profile likelihood for θ ends to be very flat. To try to improve on profile likelihood inference, one may solve the orthogonalization equations $\alpha = \alpha(\theta, \lambda, \mu)$, $\phi = (\theta, \lambda, \mu)$ such that

$$\frac{\partial \phi}{\partial \theta} = f_1(\alpha),\ \frac{\partial \alpha}{\partial \theta} = \frac{f_2(\alpha)}{\phi} \tag{1}$$

where

$$f_1(\alpha) = \frac{6}{\pi^2}\,\Gamma\left(2 - \frac{1}{\alpha}\right)\left\{\gamma(1-\gamma) - \frac{\pi^2}{6}\right\}\left\{1 + (1-y)\Psi\left(1 - \frac{1}{\alpha}\right)\right\},$$

$$f_2(\alpha) = \frac{6\alpha^2}{\pi^2}\,\Gamma\left(2 - \frac{1}{\alpha}\right)\left\{\gamma(1-\gamma) - \frac{\pi^2}{6} + (1-\gamma)\Psi\left(1 - \frac{1}{\alpha}\right)\right\}.$$

Here γ is Euler's constant, Γ the gamma function, and Ψ the digamma function.

The solution of (1) is of course quite straightforward. Defining

$$f_3(\alpha) = \frac{f_2'(\alpha) - f_1(\alpha)}{f_2(\alpha)},$$

$$g(\alpha) = \int_b^\alpha \exp\left\{-\int_b^v f_3(u)du\right\}dv$$

where $b > 1$ is arbitrary, we have one solution of (1) in the form

$$g(\alpha) = \lambda\theta + \mu, \; \phi = \frac{f_2(\alpha)g'(\alpha)}{\lambda}, \tag{2}$$

and another, putting the contants of integration in a different place, in the form

$$g(\alpha) = \frac{\theta - \lambda}{\mu}, \quad \phi = \mu f_2(\alpha)g'(\alpha). \tag{3}$$

This defines two orthogonal parametrizations. A third suggested by analogy with the generalized extreme value parametrization (Prescott and Walden, 1980, 1983) is

$$F(x;\, \theta,\, \lambda,\, \mu) = 1 - \exp\left[-\left\{\frac{\lambda(x-\theta)}{\mu}\right\}^{1/\lambda}\right], \qquad x \geqslant \theta. \tag{4}$$

The resulting five forms of log profile likelihood, namely (a) the original, (b) the modified log profile likelihood, i.e. equation (10) of Cox and Reid, without any reparametrization, and (c)–(e) the modified profile likelihoods defined with respect to the three new parametrizations (2)–(4), have been tried on some data on strengths of glass fibre analyzed by Naylor and me. In Fig. D1, the unmodified log profile likelihood (a) is very flat but (b) and (c) are even worse, having no local maximum within the range of values calculated. In contrast, (d) and (e) appear to do better in discriminating among the various values of θ. Preliminary results from a simulation study confirm the picture suggested by Fig. D1, i.e. that, in terms of sampling properties, (d) and (e) are best with (b) and (c) worse than (a).

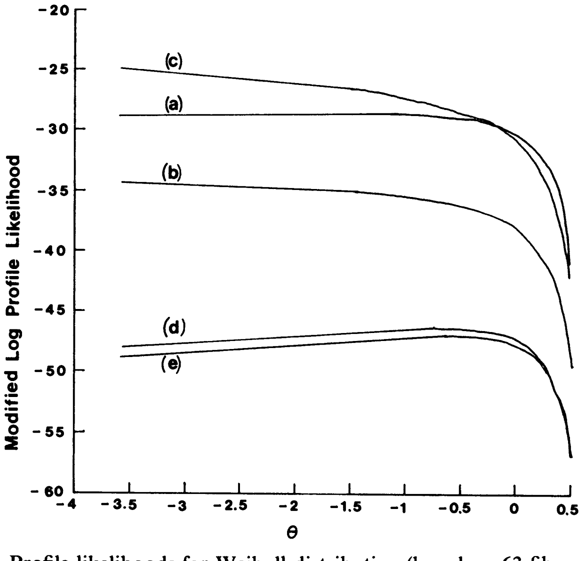

Fig. D1. Profile likelihoods for Weibull distribution (based on 63 fibre strengths).

This example may be very specialised and badly behaved but allows some general observations. Cox and Reid have performed valuable work in drawing attention to the importance of orthogonality, thereby extending Barndorff-Nielsen's definition of modified profile likelihood. However, the example, specially the bad performance of (c) contrasted with (d), shows that orthogonality is by no means the whole story. In particular, two different orthogonal parametrizations may have very different properties.

Mr D. Firth (Imperial College, London): My remarks concern the important property (iv) of Section 2.2, and perhaps have some relevance to question (x) of Section 6.

Observe first a slightly different, apparently more direct route to property (iv), based on the approximation

$$(\hat{\psi}_\lambda - \hat{\psi})\frac{\partial u}{\partial \psi}\bigg|_{(\hat{\psi},\hat{\lambda})} = (\hat{\lambda} - \lambda)\frac{\partial u}{\partial \lambda}\bigg|_{(\hat{\psi},\hat{\lambda})} + O_p(\|\theta - \hat{\theta}\|^2)$$

derived by expandng the *score function* $u(\psi, \lambda) = \partial l(\psi, \lambda)/\partial\psi$ rather than $l(\psi, \lambda)$ itself. Orthogonality implies $(\partial u/\partial\lambda)|_{(\hat{\psi},\hat{\lambda})} = O_p(\sqrt{n})$, and arguments like those following (3) apply: provided $\hat{\lambda} - \lambda = O_p(1/\sqrt{n})$ all terms are $O_p(1)$, hence $\hat{\psi}_\lambda - \hat{\psi} = O_p(1/n)$. Note that the result is 'local' in that λ is required to be within $O(1/\sqrt{n})$ of the true value; in particular, if λ is fixed it must be the true value.

The result may be extended in two stages. First it may be 'delocalized' by restricting attention to likelihoods that satisfy

$$E_{\psi,\xi}\{\partial^2 l(\psi, \lambda)/\partial\psi\partial\lambda\} = 0 \quad \textit{for all } \psi, \xi, \lambda.$$

This is a stronger condition than orthogonality and implies, in particular, that the score equation $u(\psi, \lambda) = 0$ is an unbiased estimating equation for ψ at every λ. Now consider arbitrary λ, no longer required to be near the true value; and suppose that $\hat{\lambda}$, rather than being the maximum likelihood estimate, is such that $\hat{\lambda} - \lambda = O_p(1/\sqrt{n})$. Then, with $\hat{\psi} = \hat{\psi}_{\hat{\lambda}}$, the behaviour of all quantities in the above expansion is as before, and in particular $\hat{\psi}_\lambda - \hat{\psi} = O_p(1/n)$.

An immediate further extension is to the situation where $\{u(\psi, \lambda) : \lambda \in \mathbb{R}\}$ is a more general family of estimating functions for ψ, not necessarily likelihood-based score functions; the required condition is still

$$E_\psi\{u(\psi, \lambda)\} = 0 \quad \textit{for all } \psi, \lambda,$$

i.e. $u(\psi, \lambda) = 0$ is an unbiased estimating equation for ψ at every value of λ. The result $\hat{\psi}_\lambda - \hat{\psi} = O_p(1/n)$ implies in particular that the asymptotic (normal) distribution of a solution based on any fixed value λ is the same as that of a solution based on a data-dependent value $\hat{\lambda}$, provided $\hat{\lambda} - \lambda = O_p(1/\sqrt{n})$. Consider two examples:

(i) $Y_1, \ldots, Y_n$ independent, $E(Y_i) = \psi x_i$, $\text{var}(Y_i) = \{E(Y_i)\}^{\lambda_0}$ and $u(\psi, \lambda) = \Sigma(y_i - \psi x_i)/(\psi x_i)^\lambda$. Within this class, $u(\psi, \lambda_0)$ maximizes asymptotic efficiency; the same first-order efficiency is achieved if λ_0 is replaced by a $\sqrt{n}$-consistent estimate.

(ii) $Y_1, \ldots, Y_n$ i.i.d., $E(Y_i) = \psi$, $\text{var}(Y_i) = 1$ and $u(\psi, \lambda) = \Sigma[\lambda(y_i - \psi) + (1 - \lambda)\{(y_i - \psi)^2 - 1\}]$. Provided third and fourth moments exist, asymptotic efficiency here is maximized by the choice $\lambda = (2 + \kappa_4)/(2 + \kappa_4 - \kappa_3)$; again $\sqrt{n}$-consistent estimates of κ_3 and κ_4 allow the same first-order efficiency to be achieved. This example is non-robust in the sense that the estimating equation is not generally unbiased under failure of the variance assumption.

Ms S. E. Hills (Nottingham University): I would like to make a practical point concerning the construction of orthogonal parameters. The authors have noted the problem that simple explicit solution of the differential equations for the orthogonal parameters may not be feasible, but there is also the case when an explicit solution is possible but the original nuisance parameters can not be expressed in terms of the orthogonal parameters. An example is the Michaelis-Menton model in nonlinear regression. This model is usually specified as

$$y_i = \frac{\alpha x_i}{\beta + x_i} + \varepsilon_i, \quad (i = 1, \ldots, n)$$

where $\varepsilon_i \sim N(0, \sigma^2)$ (assume σ known).

If β is the parameter of interest, then a transformation of the form $(\beta, \alpha) \to (\beta, \lambda)$ is required so that β and λ are orthogonal. The differential equation to be solved will be

$$\sum_{i=1}^n \frac{x_i^2}{(\beta + x_i)^2}\frac{\partial\alpha}{\partial\beta} = \alpha \sum_{i=1}^n \frac{x_i^2}{(\beta + x_i)^3},$$

with solution

$$a(\lambda) = \alpha^2 \sum_{i=1}^n \frac{x_i^2}{(\beta + x_i)^2}.$$

It is trivial to write α (and hence the likelihood function) in terms of β and λ.

If α is the parameter of interest, then a transformation of the form $(\alpha, \beta) \rightarrow (\alpha, \lambda)$ is required so that α and λ are orthogonal. The differential equation to be solved will be

$$\alpha \sum_{i=1}^{n} \frac{x_i^2}{(\beta + x_i)^4} \frac{\partial \beta}{\partial \alpha} = \sum_{i=1}^{n} \frac{x_i^2}{(\beta + x_i)^3}$$

with solution

$$b(\lambda) = \alpha^3 \sum_{i=1}^{n} \frac{x_i^2}{(\beta + x_i)^3}.$$

The inverse of this transformation is not explicit and therefore it is not possible to write the likelihood in the form $l(\alpha, \lambda)$.

Can the authors give any guidance as to when this type of situation will occur and how to overcome it?

Dr C. J. Skinner (University of Southampton): I should like to comment on the role of the concept of parameter orthogonality in model robustness, with particular reference to the regression example.

Let $M^0 = \{f(y;\, \phi,\, \psi^0);\ \phi \in \Phi\}$ denote a specified model. Then it appears to be of some interest to study 'orthogonal perturbations' of M^0 within broader models of the form $M = \{f(y;\, \phi, \psi);\ \phi \in \Phi, \psi \in \Psi\}$ where $\psi^0 \in \Psi$, ψ is orthogonal to ϕ locally on M^0 and ϕ retains an interpretation in M free of ψ (c.f. 3.5). For if ψ^{T}, indexing the true model assumed to lie in M, is within $0(n^{-1/2})$ of ψ^0 then, as in 2.2, $\hat{\phi}_{\psi^0}$, $\hat{\phi}_{\psi T}$, and $\hat{\phi}$ are all within $O_p(n^{-1})$ and the *MLE* of ϕ within M^0 is, in this sense, robust to local perturbations of M^0 in M.

For example, let M_R be the class of regression models $Y^{\psi} \sim N(\phi_1 + \Sigma\, x_r \phi_r,\, \phi_0)$ in Section 3.5 and let M_R^0 refer to a specific choice $\psi = \psi^0$. Then $\boldsymbol{\phi}^* = (\phi_3/\phi_2,\, \ldots,\, \phi_t/\phi_2)$ has an interpretation free of ψ (increasing x_2 by ϕ_r/ϕ_2 has the same effect on Y as increasing x_r by one unit, whatever the value of ψ) and, being a function of $\lambda_2, \ldots, \lambda_t$ in (5), is approximately orthogonal to ψ. Hence, in the sense above, *MLE* of $\boldsymbol{\phi}^*$ in M_R^0 is robust to local perturbations of M_R^0 in M_R.

This property may be compared with a seemingly stronger result for global perturbatons of M_R^0 within the wider class G_R of generalised linear models, in which Y depends on $\mathbf{x} = (x_2, \ldots, x_t)$ only *via* a linear combination $\phi_1 + \Sigma\, x_r \phi_r = \phi_1 + \mathbf{x}\boldsymbol{\phi}$, and for the wider class of point estimators of $(\phi_1, \boldsymbol{\phi})$ which solve a maximisation problem $\max(a, \mathbf{b})\ \Sigma\, \psi(Y_i,\, a + \mathbf{x}_i \mathbf{b})$, where the function $\psi(.,\, .)$ is essentially arbitrary. Subject to a suitable law of large numbers such estimators converge to $(\tilde{\phi}_1, \tilde{\boldsymbol{\phi}})$, the solution of $\max(a, \mathbf{b})$ $E\psi(Y,\, a + \mathbf{x}\mathbf{b})$ with an implied estimating equation:

$$\operatorname{cov}[\psi_2(Y,\, \tilde{\phi}_1 + \mathbf{x}\tilde{\boldsymbol{\phi}}),\ \mathbf{x}] = 0 \tag{1}$$

where $\psi_2(u, v) = \partial\psi(u, v)/\partial v$. Under G_R, (1) reduces to an equation of form $\operatorname{cov}[h\phi_1 + \mathbf{x}\boldsymbol{\phi},\, \tilde{\phi}_1 + \mathbf{x}\tilde{\boldsymbol{\phi}}),\, \mathbf{x}] = 0$ which, assuming $\|\boldsymbol{\phi}\|/\phi_1$, $\|\boldsymbol{\phi}\|/\tilde{\phi}_1$ small as in 3.5, gives a first-order approximation $\operatorname{cov}[h_1\mathbf{x}\boldsymbol{\phi} + h_2\mathbf{x}\tilde{\boldsymbol{\phi}},\, \mathbf{x}] = 0$ so that $\tilde{\boldsymbol{\phi}} \propto \boldsymbol{\phi}$ if $\operatorname{var}(\mathbf{x}) > 0$ and $\boldsymbol{\phi}^* = \boldsymbol{\phi}^*$. Hence the global robustness property that $\boldsymbol{\phi}^*$ is estimated consistently under misspecification of M_R^0 in G_R. Solomon (1984) gives a special case of this result. The small $\|\boldsymbol{\phi}\|$ condition may be replaced by a condition of elliptical symmetry on the marginal distribution of $\mathbf{x}$ (c.f. Brillinger, 1982; Ruud, 1986).

Mr N. G. Polson (Nottingham University): Tonight we have heard how to make inferences about the parameter of interest, θ, in the presence of a nuisance parameter, λ. The authors propose to use a conditional profile likelihood. We have also heard from Professor Barndorff-Nielsen who advocates the modified profile likelihood.

When the model possesses a group structure, the latter can be represented as a marginal likelihood (as mentioned on p. 10), the measure for $\lambda \mid \theta$ being the induced right invariant Haar measure.

This can be used to unify some of the comments about the Bayesian approach already made by Professor Cox, Professor Barndorff-Nielsen and Dr Sweeting. The Bayesian methodology is totally general—integrate out prior beliefs about $\lambda \mid \theta$. An appealing choice of prior when there are nuisance parameters is a reference prior, as defined by Bernardo (1979). It is used as a reference point for other inferences, also as an approximation to weak *a priori* information about $\lambda \mid \theta$. Applying Bernardo's criterion, orthogonality simplifies the asymptotic posterior for $\lambda \mid \theta$, yielding the result that the above measure is precisely the reference prior for $\lambda \mid \theta$. We therefore have the important theorem that the modified profile likelihood is precisely the Bayesian marginal likelihood with a reference prior for $\lambda \mid \theta$.

The methods of this paper are therefore essentially Bayesian. If the authors were to be consistent and also use a reference prior for θ, they would find complete numerical agreement with reference Bayesian solutions.

This has several important implications. First, we note that such priors avoid the marginalisation paradoxes of Dawid *et al.* (1973). Secondly, all of tonight's examples and previous ones given in the literature permit analytic computations for reference priors, extremely useful for Bayesians. Thirdly, questions raised in Bernardo's paper are also applicable here.

Two examples where the group structure is not present are the hyberboloid model of order 3 and the inverse Gaussian model (Barndorff-Nielsen 1983). I would like to ask the authors how their methods apply here and if there are corresponding links with a Bayesian answer?

Finally, one of the most important applications of nuisance parameters is to model elaboration (for example, the Box-Cox transformation model). The Bayesian framework allows us to view such questions in a unified manner. Do the authors think their methods can be applied in as unified a manner, for example with λ discrete or continuous, as the Bayesian methodology?

Dr Frank Critchley (University of Warwick): My reaction on reading this paper was one of awe and wonder. "Or" because the authors propose w_c^* *or* w_c *or* $\tilde{w}_c$ and wonder because I found myself genuinely wondering: "What does it all add up to?" In particular:

(i) Choice: How are we to choose among the various measures proposed? *Are* they all the same to $O_p(n^{-1})$? If so, may there be important differences in their leading coefficients (this being the basis put forward for preferring $\tilde{w}_c$ to w)? If so, when?

(ii) Practice: What are the relative and, indeed, absolute values of the measures *in practice*? By parameter orthogonality, the asymptotic conditional distribution of $\hat{\theta}_1$ given the observed $\hat{\theta}_2$ is just the asymptotic marginal distribution $N_{p_1}(\theta_1, n^{-1}i_{\theta_1\theta_1}^{-1})$. In going beyond this simple case, the authors appear to be considering sub-asymptotic situations. In any event, this is the common practical situation. The key question here is: Which values of n are sufficiently sub-asymptotic to make the more elaborate procedures worthwhile and yet sufficiently large to retain enough accuracy in the crucial approximation (2) on which rests the key advantage (iv) of parameter orthogonality?

(iii) Operation: How operational is it all? What about vector parameters of interest? How often are the differential equations (4) soluble analytically? When must the invariant w_c^* be abandoned for the more pragmatic w_c or $\tilde{w}_c$? Professor Smith's contribution contains a graphic illustration of the potential losses associated with using these alternative measures.

It would be churlish to not also offer some neutral or positive remarks:

(i) The choice among the measures is indeed a multivariate one. No one measure dominates on all criteria. There are conflicts both between and within matters of principle and matters of practice. Within this latter set, we note the criterion of communicability to the client. Not all of the entries in the criteria by measures array are known (how close is w_c^* to w_k?, . . .). Further work would be valuable here.

(ii) Noting the localness of the approximation (2), might it be worth exploring multi-parameter extensions based on *approximate* global orthogonality in which the (average) size of $i_{\theta_1\theta_2}$ is minimized over some neighbourhood of $\theta = \hat{\theta}$?

(iii) Can the freedom in choosing an orthogonal parameterisation be turned to good effect (e.g. by optimising the accuracy of (2) or the robustness in some sense of the overall procedure)?

(iv) In recently submitted papers, Critchley, Ford and co-workers have shown how strong Lagrangian theory both illuminates the theoretical properties of w and gives substantial computational benefits in calculating interval estimates based on it. It will be of great interest to see how this theory applies to tonight's paper and, in particular, to (10).

(v) There are close links between tonight's paper and the local influence work of Cook (1986).

Answers to any of the above questions would be of value. Without doubt, many of these answers will depend on the context, as with the probable advantage of $\tilde{w}_c$ over w which depends upon both $|a| > |b|$ and $i_{\lambda\lambda}$ being mathematically independent of ψ.

In sum, I found tonight's paper a valuable and thought-provoking contribution to what is one of our subject's major problems. It is, therefore, not surprising that much work remains to be done.

Dr Ann F. S. Mitchell (Imperial College, London): Amari (1982, 1985) produces the orthogonal parameters of Section 3.2. for regular exponential families by a different approach to that of this paper. For parameter space Θ, the family of densities $\{p(y;\,\theta),\,\theta\in\Theta\}$, satisfying the usual regularity conditions, is considered as a manifold in which the parameters $\theta=(\theta_1\theta_2,\ldots,\theta_r)$ play the role of co-ordinates, the information matrix entries $\{g_{ij}(\theta)\}$ form the metric tensor and the connections are the family of α-connections of Amari (1982, 1985). If the manifold is $\pm\alpha_0$-flat for some real α_0, there exist dual co-ordinate systems (θ,η) such that θ_i and η_j are orthogonal for $i\neq j$; $i,j=1,2,\ldots,r$. The dual co-ordinates are related by Legendre transformations

$$\theta_i=\frac{\partial}{\partial\eta_i}\,\phi(\eta),\quad \eta_i=\frac{\partial}{\partial\theta_i}\,\psi(\theta),$$

where the potential functions $\psi(\theta)$ and $\phi(\eta)$ are such that

$$g_{ij}(\theta)=\frac{\partial^2}{\partial\theta_i\partial\theta_j}\,\psi(\theta),\quad g^{ij}(\eta)=\frac{\partial^2}{\partial\eta_i\partial\eta_j}\,\phi(\eta)$$

and

$$\psi(\theta)+\phi(\eta)-\sum_{i=1}^{r}\theta_i\eta_i=0,$$

$\{g^{ij}(\eta)\}$ being the entries of the inverse of the information matrix in terms of the parameters η.

Since regular exponential families are ± 1 flat, the results in Section 3.2 for the case $r=2$ follow at once in general and for the particular case of the normal distribution.

The normal distribution can also be regarded as belonging to an alternative class of distributions, namely the class of elliptic distributions with densities of the form

$$p(y;\,\mu,\,\sigma)=\frac{1}{\sigma}\,h\!\left(\left(\frac{y-\mu}{\sigma}\right)^2\right)$$

for some function h and location and scale parameters, μ and σ ($\sigma>0$), respectively. The class also includes, for example, the Cauchy, Student's t on k d.f. ($k>1$) and the logistic distributions. In the multivariate context it has received much attention in studies of robustness of standard multivariate normal procedures.

The Cauchy distribution has constant negative curvature for all α-values and recent numerical work by Kyriakidis indicates that the logistic is not flat for any value of α. However, when flatness can be demonstrated, the dual co-ordinate systems are

$$\theta=\left(\frac{\mu}{\sigma^2},\,-\frac{1}{2\sigma^2}\right)\quad\text{and}\quad \eta=(a_h\mu,\,a_h\mu^2+b_h\sigma^2),$$

where a_h and b_h are constants depending on the family under consideration. In particular, the Student's t family on k d.f. is $\pm\left(\dfrac{k+5}{k-1}\right)$ flat with

$$a_h=(k+1)/(k+3),\qquad b_h=k/(k+3).$$

Full details of the differential geometry properties of the class of elliptic distributions is given by Mitchell (1986).

The following contributions were received in writing, after the meeting.

Professor Shun-ichi Amari (University of Tokyo): A statistical model $M=\{f_Y(y;\,\psi,\,\phi)\}$ forms a geometric manifold with a coordinate system $(\psi,\,\phi)$ to specify a point (a distribution) in M. When one has interests only in ψ but not in ϕ, the set of distributions $S(\psi_0)=\{f_Y\,|\,\psi=\psi_0;\ \phi\text{: arbitrary}\}$ forms a submanifold embedded in M. In such a case with nuisance parameters, geometry is more explicit in statistical inference, because the shape (more precisely the m- and e-curvatures) of $S(\psi_0)$ plays an important role. The present paper raises an interesting issue relating to the conditionality principle and geometry.

The authors propose new test statistics w_c^* and its simplified version w_c. It is interesting and important to study their characteristics. They are subject to an asymptotic chi-squared distribution, and the tests based on them are first-order efficient, and hence are automatically second-order efficient. The problem is to know the deficiency curve

$$P_T(t) = \lim_{n \to \infty} n[P^*(t) - P_T(t)],$$

of such a test T, where $P^*(t)$ is the envelope power function and $P_T(t)$ is the power function of test T at $\psi = \psi_0 + t/\sqrt{n}$.

Let us consider the problem in a curved exponential family for simplicity's sake. Then, the critical region of the test based on w_c^* (of w_c) is bounded by a hypersurface determined from

$$w_c^* = \text{const.} \quad (w_c = \text{const.})$$

in the enveloping manifold which is identified with the sample space, where the constant is to be determined from the level condition. In the case of a two-sided test, it is bounded by two submanifolds, where we use the signed root of w_c^*. The characteristics of the test depends on the geometric features (angle and curvatures) of a family of submanifolds $w_c^* = \text{const.}$ (Kumon and Amari, 1983; Amari 1985). Here, we should distinguish two problems. One is how to choose the constant. Since w_c^* (or w_c) is not subject to an exact chi-squared distribution, we need to adjust w_c^* (w_c) (or equivalently the constant), such that the level condition (and bias condition in a two-sided case) are satisfied up to the term of order n^{-1}, as we do in the Bartlett adjustment. The other problem is concerned with the deficiency curve of a test after the adjustment has been done. The deficiency curve, when we do not know the true value of ϕ, include two additional terms; one being proportional to the square of the mixture curvature of $S(\psi_0)$ and the other proportional to the square of the exponential curvature of M itself. Although we do not yet know the characteristic of the proposed tests, I believe that the differential geometrical methods developed in Amari and Kumon (1982), Kumon and Amari (1983) and Amari (1985), provide us with sufficient means of analysing these problems.

A final comment is that, even when there does not exist a global orthogonal parametrization, a locally orthogonal parmeter λ, being orthogonal at only $\psi = \psi_0$, may be sufficient for the present asymptotic purpose. Such one is easily derived as

$$\lambda_s = \phi_s + \sum_{k,r} (i^*)^{-1}_{\phi_r \phi_s} (i^*_{\psi_k \phi_r})(\psi_k - \psi_{0k}).$$

We can add quadratic terms $(\psi - \psi_0)^2$ such that not only the cross terms of the Fisher information but also its derivatives with respect to ψ vanish at ψ_0.

Professor A. C. Atkinson (Imperial College, London: A major part of my interest in the work described in tonight's paper centres on the normal transformation model which Box and Cox (1964) write $y^{(\lambda)} = (y^\lambda - 1)/\lambda$. With this background it is a nuisance that Cox and Reid choose λ to be the nuisance parameter.

1. The numerical example in Section 3.5 demonstrates the advantage of the orthogonal parameterization compared with the form investigated by Bickel and Doksum (1981). However, Box and Cox introduced the normalized transformation $z^{(\lambda)} = y^{(\lambda)}/\dot{y}^{\lambda-1}$, where $\dot{y}$ is the geometric mean of the observations. An appealing property of this transformation is that the dimensions of $z^{(\lambda)}$ is that of y. The resulting parametrization is approximately orthogonal. Are there other examples where physical arguments lead to a near orthogonal parametrization?

2. The profile loglikelihood for the factorial experiment of Section 5.2 is pleasingly parabolic, as are those for several other examples plotted by Cook and Weisberg (1982, Section 2.4). Asymptotic procedures can then be expected to behave well. The other example given by Box and Cox, a factorial experiment on the failure of worsted yarn, however yields a profile loglikelihood which is concave around $\lambda = 0$ but convex near $\lambda = \pm 1$ (Atkinson, 1985, Fig. 6.2). What results are available on the concavity of profile loglikelihoods? Do the corrections of Section 4.3 improve this curve?

3. There is a striking similarity between (26) and the extremely useful result of Patefield (1977), obtained without the use of parameter orthogonality. Lawrance (1987) uses Patefield's result to obtain a score statistic for transformations which has good distributional properties in the neighbourhood of the null hypothesis. Since this result uses observed information, rather than the expected information of (26), it gives a negative variance when $\lambda_0 = -1$ in the worsted data.

Mr B. J. R. Bailey (University of Southampton): I should like to compliment the authors on their generous provision of examples and here add yet another, but one based on discrete variates. Suppose X and Y have independent binomial distributions such that $X \sim b(m, \theta_1)$ and $Y \sim b(n, \theta_2)$. If the parameter of interest is the odds ratio $\psi = \theta_1(1-\theta_2)/(1-\theta_1)\theta_2$, then the application of equation (4) leads to the orthogonal parameter $a(\lambda) = m\theta_1 + n\theta_2$. Setting $a(\lambda) = \lambda$, and maximizing the likelihood over λ, for fixed ψ, yields the conditioning statistic $\hat{\lambda}_\psi = x + y$, the usual ancillary statistic for this problem.

On the other hand, in epidemiology, the parameter of particular interest is the risk ratio $\psi = \theta_1/\theta_2$. Orthogonal to this is $\lambda = (1-\theta_1)^m(1-\theta_2)^n$, or any function of λ such as the logarithm, and $\hat{\lambda}_\psi$ can be found explicitly as a slightly cumbersome function of ψ. However, conditioning on $\hat{\lambda}_\psi$ is extremely severe in that the possible values of the pair (x, y) generally lead to different values of $\hat{\lambda}_\psi$, for fixed ψ. This is, of course, a problem likely to arise in many discrete cases. Grouping values of $\hat{\lambda}_\psi$, and then conditioning on the particular group observed, does not seem to be practical if this has to be done for several values of ψ and for values of m and n typically in the range 100–200. Is there an easier alternative?

Professor G. A. Barnard (Retired): There can be no general method for "elimination" of "nuisance parameters". The very term harks back to the idea that statistical inference involves an act of will. In a decision problem the form of answer may be governed by our wishes. But the inferences which may be drawn from a data-model combination must be dictated by the data available, along with the logical features of this combination. We may well wish to infer something about μ without reference to λ, but the data may not permit this. And if the data do not permit it, we owe it to our clients to say so.

Logical features may smplify a problem. For example we may be interested in the correlation between scores in a verbal test and in a mathematical test. Such scores are often reasonably taken to be bivariate normal, but the means and variances of each score are clearly affected by irrelevant factors, so that analysis of the data must logically be invariant under location and scale changes. Thus invariance considerations lead directly to the sample correlation coefficient as the only quantity of interest, with its marginal distribution providing the relevant likelihood function. Similar invariance features are relevant to the Neyman-Scott problem referred to by the authors, and to other cases.

In location-scale problems, conditioning on the configuration allows us to reduce the data to two pivotals, $t = (\bar{x} - \mu)/s_x$ and $z = s_x/\sigma$, with known joint density $\phi(t, z)$. The failure of $\phi(t, z)$ to factorize means that inference about μ using the marginal distribution of t is subject to the implicit assumption, often overlooked, that the data provide the only usable information about σ. By suitable choice of μ (noting thc brief hint on p. 339 of Fisher's 1922 *Phil.Trans.* paper), μ and σ can be made orthogonal, so that possession of a little information concerning σ, in addition to the data, does not seriously affect inferences about μ. But the case is quite otherwise with the Behrens-Fisher and weighted mean problems. Here the variance ratio parameter ρ should perhaps be called a "confounded nuisance parameter", since inferences about the parameter of interest cannot be separated from statements about ρ. Insistence on rigour requires making inferences conditional on ρ; but it will often be justifiable to introduce a range of priors for ρ. Then so long as we make clear to our clients the assumptions involved, and so long as the problems have enough of a routine character to allow some check on the priors used, inferences in which ρ does not occur explicitly will be permissible.

The "top down" approach of the authors, working down from the asymptotic case, will be very useful in complex cases as indicating the kinds of assumption needed to make inferences of the form required. But an extension to several parameters of the "bottom up" approach of Sprott and Viveros (1984), who attempt to match the log-likelihood function up to the fourth term of its Taylor expansion, would also seem worth exploring.

Dr A. C. Davison (Imperial College, London): Professors Cox and Reid refer in the final section of their thought-provoking paper to the prediction of future observations. I would like to point out a curious similarity between the modified profile log likelihood (10) and recent work on predictive likelihood.

Suppose that the random variable $Y = y$, with probability density function $f(y \mid \theta)$ has been observed, and that the unobserved random variable Z with conditional density $f(z \mid y, \theta)$ is to be predicted. The parameter θ is unknown. In Davison (1986, equation 6) I suggest as an approximate predictive likelihood for the predictand the exponent of

$$\log f(z, y \mid \hat{\theta}_z) - \tfrac{1}{2} \log \det j_{\theta\theta}(\hat{\theta}_z), \qquad (*)$$

regarded as a function of z. Here $\hat{\theta}_z$ and $j_{\theta\theta}$ are the maximum likelihood estimate of θ and observed information based on both y and z. Expression (*) looks very like the modified profile log likelihood (10) with z and θ replacing the ψ and λ of Cox and Reid—though of course the logical status of the unknown value z of the random variable Z differs from that of the unknown but fixed parameter ψ.

In replying to the discussion of his paper, Butler (1986a) shows how to make (*) invariant to reparametrization of θ by adding to it

$$\log \det j_{\theta\theta}(\hat{\theta}_z) - \tfrac{1}{2} \log \det KK', \tag{†}$$

where $K = d^2 \log f(z, y\,|\,\theta)/d\theta d(y, z)$, evaluated at $\theta = \hat{\theta}_z$.

The first term of (*) is the profile predictive log likelihood suggested by Mathiasen (1979) and Lejeune and Faulkenbery (1982). In many situations the relative extra contributions to the predictive log likelihood from the second term of (*) and from (†) are small. Fig. D2 shows such a case, comparing the 3 terms for the prediction of the maximum of $m = 10$ annual maximum daily river flows, based on a sample of 35 such flows of the River Nidd at Hunsingore Weir. The model is that the annual maxima are a sequence of independent observations with a common generalized extreme-value distribution. The modifications to the profile predictive likelihood from (†) and the second term of (*) are in this case negligible, though as m increases so does the effect of (†).

As far as I know it it not yet clear how in general to base predictive confidence regions for Z on (*), though some progress in this direction has recently been made by Butler (1986b). Perhaps soon a first- and second-order asymptotic theory for prediction, as well as estimation, will be available for the likelihood function.

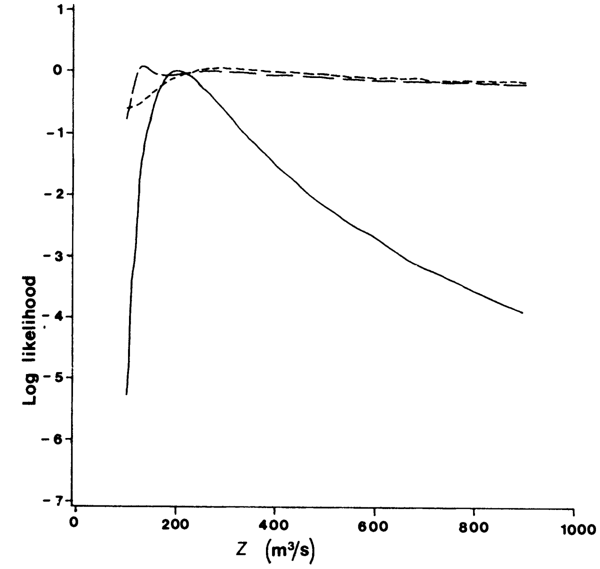

Fig. D2 Information comparison for River Nidd data, $m = 10$. Shown are profile predictive log likelihood (solid line), information matrix contribution $-\frac{1}{2} \log \det j_{\theta\theta}(\hat{\theta}_z)$ (small dashes), and Jacobian contribution $\log \det j_{\theta\theta}(\hat{\theta}_z) - \frac{1}{2}$ log det KK' (longer dashes).

Professor D. A. S. Fraser (York University; Universities of Toronto and Waterloo): Conditional inference, introduced by Fisher, generally neglected, but nurtured tenuously through connections to fiducial, ancillarity, and structural, is now receiving the attention it has seemingly long deserved and the present paper is a welcome and thorough examination of aspects of the topic.

The first three sections propose the information orthogonalization of nuisance parameters to a primary real parameter in order to obtain asymptotic independence for the corresponding m.l.e. estimates; this leads to the analysis of the primary parameter conditional on estimates of the nuisance parameters. Unfortunately, the authors do not directly pursue such conditional inference, which involves a real variable and real parameter with some minimum effect from nuisance parameters. Such inference on-the-real line is direct and straight forward leading to tests and confidence regions, and a likelihood criterion is not needed.

Some recent work on conical tests (Massam and Fraser, 1985; Skovgaard, 1986) and on fibre analysis (Fraser, 1986) lead (in joint work with a Toronto colleague) to a sample space development of one-dimensional conditional tests; these seem to show agreement with the orthogonal-parameter approach when it is available. The majority of the examples in Section 3 are location/transformation models; some compounding of conditional distributions shows promise for further reducing the effect of nuisance parameters.

Sections 4 and 5 develop modifications to profile likelihood to obtain a likelihood assessment of parameter values, but do not provide conditional tests or confidence regions in any direct sense: the 'conditional inference' in the title of the paper might reasonably be changed to 'conditional likelihood'. The modifications to profile likelihood represent an insightful use of conditional distributions to address the difficulties found with profile likelihood itself.

The vectors for the regression model as given in Section 3.5 are of length n which indicates a modification to some formulas. The log-likelihood ratio statistic is essentially a *negative* of log likelihood; thus in several places 'conditional (profile) likelihood' needs to have 'ratio statistic' added to be correct and not misleading.

Dr P. Harris (Liverpool Polytechnic): I have enjoyed reading this paper, and would like to make two brief comments. The first concerns discussion point (i) of Section 6 of the paper, namely the possibility that a Bartlett adjustment factor might exist for the test statistics introduced in Section 4. In particular consider the test statistic, $\hat{w}_c(\psi^0)$ say, given at (11)

$$\hat{w}_c(\psi^0) = w(\psi^0) + S,$$

where $w(\psi^0)$ is given at (6) and $S = \log \det\{j_{\lambda\lambda}(\psi^0, \hat{\lambda}_0)\} - \log \det\{j_{\lambda\lambda}(\hat{\psi}, \hat{\lambda})\}$.

For convenience let λ be a scalar parameter, and expand S about the true parameter value (ψ^0, λ) to give

$$S = (j_{\lambda\lambda})^{-1} y_{\lambda\lambda\psi} u_\psi + (j_{\lambda\lambda\lambda})^{-1} y_{\lambda\lambda} u_\lambda + Y + O_p(n^{-3/2})$$

where $u_\psi = \sqrt{n}(\hat{\psi} - \psi)$, $u_\lambda = \sqrt{n}(\hat{\lambda} - \hat{\lambda}_0)$, $y_{\psi\lambda\lambda} = n^{-3/2}\partial^3 l/(\partial\psi\partial\lambda^2)$, $y_{\lambda\lambda\lambda} = n^{-3/2}\partial^3 l/\partial\lambda^3$ and Y denotes all of the $O_p(n^{-1})$ terms arising in the expansion.

The term $v = (j_{\lambda\lambda})^{-1} y_{\lambda\lambda\psi} u_\psi$, which remains $O_p(n^{-1/2})$ when ψ and λ are orthogonal, appears to introduce into the $O(n^{-1})$ part of the null distribution of $\hat{w}_c(\psi^0)$, a quantity which not only prevents the calculation of a Bartlett factor, but also prevents the $O(n^{-1})$ terms in the null distribution of $\hat{w}_c(\psi^0)$ being expressed as sums of chi-squared variables. The difficulty arises because the $O_p(n^{-1/2})$ part of the expansion of S contains a term linear in $u_i (i = \psi$ or $\lambda)$, rather than the more usual $u_i u_j u_k$ $(i, j, k = \psi$ or $\lambda)$.

Assuming the orthogonality of ψ and λ, the moment generating function of $\hat{w}_c(\psi^0)$ has the form $M(t) = (1-2t)^{-1/2}\{1 + (24n)^{-1}(\phi P + t\phi Q)\}$ where $Q = 6(i_{\lambda\lambda\psi})^2 i_{\lambda\lambda}^{-2} i_{\psi\psi}^{-1}$, $i_{\lambda\lambda\psi} = \sqrt{n}E(y_{\lambda\lambda\psi})$, $\phi = 2t(1-2t)^{-1}$ and P is a complicated function of the cumulants of the derivatives of the log likelihood function. The term Q prevents the calculation of a Bartlett adjustment factor, ρ; if $Q = 0$ then $\rho = 1 + (12n)^{-1}P$. For Q to be zero we need $i_{\lambda\lambda\psi} = 0$, so that the array of expected values of the third derivatives of the log likelihood function needs to satisfy an orthogonality condition for a Bartlett factor to be available.

My second comment is that if ψ is orthogonal to λ, then in testing $H_0 : \psi = \psi^0$

$$w_A(\psi^0) = 2\{l(\hat{\psi}, \hat{\lambda}) - l(\psi^0, \hat{\lambda})\} \tag{1}$$

has the asymptotic chi-squared distribution with one degree of freedom when the null hypothesis is true. Have the authors any comments upon the use of (1), or a conditional version based upon the conditional distributions of y given $\hat{\lambda}$, in testing $\psi = \psi^0$?

As the restricted maximum likelihood estimator $\hat{\lambda}_0$ is not used in either test statistic they may be convenient in situations where $\hat{\lambda}_0$ is awkward to estimate.

Dr C. J. Lloyd (University of Melbourne): Barndorff-Nielsen's (1983) modified profile likelihood

$$L^0(\psi) = \frac{\partial\hat{\lambda}}{\partial\hat{\lambda}_\psi} \,|\hat{j}_{\lambda\lambda}|\,\tilde{L}(\psi)$$

has the nice property that it can be seen as an approximation to either a suitable conditional or marginal likelihood function. In particular, if we condition on $\hat{\lambda}_\psi$ then the approximate conditional likelihood is $L^0(\psi)$ as discussed in 4.1. The orthogonal parametrisation corresponds to choosing λ so that $\partial\hat{\lambda}/\partial\hat{\lambda}_\psi = 1 + 0(n^{-1})$ and $j_{\lambda\lambda}$ is the λ information either for fixed ψ or ignoring ψ. This alone would seem to justify the idea of orthogonality for giving the profile likelihood a standard form. It would be nice to look at the conditional profile likelihood from the dual/marginal point of view. The most likely course is to let $Z(\hat{\psi}, \hat{\lambda}_{\psi}, \psi, \lambda) = F(\hat{\psi} \mid \hat{\lambda}_{\psi}; \psi, \lambda)$ where F is a conditional distribution function and substitute $\hat{\lambda}_{\psi}$ for λ and consider the marginal likelihood of Z.

The distribution of $\hat{\psi}$ given $\hat{\lambda}_{\psi}$ does not generally lead to an unbiased score function as pointed out by Lindsay (1983). The bias term is in fact $\log(\partial\hat{\lambda}/\partial\hat{\lambda}_{\psi})$ so that bias is reduced when ψ, λ are orthogonal. (This term is omitted in w_c^*). There is also a problem of ambiguous inference in choosing the free statistic to be conditioned on. For example if (X, Y) are independent normal with means $(\lambda \cos \psi, \lambda \sin \psi)$ and variances 1 then $\hat{\lambda}(\psi) = \bar{X} \cos \psi + \bar{Y} \sin \psi$ and $Z(\psi) = \bar{X} \sin \psi - \bar{Y} \cos \psi$ with $N(0, 1/n)$ distribution. Now λ and ψ are orthogonal and it turns out

$$w_c(\psi) = -nZ(\psi)^2$$

so the estimating equation is $Z(\psi) = 0$ which is unbiased. The density of $\bar{X}$ given $\hat{\lambda}(\psi)$ is $N(\hat{\lambda}(\psi) \cos \psi, n^{-1} \sin^2\psi)$ giving the conditional loglikelihood

$$-2 \log|\sin \psi| - nZ(\psi)^2$$

so the estimating equation is

$$2\partial \log|\sin \psi|/\partial\psi + nZ(\psi)\hat{\lambda}(\psi) = 0$$

which is not unbiased. Also, if we use the density of $\bar{Y}$ given $\hat{\lambda}(\psi)$ a different likelihood results. The distribution of $\hat{\psi}$ given $\hat{\lambda}(\psi)$ leads to a conditional likelihood which differs from w_c by the term $\cos(\psi - \hat{\psi})$ which is the $0(n^{-1})$ term $\log(\partial\hat{\lambda}(\psi)/\partial\hat{\lambda})$. Finally, an unbiased equation can be obtained by substituting for $\hat{\lambda}(\psi)$ *after* differentiation of the conditional likelihood and this gives

$$-2\tfrac{\partial}{\partial\psi} \log|\sin \psi|[Z^2(\psi) - 1] - nZ(\psi)\hat{\lambda}(\psi)$$

in the present example. This procedure always gives an unbiased equation under regularity, however it is not clear whether its solution corresponds to maximising any sensible objective function.

Professor T. A. Louis (Harvard School of Public Health): Whoever invented the label "nuisance parameter" was right on the mark. Parameters that are not of direct interest plague the analyst, whatever approach is taken. The frequentist needs to condition, profile, or ignore; the Bayesian requires priors that may be difficult to pin down. Cox and Reid have helped expose these difficulties and the simplification provided by parameter orthogonality. For example, their analysis of the normal transformation model helps clarify the controversy concerning the decision to incorporate or ignore the uncertainty in estimating ψ (the transformation parameter). I read their message as suggesting that any reasonable approach will work just fine if the parameters of interest are orthogonal to ψ. Individuals may have philosophical arguments, but their inferences will be similar.

Divorced from computational issues, these results imply that putting weak (ignorance) priors on nuisance parameters, and marginalizing should produce reasonable frequentist inferences, when (approximate) orthogonality holds. When it does not hold, uncertainty due to the nuisance parameters gets incorporated but results depend to a greater degree on the prior. Generally, the applied context dictates parameters of interest, though (approximate) orthogonality may have a role in teaching us how to think about the application, much as do natural parameters in exponential families. This viewpoint makes the vector parameter case similar to the scalar case. Either approximate orthogonality holds or inferences are more difficult.

The paper succeeds at identifying an important research agenda. Section 4.1 presents the technical development underlying a radical form of double conditioning that should generate research intensity similar to that following Cox's 1972 approach to survival analysis. The relation among profile, conditional, marginal, partial, and canonical likelihoods begs further study, as does the success of the new approach in incorporating the effects of vector nuisance parameters, and analysing parametric empirical Bayes models. Though this report is unlikely to have immediate impact on statistical practice, it adds to our understanding of the issues and approaches for inference in the presence of nuisance parameters. Generated research should have a large impact, and the authors are to be congratulated.

Dr J. N. S. Matthews (University of Oxford): I would like to raise a couple of issues from an area where the methods of this paper may find application.

In the analysis of continuous data from crossover trials (n subjects, p periods with p typically between 3 and about 10), a model that is frequently used is:

Here $X\beta$ includes period and treatment effects and, crucially, a parameter for each subject. Moreover it is usually assumed that $\varepsilon \sim N(0, \sigma^2 I_{np})$. However, it is often felt that this could lead to an inefficient analysis as there is likely to be within-subject dependence in the error term, making

$$\text{var}(\varepsilon) = \sigma^2 I_n \oplus V(\rho_1, \ldots, \rho_k)$$

more realistic.

As was pointed out in the paper, β and $(\sigma^2, \rho_1, \ldots, \rho_k)$ are orthogonal. This gives some comfort to users of the simpler model, as it means that changes in the specification of the dispersion matrix will have little effect on estimates of treatment contrasts. However we need estimates of the standard errors of these contrasts; can the authors clarify how much comfort they can offer on this point?

The second issue concerns the estimation of the parameters in the dispersion matrix, in particular when $k = 1$ and $V = V(\rho)$ corresponds to a stationary first-order autoregressive process.

The need to estimate subject effects leads profile likelihood methods astray, giving badly biassed estimates of ρ. Patterson and Thompson (1971, 1974) overcame a similar problem in the estimation of variance components by using their method of restricted maximum likelihood. Applying this type of approach to the problem of estimating ρ gives some interesting results, but there is room for considerable improvement, especially for larger positive values of the true autocorrelation coefficient and smaller values of p.

As the root of the problem is the presence of so many nuisance parameters, it seems possible that conditional profile likelihood methods may be able to contribute to this problem; I would be interested in the authors' views.

Dr P. McCullagh (University of Chicago): This is a comprehensive and detailed paper that deserves careful study. I have only two comments to make at this stage, both very brief.

First, the orthogonality of the canonical parameter and the complementary expectation parameter in exponential families was demonstrated by Huzurbazar (1956), at least for two-parameter families.

Second, I'd be grateful if the authors would elaborate on the non-invariance of (8) under transformation of λ. How much leeway for transformation does orthogonality permit?

Professor Donald A. Pierce (Oregon State University): There is one general point, which though simple, may have some significant and practical bearing on the issues here. This has to do with extensions where the parameter of interest ψ is of more than one dimension and includes comparative effects. For example, consider the Weibull setting of Section 3.4, extended to the case of samples from several Weibull populations with the same shape parameter. It is easily seen that the vector of parameters consisting of *ratios* of the various scale parameters is orthogonal to the shape parameter. This is an instance of a more general result on location-scale parameter models, and those such as the Weibull which may transformed to such. By the method of Section 2.3 a conventional location parameter ψ of any location-scale family can be replaced by a certain quantile, say $\psi + k\sigma$, which is orthogonal to the scale parameter σ. But then for a multi-sample problem with common scale parameter, differences of these new quantiles are the same as differences of the original ones.

This remains true, I believe, if there are additional nuisance parameters defining the "shape" of the location-scale family. That is, the vector of differences of location parameters is orthogonal to both the scale and shape parameters. The result also extends to more general regression models in that if the location parameters are modelled as $\psi_i = \mu + z_i'\beta$, where $\Sigma z_i = 0$, then the vector β is orthogonal to the scale and shape parameters.

I suspect that these observations, which have been made in special cases before, if not more generally, are of practical importance in the general consideration of orthogonality. It would be helpful for the authors to comment on whether they might be of any particular theoretical interest in the relation to their interesting results on conditional inference.

Mr G. J. S. Ross (Rothamsted Experimental Station): In advocating parameter transformations I have regarded orthogonality in itself as being of minor importance. The shape of the log likelihood function may deviate dramatically from the quadratic approximation on which many optimisation procedures depend, whereas these procedures compute their own local orthogonalisations. For a Normal sample in the space of (μ, σ^2) likelihood contours are rounded triangles: a cube root transformation on σ^2 creates symmetry if $\mu = \bar{x}$ but not elsewhere; a better transformation is based on the 8th and 92nd percentiles which are also orthogonal. For Negative Binomial samples the parameters μ and κ are orthogonal but contours are extremely skew with respect to the *MLE*: a reciprocal transformation of κ is a great improvement (Ross and Preece, 1985).

In anticipating suitable transformations to improve numerical estimability a concept of 'unrelatedness' rather than of orthogonality is invoked. True orthogonality may depend on both the model and the data, and cannot be achieved until it is too late to be of use, after the *MLE* has been found. Unrelatedness is like property (iv) of (2.2): a qualitative description of the expected shape of the likelihood function because there is no reason why the estimate of ψ should be seriously affected by the estimate of λ. With three or more parameters it is essential to think in this way because graphical aids are of little use. In fitting non-linear curves use can be made of widely spaced points on the curve representing the local position of the curve: they may not be perfectly orthogonal but they are a great improvement on the algebraic defining parameters. In fitting a mixture of two Normal distributions with equal variances but unknown means and proportions we can anticipate that a set of unrelated parameters would take account of (i) the general location of the data, (ii) the overall spread, (iii) the asymmetry relative to a single Normal, and (iv) the separation of the modes. Within the qualitative framework the actual parameters chosen are those that lead to tractable algebraic or algorithmic procedures. (Ross 1970, 1975).

The property (ii) of (2.2) is related to measures of ill-conditioning of the dispersion matrix which are more helpful than a simple inspection of correlation coefficients but less complicated than an eigenvector analysis. The product of $i^*_{\phi_r\phi_r}$ with the corresponding element of the dispersion matrix gives an absolute quantity which is 1 for orthogonal parameters and infinity for totally dependent parameters. It is the ratio by which the variance of ϕ_r would be reduced if the values of the other parameters were known and is thus a very important diagnostic quantity. Provisionally I call this a 'variance multiplier'.

Dr D. A. Sprott (University of Waterloo): It is worth mentioning that Fisher (1922) defined the *centre of location* of a location-scale (θ, σ) family to be $\phi = \theta + k\sigma$ such that (1) is satisfied by (ϕ, σ).

Fisher (1961a, 1961b) also presented an "exact" solution to the weighted means problem for $q = 2$ samples. In this solution, the factor $n_j - 2$ in the expression for $w_c(\mu^0)$ is replaced by n_j. The likelihood produced by Kalbfleisch and Sprott (1970) has $n_j - 1$. Thus both of these solutions avoid the logical difficulty of being unable to cope with samples of size $n_j = 2$. It would be therefore of some interest to compare the frequency properties of these three solutions.

The solution involving $n_j - 2$ seems to have gained support because of Neyman and Scott's (1948) demonstration that it is more "efficient". However, their definition of the efficiency of an estimate $\hat{\mu}$ is in terms of its asymptotic variance $\sigma^2_{\hat{\mu}}$, a function of the τ_j's. This is relevant only for asymptotic $N(0, 1)$ pivotals $(\hat{\mu} - \mu)/\sigma_{\hat{\mu}}$, which, being functions of the τ_j's, are appropriate for estimating μ only when the τ_j's are *known*. Such pivotals, and their efficiencies, would seem irrelevant for estimating μ when the τ_j's are *unknown*, as in the weighted means problem. Thus the problem of assessing the behaviour of various solutions to this problem still seems open.

Finally, Viveros (1985) and Viveros and Sprott (1986) have used a different approach based on approximating the observed family of log likelihoods, up to the quartic term of their Taylor expansion, by simple functions like t or log F. This results in pivotals like $(\hat{\theta} - \theta)I_\theta^{1/2}$ having approximate log F distributions, where I_θ is the observed Fisher information. This has produced very accurate results in small samples. In fact, in a location-scale model similar to that of Section 4.2.2, the approximating log F distribution of one of the pivotals is graphically indistinguishable from its exact conditional distribution.

Mr Jonathan Tawn (University of Surrey): I would like to congratulate the authors on this interesting and motivating paper. The aspect of the paper which has particularly interested me is the use of orthogonal parameters.

The authors suggest that we orthogonalise the nuisance parameters to the parameter of interest by using global orthogonality with respect to expected Fisher informaton. The key property of this orthogonality for conditional inference being

$$\hat{\psi}_\lambda - \hat{\psi} = O_p\left(\frac{1}{n}\right) \quad \text{if} \quad \hat{\lambda} - \lambda = O_p\left(\frac{1}{\sqrt{n}}\right). \qquad (*)$$

The obvious practical problem with this form of orthogonality being equations (4) are often impossible to solve in closed form. This suggests that maybe we should look for another concept of reparametrization which has property (*).

An interesting example which motivates the rest of my discussion arises in the non-regular estimation problem when the parameter of interest is the endpoint of a distribution. Suppose that the probability density $f(x; \psi, \lambda)$ is the form

$$f(x; \psi, \lambda) = \begin{cases} \alpha c(x - \psi)^{\alpha - 1} & \text{as } x \downarrow \psi, \quad 1 < \alpha < 2 \\ 0 & \text{if } x < \psi. \end{cases}$$

Here the endpoint of the distribution exhibits 'orthogonality' properties, namely

$$\hat{\psi}_\lambda - \hat{\psi} = O_p(n^{-1/\alpha})$$

and $\hat{\psi}$ and $\hat{\lambda}$ are asymptotically independent (Smith, 1985). This situation is not unique, similar orthogonality is obtained in bivariate extreme value theory involving a parameter on the boundary of the parameter space. In each case no concept of orthogonality in the expected information is possible as this is infinite.

What the examples suggest is that we work with orthogonality of the observed information, leading to data dependent parametrization. If we had global orthogonality of the observed information then

$$\hat{\psi}_\lambda = \hat{\psi}$$

unfortunately such a situation can rarely be achieved. One particular case in which this can be achieved is example 3.1 in the paper.

Here $$\phi = \lambda_n\left(\bar{Y}_1 + \frac{\bar{Y}_2}{\psi}\right), \quad \hat{\lambda}_n = \tfrac{1}{2}, \quad \hat{\psi} = \frac{\bar{Y}_2}{\bar{Y}_1}.$$

As λ_n does not tend to the authors parametrization this leads us to question the optimality of their parametrization.

I suggest the use of local orthogonality of the observed information at $(\hat{\psi}, \hat{\lambda})$. Under certain conditions property (*) still holds, hence we obtain an equivalent of (4)

$$\sum_r j^*_{\phi_r\phi_s}(\hat{\psi}, \hat{\phi}) \left.\frac{\partial\phi_r}{\partial\psi}\right|_{(\hat{\psi},\hat{\phi})} = -j^*_{\psi\phi_s}(\hat{\psi}, \hat{\phi}) \; s = 1, \ldots, q.$$

We illustrate the flexibility of the solution is the two parameter case

$$\left.\frac{\partial\phi}{\partial\psi}\right|_{(\hat{\psi},\hat{\phi})} = \frac{-j^*_{\psi\phi}(\hat{\psi}, \hat{\phi})}{j^*_{\phi\phi}(\hat{\psi}, \hat{\phi})} = s(\mathbf{Y}) = \frac{\hat{\phi}^\alpha}{\hat{\psi}^\beta} S^*(\mathbf{Y}) \quad \text{say.}$$

Hence, if $\alpha \neq 1$, $\beta \neq 1$.

$$\frac{1}{(1-\alpha)} \phi^{1-\alpha} = \frac{1}{1-\beta} \psi^{1-\beta} S^*(\mathbf{Y}) + \lambda_n. \quad \text{say.}$$

where α and β can be chosen to make the parametrization a valid one, and to satisfy the conditions on λ.

With such flexibility in the solution can other conditions be imposed on the parametrization to make it more optimal?

Drs S. H. Moolgavkar and R. L. Prentice (Fred Hutchinson Cancer Research Centre, Washington State, USA): The problem of parameter orthogonalization has a general solution provided by the theorem of Frobenius in differential geometry (Boothby, 1975, page 159). Let (ψ, ϕ) be a parametrization with ψ the k-dimensional parameter of interest, ϕ a $(n-k)$-dimensional nuisance parameter.

Then, it follows from Frobenius' theorem that a necessary and sufficient condition for the existence of a parametrization (ψ, λ) (with ψ orthogonal to λ is that the (vector) space of all vector fields orthogonal to ϕ (with respect to the Fisher information metric) be a Lie algebra. When ψ is one-dimensional this condition is trivially satisfied. As an example, consider an exponential family with density $f(y, \theta) = g(y) \exp\{y \cdot \theta - K(\theta)\}$, and suppose $\psi = (\theta_1, \theta_2, \ldots, \theta_k)$, $\phi = (\theta_{k+1}, \ldots, \theta_n)$. Then, for any (coordinate) tangent vector $\partial/\partial\theta_i$ and any arbitrary vector field X, the inner product with respect to the Fisher information metric, $\langle X, \partial/\partial\theta_i\rangle$, is given by $\langle X, \partial/\partial\theta_i\rangle = \langle \partial/\partial\theta_i, X\rangle = X\, \partial/\partial\theta_i(K(\theta))$. Now consider vector fields X_1, X_2 such that $\langle X_1, \partial/\partial\theta_i\rangle = \langle X_2, \partial/\partial\theta_i\rangle = 0$ for $i = k+1, \ldots, n$. Then $\langle [X_1, X_2], \partial/\partial\theta_i\rangle = \langle (X_1X_2 - X_2X_1), \partial/\partial\theta_i\rangle = X_1(X_2\partial/\partial\theta_iK) - X_2(X_1\partial/\partial\theta_iK) = 0$. Thus the vector fields orthogonal to ϕ form a Lie algebra and by Frobenius' theorem there exists a parametrization (ψ, λ) with ψ orthogonal to λ, which is a well known result.

In general, of course, in an orthogonal parametrization (ψ, λ), the Fisher information for ψ will depend upon λ. The theorem of de Rham (Kobayashi and Nomizu, 1963, page 187) provides necessary and sufficient conditions for the Fisher information for each parameter to be independent of the other parameter.

The asymptotic distribution theory for ordinary profile likelihood procedures can be thought of as deriving from the asymptotic marginal distribution of $\hat{\psi}$. Inference procedures deriving instead from the asymptotic distribution of $\hat{\psi}$ given $\hat{\lambda}$ may provide more accurate approximations in moderate-sized samples since such conditioning makes an accommodation for the difference between the estimated and the true λ-values. Orthogonality is a natural requrement in this setting in order to minimize any loss of information on ψ. Conditioning on $\hat{\lambda}$ leads to a likelihood ratio statistic which differs from Cox and Reid's expression (8) only in the final term; that is,

$$\omega_{\hat{\Lambda}}(\psi^0) = 2[l_Y(\hat{\psi}, \hat{\lambda}) - l_{\hat{\Lambda}}(\hat{\psi}, \hat{\lambda}) - \{l_Y(\psi^0, \hat{\lambda}_0) - l_{\hat{\Lambda}}(\psi^0, \hat{\lambda}_0)\}].$$

This statistic behaves like (8) for Cox and Reid's examples, but is also invariant under transformations of λ.

Another approach would be to improve the approximation to the marginal distribution of $\hat{\psi}$. For example, suppose that the score $S(\psi_0) = \partial l_Y(\psi_0, \hat{\lambda}_0)/\partial\psi_0$ has mean $E_0 = E(\psi_0)$ and variance $V_0 = V(\psi_0)$ at $\lambda = \hat{\lambda}_0$. One can then readily obtain an asymptotic χ^2_k distribution for

$$\omega_M(\psi^0) = 2\{l_Y(\hat{\psi}, \hat{\lambda}) - l_Y(\psi^0, \hat{\lambda}_0)\} + E_0^t V_0^{-1} E_0.$$

Calculation of, or approximation to, E_0 and V_0 leads to another possibility for adjusting profile likelihood ratio tests.

The **authors** replied briefly at the meeting and subsequently more fully in writing as follows:

We are very grateful to all those who took part in the discussion for their thoughtful and wide-ranging comments.

Orthogonality of parameters is a major theme of the paper and the discussion contains many valuable comments on this. As we stressed in Section 2.2 a number of distinct, if interrelated, ideas are involved.

Global orthogonality is helpful for interpreting the model, and in particular for studying the important topic of model robustness (Sweeting, Skinner, Atkinson, Barnard, Pierce). For example, the orthogonalized expression of the Behrens-Fisher model clearly indicates that the variance ratio is what Professor Barnard calls a confounded nuisance parameter. We agree with Mr Ross's valuable points and particularly that orthogonality on its own may not provide definitive solutions.

We agree with Professor Atkinson that ultimately physical interpretability is more important and that from this point of view approximate orthogonality may often be enough. But from the point of view of interpreting models, data dependent and local definitions are less than ideal, which is one reason for preferring our formulation of the transformation problem to that based on a data dependent scale of measurement. It is clear from Mr Tawn's discussion of Example 3.1 that either observed or expected information orthogonality is more important than property (iv).

As suggested by Professor Amari, for the more technical details involved in deriving "improved" inference for ψ it is likely that local orthogonality is enough, and this may well be the basis of any numerical method of implementing the procedures in generality.

A version of approximate orthogonality based on local expansions can be used to study Miss Hills's question about the second formulation of the Michaelis-Menten model. We write $x_i = \bar{x} + d_i$, $c^2 = n^{-1} \Sigma\, d_i^2/\bar{x}^2$ and assume in the expansions that c^2 is small. One finds with the particular choice in Miss Hills's notation of $\alpha^3 b(\lambda) = n\bar{x}^2\lambda^2$ that we may take

$$\beta = (\bar{x}/\lambda)\{1 - \lambda + c^2(1 - 6\lambda + 6\lambda^2)\}.$$

Note that for some purposes, such as computing $\tilde{w}_c$, an explicit expression for β is not needed. From the point of view of model interpretation the dependence of the parameterization on the design is a misfortune.

An important aspect of orthogonality that we did not discuss is its interpretation via estimating equations and the comments of Mr Firth and Dr Lloyd on this topic are most welcome. In a recent paper Liang (1986) discusses this in some detail; he shows that $|d\hat{\lambda}_\psi/d\psi| = o_p(1)$ and that the score function based on the conditional profile likelihood is, at least in special cases, more nearly unbiased than the score function from the profile likelihood.

As a number of contributors (Barndorff-Nielsen, Critchley, Smith and McCullagh) emphasize, and as we mentioned in the paper, our discussion is not exactly invariant under nonlinear changes in the orthogonalized nuisance parameter λ, although it is invariant to the order considered in the asymptotic expansions. Professor Smith's example is one where the choice does affect inference about ψ, presumably because there is so much uncertainty in the data that the form of the $O_p(n^{-1})$ term is important. Nearness of the problem to nonregularity may also be relevant. It is natural to aim to resolve this nonuniqueness by higher order expansions. While we have explored a number of such possibilities none we have found so far is totally satisfactory and easily implemented. From one point of view the inclusion of the term $|\partial\hat{\lambda}_\psi/\partial\hat{\lambda}|$ is the natural way to restore invariance, thus leading to Barndorff-Nielsen's modified profile likelihood, although, as we discuss below, it is not clear that this is the most appropriate objective. As $|\partial\hat{\lambda}_\psi/\partial\hat{\lambda}|$ is often very difficult to calculate, we have investigated a transformation of λ to make the term vary as slowly as possible with ψ. This leads for one-dimensional λ to a differential equation for the preferred parameterization with solution

$$\lambda^* = \int^{\lambda} \{i_{\lambda\lambda}(\psi, \chi)/i_{\psi\psi\lambda}(\psi, \chi)\} d\chi,$$

where $i_{\psi\psi\lambda} = n^{-1}E(\partial^3 l/\partial\psi^2\partial\lambda)$. For multidimensional λ there results a system of q partial differential equations to be solved by the method of characteristics.

We do not know the answer to Professor Barndorff-Nielsen's question whether or when modified profile likelihood is preferable to one of the conditional versions we proposed. It is possible to compute our (8) or (9) from the data without explicitly assuming that one of the four sufficiency reductions holds, and if (8) is used the resulting likelihood is constructed from an exact conditional density. However, it may be important for the general theory to be rather explicit about the transformation from the minimal sufficient statistic to the maximum likelihood estimate; furthermore the error in using the approximation to the conditional likelihood in deriving modified profile likelihood may be negligible for all practical purposes. In at least two examples the conditional likelihood seems to give better results than modified profile likelihood. The first is the weighted means example of Section 4.2.1, begging Professor Sprott's pertinent query about the "correct" solution for this, and the second is the very interesting example suggested by Dr Lloyd, a version of Fieller's problem concerning the ratio of normal means. In this example the exact versions of w, w_c, $\tilde{w}_c$ and w_c^* give identical likelihoods for ψ, whereas the modified profile likelihood, calculated from $p(\hat{\psi} \mid \hat{\lambda}_{\psi})$, gives a different and apparently inferior version; see Dr Lloyd's remarks. However, w_c and $\tilde{w}_c$ must be calculated in the λ parameterization in which the problem is formulated.

With the current strong interest in the differential geometric aspects of statistics it is pleasing, although not surprising, that there were a number of comments on geometry (Barndorff-Nielsen, Mitchell, Amari, Moolgavkar and Prentice). It seems likely that such considerations will be qualitatively helpful over the general second-order choice of test statistics, and hopefully on the issue discussed below concerning the particular version of orthogonalized nuisance parameter most appropriate. Some aspects of this are described by Amari (1985, Ch. 8). Our discussion, however, has been strongly influenced by the desire to handle particular examples, and here the role of differential geometry is less clear. The complexity of the calculations behind Dr Mitchell's interesting results serves to emphasize the difficulty of handling statistically simple situations. Thus in discussing multidimensional parameters of interest, it is probably easier to check the compatability equations directly to investigate the possible existence of orthogonalized nuisance parameters than to use the geometric considerations so clearly summarized by Professors Moolgavkar and Prentice.

Professor Amari also raises the possibility of studying the characteristics of the proposed test statistics via differential geometric techniques. We find this a rather daunting task but look forward to further important results from him and his associates. One key queston is whether w_c can be adjusted by a Bartlett factor to satisfy Professor Amari's level condition. Dr Harris has summarized his very detailed calculations on this matter and it seems that the form of the $O_p(n^{-1/2})$ term in (21) means that w_c in general cannot be simply adjusted to improve the χ^2 approximation in the required way. In fact it can be shown that the addition of an $O_p(n^{-1/2})$ term to a chi-squared random variable can be corrected by a Bartlett factor only under a very special condition on the conditional variance of the added term.

Several contributions, especially those of Dr Sweeting, Mr Polson and Professor Louis, deal with the relation between our results and a Bayesian approach. Such parallels are valuable and are a natural extension of the work of Welch and Peers (1963). It must always be of interest to examine problems from different viewpoints, but if the notion of a prior distribution is taken seriously as a way of injecting further knowledge into a discussion, the treatment of orthogonal parameters as independent will be far from inevitable. It would take us too far afield to treat Mr Polson's final question as other than rhetorical. Clearly the Bayesian *formalism* is very appealing.

Dr Bailey gives an interesting discrete example: see also Cox (1984), where the difference of probabilities, also of epidemiological interest, is briefly discussed. A slightly simpler version concerns the comparison of two Poisson variables, where interest in the difference of the means demands an approximate discussion, the orthogonal parameter being the ratio of the means. This would be appropriate if the first Poisson variable represented background emission alone and the second source plus background. The question raised by Dr Bailey concerning the degree of conditioning appropriate in discrete problems is important and puzzling. In a more general setting the amount of conditioning involved in determining the probability of some specified event is settled by a balance between the selectivity achieved in conditioning and the "noise" introduced by overconditioning, but it is hard to make that notion precise in the present context.

We agree with Professor Barnard that certain questions may have to be regarded as unanswerable, or that there may be virtually no relevant information in the data; Professor Smith's flat likelihood may be an expression of this. Nevertheless we would be very reluctant to draw a strong distinction between questions that have an "exact" answer within some mathematical formalism and those that have only an approximate answer. Indeed one goal of our paper is to extend somewhat the availability of good approximations and to treat these and exact cases in a way that is unified. The example mentioned above concerning the difference of Poisson means is a case in point. Recent work shows that this is a case where the adjustment needed to ordinary profile likelihood is for most purposes negligible and that the confidence intervals obtained from the profile likelihood are satisfactory.

On the other hand, Dr Matthews's application is an important one where direct use of profile likelihood may be grossly misleading. Ms Marie Cruddas has investigated the simpler problem of estimating the autoregressive parameter ρ in a first-order autoregressive process on the basis of m small samples, the samples having different means but common ρ and variance. Confidence intervals from the modified likelihood procedures perform well in simulations even for m as small as 10, whereas intervals based on unmodified profile likelihood are strongly negatively biased.

We are grateful to Professor Fraser for his sympathetic comments on alternative but related viewpoints. We agree that if a reasonably simple one-dimensional statistic can be found whose distribution, conditional or otherwise, depends only on the parameter of interest then that provides an attractively direct and readily interpreted route to inference. One of the features of the use of confidence intervals via likelihood in such problems as the normal transformation model is that the search for such statistics is by-passed.

Several people (Barndorff-Nielsen, Mitchell, Barnard, Pierce and Sprott) pointed out the interesting orthogonality of $\mu + k\sigma$ and σ for suitable k in the location-scale model. As k is a fairly complicated function of the distribution of the ancillary statistic (or of the ancillary statistic itself if observed information is used) Dr Pierce's observation that contrasts in the means are also orthogonal to σ is particularly relevant. In the case of just two samples this is a special case of Dr Mitchell's result, because taking differences induces symmetry in the underlying distribution. We are not clear how this applies to the problem of more than two samples.

One motivation for our work was to extend methods that work well for exponential models and it would be interesting to examine their success for transformation models. Related to this are Mitchell's (1986) discussion of the geometry of elliptical models and the intriguing result that Barndorff-Nielsen's formula for the distribution of the maximum likelihood estimate is exact in transformation models and accurate to $O(n^{-3/2})$ in full exponential families. This relates to Mr Polson's question on the hyperboloid and inverse Gaussian models: the group structure is not needed to derive the modified profile and conditional profile likelihood from the saddlepoint approximation.

We are grateful to Dr Davison for a clear summary of the predictive likelihood approach. Formally at least one can identify the random variable to be predicted with the parameter to be estimated and obtain a correspondence between various approximate predictive likelihoods and modified or conditional profile likelihoods. This identification may prove valuable for clarifying aspects of both inference and prediction, and the relationship between them.

Finally it may help, in particular partly to assuage Dr Critchley's anxieties, to set out the broad qualitative objectives of our paper.

(i) Many problems of formal inference can be tackled only via approximate arguments. There are many different procedures equivalent to the first order of asymptotic theory. Often these procedures will in practice give virtually the same answer, but this is not always so and there is thus a need for a second-order approach to clarify the choice.

(ii) We are guided in part by the conditioning procedure that is often effective in exponential family problems, and in part by the need to adjust ordinary profile likelihood for its defects when there are appreciable numbers of nuisance parameters. In some examples the modification of ordinary profile likelihood is negligible, but in others such modification leads to improved likelihood-based inference procedures.

(iii) These considerations lead first to a formulation of the model in which nuisance parameters λ are orthogonal to the parameter of interest, and then to a variety of conditional test statistics. Major outstanding questions include whether a requirement in addition to orthogonality can sensibly define λ uniquely, and under what conditions any of the conditional procedures can be shown to have good statistical properties. We have concentrated on version (11), $\tilde{w}_c$, because it is often straightforward to calculate, and because its stochastic expansion can be directly compared with that of the profile likelihood ratio. The justification put forth is that among procedures equivalent to the first order (11) is in a certain sense as close as possible to the likelihood procedure for a known value of the nuisance parameters.

While further work is certainly needed before the central objectives are fully met in a simple, easily implemented and conceptually compelling way, we have been much encouraged by the breadth and depth of the comments on our paper.

One of us (D.R.C.) is grateful for the hospitality of Department of Statistics, University of Toronto, during some of the work on the reply.

REFERENCES IN THE DISCUSSION

Amari, S. (1985) *Differential Geometrical Methods in Statistics.* New York: Springer Verlag. (Springer Lecture Notes in Statistics, 28.)

Amari, S. and Kumon, M. (1983) Differential geometry of Edgeworth expansions in curved exponential family. *Ann. Inst. Statist. Maths*, **34A**, 1–24.

Atkinson, A. C. (1985) *Plots, Transformations and Regression.* Oxford: University Press.

Barndorff-Nielsen, O. E. (1986a) Likelihood and observed geometries. *Ann. Statist.*, **14**, 856–873.

———(1986b) Differential and integral geometry in statistical inference. In *Differential Geometry in Statistical Inference* (IMS Monograph, to appear).

Bernardo, J. M. (1979) Reference posterior distributions for Bayesian inference. (with Discussion). *J. R. Statist. Soc.* B, **41**, 113–147.

Bickel, P. and Doksum, J. (1981) An analysis of transformations revisited. *J. Amer. Statist. Ass.*, **76**, 296–311.

Boothby, W. M. (1975) *An Introduction to Differentiable Manifolds and Riemannian Geometry.* New York: Academic Press.

Brillinger, D. R. (1982) A generalised linear model with "Gaussian" regressor variables. In *A Festschrift for Erich Lehmann* (P. J. Bickel, K. A. Doksum and J. L. Hodges, eds). San Francisco: Wadsworth.

Butler, R. W. (1986a) Predictive likelihood inference with applications (with Discussion). *J. R. Statist. Soc.* B, **48**, 1–38.

———(1986b) Approximate pivots for prediction. Unpublished.

Cook, R. D. (1986) Assessment of local influence (with Discussion). *J. R. Statist. Soc.* B, **48**, 133–169.

Cook, R. D. and Weisberg, S. (1982) *Residuals and Influence in Regression.* London: Chapman and Hall.

Cox, D. R. (1972) Regression models and life tables (with Discussion). *J. R. Statist. Soc.* B, **34**, 197–220.

———(1984) In discussion of paper by Yates, F. *J. R. Statist. Soc.* A, **147**, 451.

Davison, A. C. (1986) Approximate predictive likelihood. *Biometrika*, **73**, 323–332.

Dawid, A. P., Stone, M. and Zidek, J. (1973) Marginalization paradoxes in Bayesian and structural inference (with Discussion). *J. R. Statist. Soc.* B, **35**, 189–233.

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

———(1961a) Sampling the reference set. *Sankhya*, **23**, 3–8.

———(1961b) The weighted mean of two normal samples with unknown variance ratio. *Sankhya*, **23**, 103–114.

Fraser, D. A. S. (1986) Fibre analysis and tangent analysis. Submitted to *Statistical Papers.*

Huzurbazar, V. S. (1956) Sufficient statistics and orthogonal parameters. *Sankhya*, **17**, 217–220.

Kobayashi, S. and Nomizu, K. (1963) *Foundations of Differential Geometry. Vol. 1.* New York: Interscience.

Kumon, M. and Amari, S. (1983) Geometrical theory of higher order asymptotics of test, interval estimators and conditional inference. *Proc. Roy. Soc. London*, A, **397**, 429–458.

Kyriakidis, E. (1986) M. Sc. Report, Imperial College, London.

Lawrance, A. J. (1987) The score statistic for regression transformation. *Biometrika*, **74**, in the press.

Lejeune, M. and Faulkenberry, G. D. (1982) A simple predictive density function. *J. Amer. Statist. Ass.*, **77**, 654–657.

Liang, K. Y. (1986) Estimating functions and approximate conditional likelihood. Technical Report 610, Dept of Biostatistics, John Hopkins University.

Massam, H. and Fraser, D. A. S. (1985) Conical tests: observed levels of significance and confidence regions. *Statistical Papers*, 26, 1–17.

Mathiasen, P. E. (1977) Prediction functions. *Scand. J. Statist.*, **6**, 1–21.

Mitchell, A. F. S. (1986) Statistical manifolds of univariate elliptic distributions. *Int. Statist. Rev.*, to appear.

Patefield, W. M. (1977) On the maximised likelihood function. *Sankhya* B, **39**, 92–96.

Patterson, H. D. and Thompson, R. (1974) Maximum likelihood estimation of components of variance. *Proc. 8th Int. Biometrics Conf.*, 197–207.

Prescott, P. and Walden, A. T. (1980) Maximum likelihood estimation of the parameters of the generalized extreme-value distribution. *Biometrika*, **67**, 723–724.

———(1983) Maximum likelihood estimation of the parameters of the three-parameter generalised extreme-value distribution from censored samples. *J. Statist. Comput. Simul.*, **16**, 241–250.

Ross, G. J. S. and Preece, D. A. (1985) The negative binomial distribution. *The Statistician*, **34**, 323–336.

Ross, G. J. S. (1975) Simple non-linear modelling for the general user. *Proc. 40th Session Int. Statist. Inst.*, **2**, 585–593.

Ruud, P. A. (1986) Consistent estimation of limited dependent variable models despite misspecification of distribution. *J. Econometrics*, **32**, 157–187.

Skovgaard, H. M. (1986) Saddlepoint expansions for directional test probabilities.

Smith, R. L. (1985) Maximum likelihood estimation in a class of nonregular cases. *Biometrika*, **72**, 67–90.

Solomon, P. J. (1984) Effect of misspecification of regression models in the analysis of survival data. *Biometrika*, **71**, 291–298.

Sprott, D. A. and Viveros, R. (1984) The interpretation of maximum likelihood estimation. *Can. J. Statist.*, **12**, 27–38.

Sweeting, T. J. (1985) Consistent prior distributions for transformed models. *Bayesian Statistics 2*, 755–762.

Tierney, L. and Kadane, J. B. (1986) Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Ass.*, **81**, 82–86.

Viveros, R. (1985) Estimation in small samples. *Ph.D. Thesis, Dept of Statist. and Actuarial Science*, University of Waterloo.

Viveros, R. and Sprott, D. A. (1986) Allowance for skewness in maximum likelihood estimation with application to the location-scale model. Submitted for publication.

Welch, B. L. and Peers, H. W. (1963) On formulae for confidence points based on integrals of weighted likelihoods. *J. R. Statist. Soc.* B, **25**, 318–329.