Evolution in Statistical Inference: From Sufficiency to Likelihood Asymptotics

D.A.S. Fraser
N. Reid
Department of Statistics
University of Toronto
Toronto, Ontario
Canada M5S 1A1

1. SUMMARY AND INTRODUCTION

Sufficiency and conditionality are long standing basic methods central to statistical inference; the first is effective primarily with exponential models and the second with location and transformation models. Recent asymptotics involving the likelihood function has led to highly accurate test procedures even for small sample sizes and provides well defined reduction methods for general models. Interestingly the contradictory methods of sufficiency and conditionality are seen in fact to both be in harmony with the recent general methods. We survey this evolution in statistical inference.

Sufficiency, minimum variance unbiased, and uniformly most powerful have a prominent place in theoretical statistics; likelihood methods and the maximum likelihood estimates occupy a related prominent place in dealing with new applications. The present survey is to highlight recent asymptotic theory that places the likelihood methods in a very important and central place for both applications and theory.

In Section 2 we give a brief overview of the roles of sufficiency and conditionality in

statistical inference. Section 3 discusses the observed likelihood function and the standard first order measures of departure. The determination of the canonical representation of exponential models is discussed in Section 4; this is a basic ingredient in recent approximation methods. Third order density approximation results are outlined in Section 5, and third order distribution function approximations in Section 6. These form the basis for dimension reduction (Section 7) and lead to third order p values for component scalar parameters. Section 8 contains some brief discussion.

2. SUFFICIENCY AND CONDITIONALITY

Consider a statistical problem where the basic variable y has dimension n, the overall parameter θ has dimension p, and the parameter of interest ψ has dimension r; then either directly or implicitly we have $\theta = (\lambda, \psi)$ where λ is a nuisance parameter indexing the possible distributions for each value of ψ.

Sufficiency was introduced by Fisher (1920, 1922) as a reduction method to simplify statistical models. In many cases the reduction is from the model for the original variable to a model for the sufficient statistic of dimension p. This happens typically for exponential models where the range for the canonical parameters has full dimensions p:

$$f(y;\theta) = \exp\{\varphi'(\theta)s(y) - k(\theta)\}h(y) \ . \tag{2.1}$$

As an example consider y_1, y_2 from $\theta e^{-\theta y}$ on $(0, \infty)$. The statistic $s = y_1 + y_2$ is sufficient and the distribution of, say, $y_1|s$ is uniform $(0, s)$ and accordingly is free of θ.

We note that in going for y to $s(y)$ we have reduced the dimension of the variable from n to p and replaced the original model by the marginal model for the new variable. If we refer to the θ-free distribution of $y|s$ as error, then we have eliminated the error by marginalization.

A less well known aspect of the usual sufficiency reduction concerns inference for an interest parameter ψ. For the case that ψ is a canonical parameter in (2.1), say

$$f(y;\theta) = \exp\{\lambda' s_1 + \psi' s_2 - K(\theta)\}h(y) \ , \tag{2.2}$$

we have that the distribution of $s_2|s_1$ is free of the nuisance λ and thus depends only on ψ; this conditional model for $s_2|s_1$ is widely recognized as being the appropriate basis for inference concerning ψ free of the nuisance parameter λ. As an example consider y_1 from $\theta_1 e^{-\theta_1 y_1}$ and y_2 from $\theta_2 e^{-\theta_2 y_2}$, both on $(0, \infty)$; with $\psi = \theta_2 - \theta_1$ as interest parameter, we have that the distributions of $y_2 - y_1 | y_2 + y_1$ depends only on ψ and thus is free of say $\lambda = \theta_1 + \theta_2$.

Conditionality was introduced by Fisher (1934) as a reduction method for location models where there is no loss of information as measured by Fisher information. The method has been generalized to transformation models and is well illustrated by the regression model

$$f(y - X\beta) \tag{2.3}$$

say with known error scaling; the reduction is from the full model on R^n to the conditional model for say $b(\mathbf{y}) = (X'X)^{-1}X'y$ given the residuals $a(y) = y - Xb(y)$, and thus is from dimension n to dimension p.

We note that the original model is replaced by a conditional model for the new variable. Also if we refer to the β free distribution for the residuals as error, then we have eliminated the error by conditioning.

Now consider the follow-up aspect of inference for the interest parameter ψ. For the case that ψ is canonical, say with

$$\beta = \begin{pmatrix} \lambda \\ \psi \end{pmatrix} \ , \ b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \ , \tag{2.4}$$

then the marginal distribution (within the preceding conditional distribution) of $b_2 - \psi$ is free of the nuisance parameter λ and is recommended for inference concerning ψ. Note that in going from b to b_2 we have reduced the dimension from p to r and replaced the initial model by the marginal model for b_2. In some sense we have eliminated the nuisance parameter effect by marginalization.

As an overall summary of the sufficiency and conditionality methods, let e be a variable associated with error, u be a variable associated with the nuisance parameter, and v be one for the interest parameter. We then have the pattern:

Reduction method	Original variable	Eliminate error use	Eliminate nuisance use
Sufficiency	(e,u,v)	(u,v)	$v\|u$
Conditionality	(e,u,v)	$(u,v)\|e$	$v\|e$

For sufficiency we marginalize and then condition; for conditionality we condition and then marginalize. Some statisticians feel that it is inappropriate to have different procedures for different models. By contrast we feel that different models are presenting information in different ways and that one can then reasonably expect to see different procedures to elicit the information. We will see that recent asymptotics presents a unifying thread to inference procedures allowing them to be examined from a common viewpoint. Interestingly, the unifying pattern corresponds superficially more to the second procedure, the conditionality procedure.

3. LIKELIHOOD AND FIRST ORDER ORDER ASYMPTOTICS

The likelihood function records the probability at a data point y as a function of the parameter that indexes possible distributions. For a model $f(y;\theta)$ the likelihood function in logarithmic form is

$$\ell(\theta;y) = a + \ln f(y;\theta) \tag{3.1}$$

where a is an arbitrary constant with the effect that only relative likelihood from one θ value to another value is recorded. For some discussion, see for example Fraser (1976, Ch.8).

Two key characteristic of a likelihood function are the maximum likelihood value $\hat{\theta} = \arg\sup \ell(\theta;y)$ which gives the maximum value to likelihood and the observed information

$$\hat{\jmath} = -\ell_{\theta\theta}(\hat{\theta};y) = -\frac{\partial^2}{\partial\theta\partial\theta'}\ell(\theta;y)\Big|_{\hat{\theta}} \tag{3.2}$$

which records the negative curvature or Hessian matrix at the maximum likelihood value.

Now consider a statistical model $f(y;\theta)$ with n dimensional variable y and p dimensional parameter θ, and suppose that the model has asymptotic properties as $n \to \infty$ and that the per coordinate information matrix $i(\theta)/n$ is bounded below by a positive definite matrix, where $i(\theta) = V\{\ell_\theta(\theta;y);\theta\}$ is the variance matrix of the score function $\ell_\theta(\theta;y) = (\partial/\partial\theta)\ell(\theta;y)$.

Familiar first order asymptotics uses the Central Limit Theorem to show that $\ell_\theta(\theta;y)$ is asymptotically normal $\big(0;i(\theta)\big)$. Taylor Series expansion of $\ell_\theta(\theta;y)$ about $\hat\theta$ then shows that the normal limiting distribution of the score function transfers linearly to a normal limiting distribution for $\hat\theta$ with mean θ and variance $i(\theta)$; for this, the linear adjustment is given by the per observation observed information $\hat\jmath/n$, which converges by the Law of Large Numbers to the per observation information $i(\theta)/n$. A further analysis then in which likelihood $\ell(\theta;y)$ is expanded about $\hat\theta$ to quadratic terms shows that likelihood expressed in standardized coordinates has quadratic limiting form in an $O(n^{-1/2})$ neighbourhood of $\hat\theta$. As a consequence we have that the following quantities

$$z^2 = \ell_{\theta'}(\theta;y)\hat\jmath^{-1}\ell_\theta(\theta;y) \tag{3.3}$$

$$q^2 = (\hat\theta-\theta)'\hat\jmath(\hat\theta-\theta) \tag{3.4}$$

$$r^2 = 2\{\ell(\hat\theta;y) - \ell(\theta;y)\} \tag{3.5}$$

are each asymptotically chi-square with p degrees of freedom. The Central Limit Theorem produces the chi-square distribution for the first and the successive approximations lead to the limiting distributions for the second and third. In applications however the accuracy of the approximation is usually best for the last and degrades through the second to first.

With data y and a parameter value θ of interest we refer to the quantities as measures of departure of data y from what is expected under θ. We find it of particular interest that the observed values for r^2, q^2, z^2 can all be calculated from the likelihood function $\ell^\circ(\theta) = \ell(\theta;y^\circ)$ at the data point of interest. Also we note that only r^2 is parameterization

invariant although z^2 could be made invariant by replacing $\hat{j}$ by the expected information $i(\theta)$ at the parameter value being tested.

If θ is a scalar parameter it is usually advantagious to work with the square roots of (3.3), (3.4), (3.5) and attach an appropriate sign:

$$z = \ell_\theta(\theta; y)\hat{j}^{-1/2} \tag{3.6}$$

$$q = (\hat{\theta} - \theta)\hat{j}^{1/2} \tag{3.7}$$

$$r = \text{sgn}(\hat{\theta} - \theta) \cdot \left[2\{\ell(\hat{\theta}; y) - \ell(\theta; y)\}\right]^{1/2} \tag{3.8}$$

These quantities are asymptotically standard normal. The comments following (3.5) apply perhaps more strongly to the present quantities.

Now consider a component interest parameter ψ with corresponding nuisance parameter λ; for convenience we take $\theta = (\lambda', \psi')'$. For testing ψ we have the following first order likelihood based quantities

$$z_\psi^2 = \ell'_\psi(\hat{\theta}_\psi; y)\hat{j}^{\psi\psi}\ell_\psi(\hat{\theta}_\psi; y) \tag{3.9}$$

$$q_\psi^2 = (\hat{\psi} - \psi)'(\hat{j}^{\psi\psi})^{-1}(\hat{\psi} - \psi) \tag{3.10}$$

$$r_\psi^2 = 2\{\ell(\hat{\theta}; y) - \ell(\hat{\theta}_\psi; y)\} \tag{3.11}$$

where

$$\hat{j}^{\theta\theta} = \begin{pmatrix} \hat{j}^{\lambda\lambda} & \hat{j}^{\lambda\psi} \\ \hat{j}^{\psi\lambda} & \hat{j}^{\psi\psi} \end{pmatrix} = \begin{pmatrix} \hat{j}_{\lambda\lambda} & \hat{j}_{\lambda\psi} \\ \hat{j}_{\psi\lambda} & \hat{j}_{\psi\psi} \end{pmatrix}^{-1} \tag{3.12}$$

is the inverse of the observed information matrix; these quantities have asymptotically the chi-square distribution with r degrees of freedom. Again we note that the quantities can all be calculated from the observed likelihood but only r_ψ^2 has appropriate parameterization invariance. If ψ is scalar it is appropriate to use square roots of the preceding and attach an appropriate sign:

$$z_\psi = \ell_\psi(\theta; y)(\hat{j}^{\psi\psi})^{1/2} \ , \tag{3.13}$$

$$q_\psi = (\hat{\psi} - \psi)(\hat{j}^{\psi\psi})^{-1/2} = (\hat{\psi} - \psi)\frac{|\hat{j}|^{1/2}}{|\hat{j}_{\lambda\lambda}|^{1/2}} \ , \tag{3.14}$$

$$r_\psi = \text{sgn}(\hat{\psi} - \psi) \cdot \left[2\{\ell(\hat{\theta}; y) - \ell(\hat{\theta}_\psi; y)\}\right]^{1/2} \ . \tag{3.15}$$

A slightly modified version of (3.14) arises in certain contexts:

$$q_\psi = (\hat{\psi} - \psi)\frac{|\hat{j}|^{1/2}}{|j_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2}} = (\hat{\psi} - \psi)\frac{|\hat{j}|^{1/2}}{|\hat{j}_{\lambda\lambda}|^{1/2}} \cdot \frac{|\hat{j}_{\lambda\lambda}|^{1/2}}{|j_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2}} \tag{3.16}$$

The preceding measures of departure (3.13) – (3.16) are asymptotically standard normal to first order.

4. CANONICAL EXPONENTIAL MODELS FROM LIKELIHOOD

Consider an exponential model $f(y;\theta)$ with p dimensional continuous variable y and p dimensional parameter and suppose we do not have available the actual canonical variables and parameters. Thus $f(y;\theta)$ is known to have the form

$$f(y;\theta) = \exp\{\varphi'(\theta)s(y) - K(\theta)\}h(y) \tag{4.1}$$

but the various entries on the right side are not explicitly available. We consider how to obtain the representation on the right side, using minimal information such as likelihood at a data point and sample space gradient of the likelihood function at that point.

For a data point y° let

$$\varphi(\theta) = \ell_{;y}(\theta; y^\circ) - \ell_{;y}(\hat{\theta}^\circ; y^\circ) \tag{4.2}$$

be a new parameterization based on likelihood characteristics at y°; for this we let $\ell_{;y}(\theta; y) = (\partial/\partial y)\ell(\theta; y)$ and say take $\ell(\theta; y) = \ln f(y;\theta)$ ignoring the usual additive constant which in any case disappears from the difference (4.2). Also let

$$s(y) = \ell_\varphi(\hat{\theta}^\circ; y) \tag{4.3}$$

be a score type reexpression of the initial variable y; for this we have $\partial\ell/\partial\varphi = (\partial\ell/\partial\theta)J^{-1}(\hat{\theta}^\circ)$ where $J(\theta) = (\partial\varphi/\partial\theta')$ is the Jacobian between the two parameterizations. We then have

$$f(y;\theta)dy = \exp\left\{\varphi' s + [\ell\{\theta(\varphi); y^\circ\} - \ell(\hat{\theta}^\circ; y^\circ)]\right\}H(s)ds \tag{4.4}$$

where $H(s)$ is the density corresponding to the cumulant generating function

$$c(\varphi) = \ell(\hat{\theta}^\circ; y^\circ) - \ell\{\theta(\varphi); y^\circ\}. \tag{4.5}$$

Note that as input to the expression (4.4) we have used only the observed likelihood $\ell(\theta; y^\circ)$ at a data point y° (with additive constant) and the observed likelihood gradient $\ell_{;y}(\theta; y^\circ)$ at that point. All of the ingredients to (4.4) are explicit with the exception of the Laplace inversion $H(s)$ from the cumulant generating function (4.5). In summary we obtain the canonical representation of the exponential model (4.4) using only the likelihood and its gradient at a single data point y°; the canonical variable and canonical parameter are given explicitly in (4.2), (4.3) and the cumulant generating function is essentially negative likelihood reexpressed in terms of the new parameterization.

5. DENSITY APPROXIMATIONS

Edgeworth expansions have a prominent place in the development of higher order asymptotics. Consider a p-dimensional variable y with cumulant generating function $c(t)$, the logarithm of the moment generating function. The Edgeworth expansion is an asymptotic series that uses the cumulant generating function to provide an approximation to the density $f(y)$, where typically y is a sum or average of the values from some initial sample.

For the scalar case, suppose that y has mean μ and standard deviation σ. The approximation then using the standardized variable $x = (y - \mu)/\sigma$ is

$$\varphi(x)\Big[1 + \{\gamma_3 H_3(x)\}\frac{1}{6} + \{3\gamma_4 H_4(x) + \gamma_3^2 H_6(x)\}\frac{1}{72} + O(n^{-3/2})\Big] \tag{5.1}$$

to the third order where

$$\begin{aligned} H_2(x) &= x^2 - 1\ ,\ H_3(x) = x^3 - 3x\ ,\ H_4(x) = x^4 - 6x^2 + 3 \\ H_5(x) &= x^5 - 10x^3 + 15x\ ,\ H_6(x) = x^6 - 15x^4 + 45x^2 - 15 \end{aligned} \tag{5.2}$$

are Hermite polynomial and γ_3, γ_4 are the third and fourth order cumulants of the standardized variable x. The Hermite polynomials are orthogonal polynomials for the standard normal density function $\varphi(x)$.

A direct expansion in terms of the orthogonal polynomials together with $\varphi(x)$ is called a Gram Charlier Type A expansion. The Edgeworth expansion regroups the Gram Charlier terms into those of order $O(1), O(n^{-1/2}), O(n^{-1}), \ldots$ and typically provides better approximations when terminated at a given order. For background see Feller (1971, Ch.15) and Barndorff-Nielsen & Cox (1989, Ch.4). For a survey of the multivariate Edgeworth expansion see McCullagh (1987, Ch.5).

Saddlepoint methods were introduced to statistics by Daniels (1954) and Barndorff-Nielsen & Cox (1979). For a density function $f(y)$ with cumulant generating function $c(t)$ these methods are more easily described in a statistical context in terms of the corresponding exponential model

$$f(y;\theta) = \exp\{\theta' y - c(\theta)\} f(y) \ . \tag{5.3}$$

The saddlepoint method reexpressed then in terms of statistical terminology uses likelihood function information to provide the following approximations to the density functions for y and $\hat{\theta}$

$$\frac{c}{(2\pi)^{p/2}} \exp\{\ell(\theta;y) - \ell(\hat{\theta};y)\} |\hat{\jmath}|^{-1/2} dy \ , \tag{5.4}$$

$$\frac{c}{(2\pi)^{p/2}} \exp\{\ell(\theta;y) - \ell(\hat{\theta};y)\} |\hat{\jmath}|^{1/2} d\hat{\theta} \ . \tag{5.5}$$

This uses the likelihood function $\ell(\theta;y) = a + \ln f(y;\theta)$ at the data point of interest where a is an arbitrary additive constant; for some discussion, see Fraser (1976, Ch.8). The factor $c = 1$ to order $O(n^{-1})$ and is constant to order $O(n^{-3/2})$; the formula has relative error of order $O(n^{-3/2})$.

Now consider a general statistical model $f(y;\theta)$ with continuous p dimension variable y and p dimensional parameter. The expression (4.4) with (4.2) and (4.3) produces the exponential model that coincides with the given model to first derivative at y°; it can be called the tangent exponential model at $y = y^\circ$.

Now in addition we assume that the general model $f(y;\theta)$ has asymptotic properties as some parameter $n \to \infty$. This can arise in a Central Limit Theorem context where y is

a sum or average of independent and identically distributed variables. It can also arise in a conditional context where independent variables are conditioned to bring the dimension from $n \to p$; see, for example, DiCiccio, Field & Fraser (1990). Results in Fraser & Reid (1993) show that general model $f(y;\theta)$ can be approximated by the exponential model

$$f(y;\theta)dy = \exp\{\varphi' s + \big[\ell\{\theta(\varphi); y^\circ\} - \ell(\hat\theta^\circ; y^\circ)\big]\} H(s) ds \tag{5.6}$$

as described in Section 4. In turn this model can be reexpressed to third order accuracy as

$$\begin{aligned} &\frac{c}{(2\pi)^{p/2}}\ \exp\{\ell\big(\theta(\varphi); y^\circ\big) - \ell(\hat\theta^\circ; y^\circ) + \varphi' s\} |\tilde\jmath|^{-1/2}\, ds \\ &= \frac{c}{(2\pi)^{p/2}}\ \exp\{\ell\big(\theta(\varphi); y^\circ\big) - \ell(\hat\theta^\circ; y^\circ) + \varphi' s\} |\tilde\jmath|^{1/2}\, d\hat\varphi\ ; \end{aligned} \tag{5.7}$$

the information is calculated from the tilted likelihood in the exponent. We refer to this as the tangent exponential model at the data point y°.

Now consider the accuracy of the approximation of the original statistical model $f(y;\theta)$ by the tangent model (5.7). Except for a constant $c' = 1 + d/n$ the tangent model (5.7) agrees with the original model to order $O(n^{-3/2})$ at y° and to first derivative at y°. Also to order $O(n^{-1})$ the tangent model agrees with the original model in a compact region of the standardized variable and parameter. We will see that this closeness of approximation is enough to produce third order significance values for scalar component parameters. For details, see Fraser & Reid (1989, 1993, 1994), Abebe, Cakmak, Cheah, Fraser, Kuhn & Reid (1994), Cakmak, Fraser, McDunnough, Reid & Yuan (1994), Cakmak & Fraser (1995).

6. DISTRIBUTION FUNCTION APPROXIMATIONS

For statistical applications we are commonly interested in tail probabilities and thus in distribution functions rather than density functions. The distribution function approximation corresponding to the Edgeworth density approximation (5.1) is easily derived:

$$\Phi(x) - \varphi(x)\Big[\gamma_3 H_2(x)\frac{1}{6} + \{3\gamma_4 H_3(x) + \gamma_3^2 H_5(x)\}\frac{1}{72} + O(n^{-3/2})\Big] \tag{6.1}$$

In practice this formula is often unreliable for larger values of $|x|$, and can give values outside the $[0,1]$ range for probabilities.

The saddlepoint approximation (5.5) to an exponential model (4.1) has a corresponding distribution function approximation (Lugannani & Rice, 1980)

$$\Phi(r) + \varphi(r)\Big\{\frac{1}{r} - \frac{1}{q}\Big\} \tag{6.2}$$

where r and q are the signed likelihood ratio (3.8) and standardized maximum likelihood departure (3.7) calculated in terms of the canonical variable φ as determined (Fraser, 1990) by (4.2); the approximation has relative error of order $O(n^{-3/2})$. In practice this approximation is found to be much more accurate than the Edgeworth (6.1); in some sense it uses the full likelihood function at the data point rather than in effect just several derivatives of the likelihood function at the parameter value of interest.

An asymptotically equivalent version of (6.2) has the form

$$\Phi(r - r^{-1}\log\frac{r}{q}) \ ; \tag{6.3}$$

the argument $r - r^{-1}\log(r/q)$ is called r^* by Barndorff-Nielsen (1991) and is standard normal to third order. From personal experience the formula (6.2) seems often to produce better approximations but (6.3) has the advantage that it does not produce p-values outside the $[0,1]$ range.

Now consider the general asymptotic model with a scalar variable y and a scalar parameter θ. Fraser & Reid (1989, 1993, 1994) show that the probability left of the data

point,

$$\begin{aligned} p(\theta) &= P(y \leq y^\circ; \theta) = P(\hat{\theta} \leq \hat{\theta}^\circ; \theta) \\ &= \Phi(r) + \varphi(r)\Big\{\frac{1}{r} - \frac{1}{q}\Big\} \ , \end{aligned} \tag{6.4}$$

for the tangent exponential model (5.7) coincides to third order with the probability left of the data point for the general model. If $\hat{\theta}(y)$ is monotone decreasing we would use $P(y \geq y^\circ; \theta)$ in expression (6.4).

The preceding result can be obtained by expanding the log-density about y° and $\hat{\theta}^\circ$ and showing that the probability left of the data point y° is independent of a constant in the expanded model that measures the amount the model differs from being exponential. In the preceding expression r is the signed likelihood ratio (3.8) and q is the standardized maximum likelihood departure (3.7), but calculated using the reparameterization φ,

$$q = (\hat{\varphi} - \varphi)\hat{j}_{\varphi\varphi}, \tag{6.5}$$

where φ is the nominal reparameterization (4.2).

In some contexts we will obtain an expression as in (5.7) but with an additional adjustment factor $A(\hat{\theta}; \theta)$ which is equal to 1 to order $O(n^{-1/2})$. In this case we then obtain the third order significance function $p(\theta)$ by adjusting q in (6.5)

$$Q = q/A \ . \tag{6.6}$$

and replacing q in (6.4) by Q. The additional factor typically arises from averaging over a nuisance parameter distribution but its effect is easily incorporated by the adjustment (6.6). See Cheah, Fraser, and Reid (1994).

Now consider the general asymptotic model with p dimensional variable y and p dimensional parameter θ and suppose we are interested in some scalar component parameter ψ. We find it convenient to use the abbreviation $\ell^\circ(\theta) = \ell(\theta; y^\circ)$.

At a data point y° we have the approximating tangent exponential model (5.7). For the particular value ψ of interest, we consider the contour $\hat{\theta}_\psi = \hat{\theta}^\circ_\psi$. In the space of the

variable s this is a straight line through $s^\circ(=0)$ and parallel to the line in φ space through $\hat{\varphi}^\circ_\psi$ perpendicular to the contour with $\psi =$ constant. It can be shown (Barndorff-Nielsen, 1980, 1983) that there is a one-dimensional ancillary free of the nuisance parameter. Also it can be shown (Fraser & Reid, 1994) that, whatever the ancillary is, its distribution recorded on the line $\hat{\theta}_\psi = \hat{\theta}^\circ_\psi$ is unique to third order with density

$$\frac{c}{(2\pi)^{1/2}} \exp\{\ell^\circ(\hat{\theta}^\circ_\psi) - \ell^\circ(\hat{\theta}^\circ) + \bar{s}\bar{\varphi}\} \frac{|\tilde{j}_{(\theta\theta)}|^{1/2}}{|\tilde{j}_{(\lambda\lambda)}|^{1/2}} \cdot \frac{|\tilde{j}_{(\lambda\lambda)}|^{1/2}}{|\tilde{j}_{(\lambda\lambda)}(\hat{\theta}_\psi)|^{1/2}} \cdot d\bar{s} \tag{6.7}$$

where $\bar{s}$ is length along the line corresponding to $\hat{\theta}_\psi = \hat{\theta}^\circ_\psi$ and $\bar{\varphi}$ is the scalar parameter obtained by measuring φ in the increasing ψ direction for that line. The parentheses on the information subscripts are to indicate that they are to be recalibrate in the nominal φ parameterization. The model (6.7) has the form of an exponential model with adjustment factor A as described preceding (6.6). It follows that the significance function $p(\psi)$ is given to third order by formulas (6.2) or (6.3) using the signed likelihood ratio $r = r_\psi$ in (3.15) and the maximum likelihood departure q given by

$$Q = \left[\bar{\varphi}(\hat{\theta}^\circ) - \bar{\varphi}(\hat{\theta}^\circ_\psi)\right] \frac{|\tilde{j}_{(\theta\theta)}|^{1/2}}{j_{(\lambda\lambda)}(\hat{\theta}_\psi)|^{1/2}} \tag{6.8}$$

as in (3.16) but calibrated in the φ scaling. For details, see Fraser & Reid (1994). An alternative third order approximation is developed by Barndorff-Nielsen (1991).

7. DIMENSION REDUCTION

Consider a statistical model $f(y;\theta)$ with n dimensional variable y and p dimensional parameter θ, and suppose that the model has asymptotic properties as $n \to \infty$ and that the per coordinate information matrix $i(\theta)/n$ is bounded below by a positive definite matrix, where $i(\theta) = V\{\ell_\theta(\theta;y);\theta\}$ is the variance matrix of the score function $\ell_\theta(\theta;y) = (\partial/\partial\theta)\ell(\theta;y)$.

Barndorff-Nielsen's (1980, 1983) extensive analysis of likelihood assumed the presence of a third order ancillary of dimension $n-p$ for the parameter θ; no accessible procedure was available for the construction of such ancillary in a general context.

For a p dimensional statistical model we noted in the preceding section that for third order inference we need only the observed likelihood function and the gradient of the likelihood function calculated at the data point. What are the implication from this in the context of a third order ancillary?

Directly from the property of ancillary we have the factorization

$$f(y;\theta) = g(a)h(s|a;\theta)J(a,s) \tag{7.1}$$

in obvious notation, and thus have that the overall likelihood is equal to the conditional likelihood,

$$\ell(\theta;y) = \ell(\theta;s|a) \tag{7.2}$$

We then see that the gradient of likelihood in the conditional distribution is given by the gradient of the full likelihood $\ell(\theta;y)$ tangent to the ancillary surface. Let $V = (v_1,\ldots,v_p)$ be p vectors tangent to the ancillary surface at the data point and let $W = (w_1,\ldots,w_p)$ be the corresponding vectors in terms of the coordinates $s|a$ at the data point; then

$$\begin{aligned} \ell_{;V}(\theta;y^\circ) &= \left\{\frac{d}{dv_1}\ell(\theta;y),\ldots,\frac{d}{dv_p}\ell(\theta;y)\right\}\Big|_{y^\circ} \\ &= \left\{\frac{d}{dw_1}\ell(\theta|s,a),\ldots,\frac{d}{dw_p}\ell(\theta|s,a)\right\}_{y^\circ} = \ell_{;W}(\theta|s^\circ,a^\circ) \end{aligned} \tag{7.3}$$

Thus for third order inference from the conditional distribution we need the observed likelihood $\ell^\circ(\theta)$ and the observed likelihood gradient $\ell^\circ_{;V}(\theta)$ tangent to the ancillary surface.

A function $a(y)$ is said to be a first derivative ancillary at θ_0 if $(d/d\theta)\ln g(a;\theta)|_{\theta_0} = 0$ where $g(a;\theta)$ is the density for $a(y)$. Also a first derivative $n-p$ dimensional ancillary at $\hat\theta^\circ$ is a first order ancillary at the data point. A first derivative ancillary at $\hat\theta^\circ$ can be adjusted to become second or third order ancillary without changing its tangent directions at the data point. Skovgard (1986) discusses the successive improvement of the order of ancillarity. Fraser & Reid (1994) shows how a first derivative ancillary can have its order increased without altering the tangent directions at the data point. Thus it suffices to

obtain the likelihood gradient $\ell_{;V}(\theta)$ tangent to a first derivative ancillary at the data point.

Fraser (1964) derived a first derivative ancillary with a scalar parameter using location model theory. Consider a continuous model $f(y;\theta) = \Pi f_i(y_i;\theta)$. For the parameter value θ_0 we transform each coordinate y_i to a new variable x_i; for this let f and F be the density and distribution functions for y_i stochastically increasing in θ, and define

$$x_i = \int^{y_i} \Big\{ -F_y(y;\theta_0)/F_{;\theta}(y;\theta_0) \Big\} dy \ ; \tag{7.4}$$

then the density function say g for x_i satisfies $\partial g(x;\theta)/\partial\theta = -\partial g(x;\theta)/\partial x$ at $\theta = \theta_0$ and thus is a location model to first derivative at θ_0.

We now write the full model in terms of the coordinates x_i and have

$$g(x;\theta) = \Pi g_i(x_i;\theta) \ . \tag{7.5}$$

The location model determined at the parameter value θ_0 is

$$g^*(x;\theta) = \Pi g_i\{x_i - (\theta - \theta_0);\theta_0\} \tag{7.6}$$

and it agrees to first derivative with the original model (7.5). The model (7.6) then has an exact ancillary $(x_i - \bar{x}, \ldots, x_n - \bar{x})$ and correspondingly the model (7.5) has a first derivative ancillary at θ_0. We call this a tangent location model.

In terms of the original coordinates the ancillary direction vector $V = v$ has ith coordinate

$$v_i(y) = -\frac{\partial F_i(y;\theta)/\partial\theta}{\partial F_i(y;\theta)/\partial y}\Big|_{\hat{\theta}^\circ} \tag{7.7}$$

This approach has been extended to the vector parameter case in Fraser & Reid (1994) and Fraser, Monette, Ng & Wong (1992).

8. DISCUSSION

For a general asymptotic model with n dimensional variable y and p dimensional parameter we are concerned with inference for a scalar parameter $\psi = \psi(\theta)$.

The tangent location model in Section 7 at a data point y° gives the tangent directions $V = (v_1, \ldots, v_p)$ to a third order ancillary.

The tangent exponential model to the conditional distribution is then available from (5.7) using $\ell^\circ(\theta)$ and $\ell^\circ_{;V}(\theta)$. Third order significance $p(\psi)$ for testing ψ is then given by (6.1) or (6.2) using the signed profile likelihood (3.15) and the adjusted maximum likelihood departure (6.8).

For an indication of the high accuracy with these current likelihood based approximations, see for example Fraser (1990), DiCiccio, Field & Fraser (1990), Abebe, Fraser, Reid & Wong (1994), Cheah, Fraser, & Reid (1994), Fraser, Reid & Wong (1994).

REFERENCES

Abebe, F., Cakmak, S., Cheah, P.K., Fraser, D.A.S., Kuhn, J., and Reid, N. (1994). *Parisankhyan Samikhha*, to appear.

Abebe, F., Fraser, D.A.S., Reid, N. and Wong A. (1994). Nonlinear regression: Third order significance, submitted *Can. J. Statist.*

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

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

Barndorff-Nielsen, O.E. (1991). Modified signed log likelihood ratio. *Biometrika* **78**, 557-563.

Barndorff-Nielsen, O.E. and Cox, D.R. (1979). Edgeworth and saddlepoint approximations with statistical inference. *J.R. Statist. Soc. B* **41**, 279-312.

Barndorff-Nielsen, O.E. and Cox, D.R. (1989). Asymptotic Techniques for use of Statistics, London: Chapman and Hall.

Cakmak, S., Fraser, D.A.S., McDunnough, P., Reid, N. and Yuan (1995). Likelihood centered asymptotic model: exponential and location model version. *Festschrift for A.M. Mathai*, Ed: S. Provost, to appear.

Cakmak, S., and Fraser, D.A.S., (1995). Multivariate asymptotic model: exponential and location approximations, *Utilitas Mathematica*, to appear.

Cheah, P.K., Fraser, D.A.S., and Reid, N. (1994). Adjustment to likelihood and densities; calculating significance, submitted *Can. J. Statist.*.

Daniels, H.E. (1954). Saddlepoint approximations in statistics. *Ann. Math. Statist.* **25**, 631-650.

DiCiccio, T., Field, C., and Fraser, D.A.S. (1990) Marginal tail probabilities and inference for real parameters. *Biometrika* **77**, 77-95.

Feller, W. (1971). An Introduction to Probability Theory and its Applications, Second Edition, Vol. 2; New York: Wiley.

Fisher, R.A. (1920). A mathematical examination of the methods of determining the accuracy of an observation by the mean error, and by the mean square error. *Monthly Notes of the Royal Astonomical Society*, **80**(8), 758-770.

Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics. **A222**, 309-368.

Fisher, R.A. (1934). Two new properties of mathematical likelihood. *Proc. R. Soc.*, **A144**, 285-307.

Fraser, D.A.S. (1964). Local conditional sufficiency. *J. Roy. Statist. Soc.* **B26**, 52-62.

Fraser, D.A.S. (1976). Probability and Statistics: Theory and Applications, Toronto: ITS; New York:McGraw Hill.

Fraser, D.A.S. (1990). Tail probabilities from observed likelihoods. *Biometrika* **77**, 333-341.

Fraser, D.A.S., Monette, G., Ng, K.W., and Wong, A. (1992). Higher order approximations with generalized linear models, Symposium on Multivariate Analysis, Hong Kong, to appear.

Fraser, D.A.S. and Reid, N. (1989). ¿From multiparameter likelihood to tail probabilities for a scalar parameter. Technical Report, Department of Statistics, University of Toronto.

Fraser, D.A.S. and Reid, N. (1993). Simple asymptotic connections between densities and cumulant generating function leading to accurate approximations for distribution functions. *Statist. Sinica* **3**, 67-82.

Fraser, D.A.S. and Reid, N. (1994). Ancillaries and third order significance. *Utilitas Mathematica*, to appear.

Fraser, D.A.S., Reid, N. and Wong, A. (1994). Simple and accurate inference for the gamma model, submitted *Can. J. Statist.*

Lugannani, R. and Rice, S.O. (1980). Saddlepoint approximation for the distribution of the sums of independent random variables. *Adv. Appl. Prob.* **12**, 475-490.

McCullagh, P. (1987). Tensor Methods in Statistics, London:Chapman and Hall.

Skovgaard, I.M. (1986). Successive improvement of the order of ancillarily. *Biometrika* **3**, 516-519.