

**On conditional inference for a real parameter:  
a differential approach on the sample space**

D.A.S. Fraser

Department of Mathematics, York University, Toronto, M3T 1P3 Canada  
University of Toronto and University of Waterloo

N. Reid

Department of Statistics, University of Toronto, Toronto M5S 1A1 Canada

**SUMMARY**

A one dimensional conditional procedure defines a partition of the sample space into curves which can be represented by means of a unit vector field. A formula is given for the conditional distribution in terms of local properties of the vector field. Conditions are developed for reducing the first-order effects of nuisance parameters and reproducing to higher order the likelihood change for the parameter of interest. The emphasis is on extending exponential family methods after locally approximating the statistical model by an exponential family.

**Keywords and Phrases:**

ancillary, differential, exponential family, likelihood, orthogonality, sample space partitions, sufficiency

## 1. Introduction

A fairly general description of conditional inference is that only some portion of the sample space for the response vector is taken into consideration in constructing the inference method. The usual way of arranging this is to fix some appropriate components of the response vector, and use the resulting conditional distribution for inference.

As a simple example due to Fisher (1934), if  $y$  is a sample of size  $n$  from the location model  $f(y; \theta) \propto \prod f(y_i - \theta)$ , an  $(n - 1)$ -dimensional component of  $y$ , which can be written  $a \equiv (y_1 - \bar{y}, \dots, y_n - \bar{y})$ , has a distribution that does not depend on  $\theta$ . Although not the full substance of Fisher's discussion, it can be argued that the randomness in this  $(n - 1)$ -vector is not relevant for inference about  $\theta$ , and the conditional distribution of  $y$ , given  $a$ , is the appropriate reference distribution. This type of analysis can be developed for any transformation model: there always exists a vector of generalized residuals analogous to  $a$  above that provides an appropriate conditioning. One way to formalize this approach to conditioning is by means of Birnbaum's (1962) conditionality principle, which says to condition on a variable with a fixed distribution; various aspects of this are discussed in Berger and Wolpert (1985) and Evans, Fraser and Monette (1986). A somewhat different approach is to work with the related error or structural model in which case the conditioning is based on an identified component of error and is logically necessary; cf. Fraser (1979, Chap. 6).

Outside the class of transformation models, this approach can be approximated by, for example, finding a statistic whose distribution is approximately parameter-free (Barndorff-Nielsen, 1980; Cox, 1980; Amari, 1982) or by expressing the model so that it is approximately a transformation model (Barndorff-Nielsen, 1980; Hinkley, 1980).

In an exponential family model, conditioning is usually invoked for a different reason: to eliminate effects of nuisance parameters. In certain special cases, it is possible to find a function of the response that is sufficient for the nuisance parameters, in which case the resulting conditional distribution depends only on the parameter of interest. Extensions of and approximations to this type of conditioning are discussed in Barndorff-Nielsen (1983) and Cox and Reid (1987).

A computational advantage of any conditional approach is that it may result in a substantial reduction in the dimensionality of the problem, and so be much easier to implement numerically. Thus even if a marginal approach is preferred on theoretical grounds, a conditional approach that closely approximates it may be more useful. Examples of this type of conditioning are discussed in Fraser and Massam (1985) and Skovgaard (1987).

In this paper we investigate the construction of a one-dimensional conditional distribution for inference about a real parameter. This construction is carried out directly in the sense that instead of choosing a statistic to condition on, we search in the sample space for a one-dimensional conditional model to use for inference. The resulting conditional distribution should be insensitive to nuisance parameters, as in the exponential family models, and sensitive to the parameter of interest, as in the transformation models. Our approach is motivated by and extended from the exponential family setting, and so is expected to be more effective for models that are closer to exponential families than to transformation models.

By not specifying properties of the statistic to be conditioned on, we are effectively allowing any conditional distribution to be a candidate for inference. In Section 2 we show that this approach leads to examining local changes of the full model, where "local" means local to a point in the sample space. The parameter effects are obtained from the likelihood function, so our inference is specified by looking at the local properties of the likelihood ratio: the differential likelihood approach is developed in Section 3. For cases where the sample space dimension exceeds the parameter dimension, curved exponential models are used to provide local approximations; this raises parametrization issues that are discussed in Sections 3.3 and 3.5. Some technical aspects of implementation are discussed in Section 4.

## 2. One dimensional conditional procedures and differential likelihood

### 2.1 Sample space partitions

Consider a variable  $y = (y_1, \dots, y_k)'$  with sample space  $Y \subseteq \mathbb{R}^k$ , a parameter  $\phi = (\theta, \lambda)$  with parameter space  $\Omega \subseteq \mathbb{R}^{r+1}$  and a statistical model  $f(y; \phi)dy$ . To avoid difficulties of over-conditioning in discrete distributions, we assume that the distribution of  $y$  is continuous. We assume that  $y$  is the minimal sufficient statistic,  $\theta$  is a real-valued parameter of interest, and  $\lambda$  is an  $r$ -dimensional nuisance parameter.

Any one-dimensional conditional inference procedure can be defined by first specifying a one-to-one transformation from  $y$  to a pair  $(T, s)$ , where  $T$  is  $k-1$  dimensional, and then determining the conditional density of  $s$  given  $T$ . This density is

$$g(s|T; \phi)ds = h(T; \phi)f(y; \phi)J(y)ds \quad (2.1)$$

where  $y = y(T, s)$ ,  $J(y)$  is the Jacobian  $|\partial y/\partial(T, s)|$ , and  $1/h(T; \phi)$  is the marginal density for  $T$ .

A geometric description of the conditional procedure is that  $T$  generates a partition of  $Y$  into curves and for each curve  $s$  gives the position of  $y$  along that curve. Differentiating  $y$  with respect to  $s$  gives a tangent vector  $v(y)$  with  $i$ th component  $\partial y_i/\partial s$ ; if  $s$  is chosen to measure arc-length then  $v(y)$  will be a unit tangent vector. We thus have that any conditional procedure generates a unit vector field  $V = \{v(y)\}$  on  $\mathbb{R}^k$ . Conversely, under mild regularity conditions, a given unit vector field  $V$  can be integrated from any initial point  $y^{(0)}$  to produce a curve  $y = y(s, y^{(0)})$ , which thus gives a conditional procedure.

In (2.1)  $T$  is fixed, so only the dependence on  $s$  is needed and factors depending purely on  $T$  can be absorbed into the normalizing constant. In particular, for the Jacobian  $J = J(y) = J(T, s)$  we show below that

$$\frac{\partial \log J}{\partial s} = \sum_{i=1}^k \frac{\partial v_i(y)}{\partial y_i} = \text{div} \{v(y)\} \quad (2.2)$$

is given by the divergence function and thus find that  $J(y)$  is proportional to

$$c(y) = \exp \int_0^{s(y)} \text{div}[v\{y(s)\}]ds \quad (2.3)$$

where the initial point  $y^{(0)}$  corresponds to  $s = 0$ . The conditional density defined in (2.1) can then be re-expressed as

$$g(s|T; \phi)ds = k(T; \phi)f(y; \phi)c(y)ds \quad (2.4)$$

where  $y = y(s, y^{(0)})$  is the curve through  $y^{(0)}$  and  $k(T, \phi)$  is the norming constant.

An intuitive explanation of (2.2) follows from the transformation  $y \rightarrow y + \delta v(y)$  which, with  $\delta$  small, changes  $s$  to  $s + \delta$  leaving  $T$  fixed:  $dy$  is changed to  $|I + \delta V|dy$  where  $V = \partial v(y)/\partial y'$ , and  $dTds$  is unchanged. Thus  $dy = J(T, s)dTds$  becomes

$|I + \delta V|dy = J(T, s + \delta)dTds$ , which gives  $J(T, s + \delta)/J(T, s) = |I + \delta V|$  and thus

$$\frac{\partial \log J(T, s)}{\partial s} = \frac{\partial}{\partial \delta} |I + \delta V|_{\delta=0} = \text{tr} V = \text{div} \{v(y)\}.$$

One reason for looking at the conditional density from the point of view of equation (2.4) is that it is easily computed numerically once the vector field  $V$  has been determined. Choosing the vector field for inference about  $\theta$  is the subject of the next sections. For the moment note that a starting point  $y^{(0)}$  needs to be specified, which will typically be the observed data point. From point  $y^{(j)}$  the next point in a positive direction on the path is determined by  $y_i^{(j+1)} = y_i^{(j)} + \delta v_i(y^{(j)})$ ,  $i = 1, \dots, k$ , where  $\delta$  is the step size and  $c(y^{(j+1)}) = c(y^{(j)})[1 + \delta \text{div}\{v(y^{(j)})\}]$  is the expansion adjustment; corresponding expressions would be obtained for the negative direction. Thus the conditional density can be computed without ever specifying the variable  $T$  that is being conditioned on. The computational details may need some care in implementation, in order to obtain accurately the norming constant from the accumulated sum of the products  $f(y^{(j)}; \phi)c(y^{(j)})$ . In particular it may be necessary to allow the step size  $\delta$  to vary.

*Example 2.1: Normal circle*

Let  $(y_1, y_2)'$  be bivariate normal with mean vector  $(\theta \cos \lambda, \theta \sin \lambda)'$  and covariance matrix  $I$ . We transform  $(y_1, y_2)$  to  $(r, \hat{\lambda})$ , where  $r^2 = y_1^2 + y_2^2$  and  $\tan \hat{\lambda} = y_2/y_1$ . For illustrating the above discussion we consider conditioning on  $\hat{\lambda}$ ; recall that  $J(y) = r$ . To compute the divergence we start with the vector field  $v(y) = (\partial y_1/\partial r, \partial y_2/\partial r) = (\cos \hat{\lambda}, \sin \hat{\lambda}) = (y_1/(y_1^2 + y_2^2)^{1/2}, y_2/(y_1^2 + y_2^2)^{1/2})$ . In this example  $r$  does measure arc length along the ray determined by fixing  $\hat{\lambda}$ . Then  $\text{div}\{v(y)\} = (y_1^2 + y_2^2)^{-1/2} = r^{-1}$ , and  $c(y) = \exp \int_{r_0}^r s^{-1} ds = r/r_0$ , as required.

Note that if we condition on  $r$  instead, the vector field  $v(y) = (-y_2/(y_1^2 + y_2^2)^{1/2}, y_1/(y_1^2 + y_2^2)^{1/2})$  giving  $\text{div}\{v(y)\} = 0$ . In other words the Jacobian can be absorbed into the norming constant because it does not depend on  $\hat{\lambda}$ .

For interpretation and for graphical display it seems appropriate to rewrite (2.4) in terms of  $\hat{\theta}$ . This is relatively easy for the case of no nuisance parameters. We need to express  $\hat{\theta}$  as a function of  $y$  along the curve. Letting  $S(y; \theta)$  be the score function and using the directional derivatives

$$\frac{dS(y; \theta)}{dv(y)} \Big|_{\hat{\theta}(y)} = \Sigma \frac{\partial S}{\partial y_i} v_i(y) , \tag{2.5}$$

we obtain from  $S(y; \hat{\theta}) = 0$ ,

$$\frac{dS(y; \theta)}{dv(y)} \Big|_{\hat{\theta}} ds + \frac{\partial S(y; \theta)}{\partial \theta} \Big|_{\hat{\theta}} d\hat{\theta} = 0 . \tag{2.6}$$

This gives

$$\bar{g}(\hat{\theta}|T; \theta)d\hat{\theta} = k(T, \theta)f(y; \theta)c(y) \Big| \frac{dS(y; \theta)}{dv(y)} \Big|_{\hat{\theta}}^{-1} j(\hat{\theta})d\hat{\theta} \tag{2.7}$$

where  $j(\theta) = -S'(y; \theta)$  is the observed information function. All the factors in this expression are computationally available as indicated before the example. The relation of (2.7) to Barndorff-Nielsen's (1983) approximation to the density of  $\hat{\theta}$  is discussed in Section 5.

In the next sections we seek a conditional procedure  $g(s|T; \phi)ds$  that is insensitive to the nuisance parameters  $\lambda$  and best presents the available information for the primary parameter  $\theta$ . Inference from such a one-dimensional distribution with one-dimensional parameter is fairly

straightforward. Tests and observed levels of significance can be calculated by one-dimensional integrations, and confidence intervals are available by iterations on the calculations of the observed levels of significance. Note also as just described that the conditional density can typically be re-expressed as  $\bar{g}(\hat{\theta}|T, \phi)d\hat{\theta}$  for ease of interpretation.

## 2.2 Differential likelihood

We write the log-likelihood function for  $g$  relative to a fixed value  $\phi_0$  as

$$\tilde{l}(s; \phi) = \log \{g(s; \phi)/g(s; \phi_0)\} ; \quad (2.8)$$

we also use the notation

$$l(y; \phi) = \log \{f(y; \phi)/f(y; \phi_0)\} . \quad (2.9)$$

The conditional density  $g$  is constructed by moving along a curve from the initial data point; at each point  $s$  we need to know how the conditional density changes in moving to  $s + ds$ . For this we write

$$g(s + ds; \phi) = g(s; \phi_0) \exp \{\tilde{l}(s; \phi)\} \frac{g(s + ds; \phi_0)}{g(s; \phi_0)} \exp \{\tilde{l}(s + ds; \phi) - \tilde{l}(s; \phi)\} \quad (2.10)$$

to express the density at the new point in terms of: original null ( $\phi_0$ ) density; original likelihood ratio; new null density factor; likelihood difference to the new point. Only the latter two factors depend on the new point and only the last factor depends also on the parameter. We use this likelihood-dependent factor as the basis for choosing  $v(y)$  and thus the curve for conditional inference.

We write  $d\tilde{l}(s; \phi)$  for the likelihood difference on the curve and obtain

$$\begin{aligned} d\tilde{l}(s; \phi) &= \tilde{l}(s + ds; \phi) - \tilde{l}(s; \phi) \\ &= l(y + vds; \phi) - l(y; \phi) . \end{aligned} \quad (2.11)$$

We have complete control over the choice of likelihood difference on the curve, from the likelihood differences available on the original sample space; thus we approach finding a one-dimensional conditional procedure from the viewpoint of optimally choosing the likelihood difference at each step along this curve. The differential expression of  $d\tilde{l}(s; \phi)$  is

$$\begin{aligned} d\tilde{l}(s; \phi) &= \frac{\partial}{\partial s} \tilde{l}(s; \phi) ds \\ &= dl(y; \phi)|_{dy=vds} \\ &= \sum_{i=1}^k \frac{\partial}{\partial y_i} l(y; \phi) dy_i |_{dy=vds} \\ &= \sum_{i=1}^k \frac{\partial}{\partial y_i} l(y; \phi) v_i(y) ds \end{aligned} \quad (2.12)$$

Equation (2.12) indicates the spectrum of possible log-likelihood differences corresponding to the various possible directions  $v(y)$ .

The differential  $dl(y; \phi)$  is the minimal sufficient statistic at the point  $y$ , and it maps the tangent space  $\{dy\}$  to the space  $\mathbb{R}^k$  of real-valued functions of  $\phi$ . As we have assumed that  $y$  is minimal sufficient, this is a full-rank linear mapping; i.e. the coefficients of  $dy_i$  in (2.12) are  $k$  linearly independent functions of  $\phi$ .

### 3. Choice of conditional inference procedure

#### 3.1 Full exponential families

We have argued in Section 2 that a conditional inference procedure can be generated directly on the sample space by developing an inference contour incrementally, and that this local development should focus on the sample space differential of the log-likelihood function. We now consider whether or not, and how, the conditional procedure can be developed to be insensitive to the nuisance parameters and effective for inference about  $\theta$ . In Sections 3.2 and 3.3 we will consider these two requirements separately. Although this separation is somewhat technical, there are different issues involved, even though similar techniques are used.

The conditional procedure to be discussed in the remainder of this section is motivated by the following simple special case. Suppose the density for  $y$  takes the exponential family form

$$f(y; \phi) = \exp\{\theta y_1 + \lambda_1 y_2 + \cdots + \lambda_r y_k - K(\theta, \lambda) - d(y)\}. \quad (3.1)$$

There is no essential loss of generality in assuming that  $y$  is the minimal sufficient statistic; but (3.1) also incorporates the restriction that  $k = r + 1$ , and that the parameter of interest is a single component of the canonical parameter.

From various viewpoints it is appropriate to base inference for  $\theta$  on the conditional density of  $y_1$  given  $y_2, \dots, y_k$ . For this we note that  $(y_2, \dots, y_k)$  is sufficient for  $\lambda$  for fixed values of  $\theta$  and that similar or unbiased tests of  $\theta$  are based on the conditional density (Cox and Hinkley, 1974, p. 134; Lehmann, 1986, p. 145). Alternatively the contour given by fixing  $y_2, \dots, y_k$  is the unique contour along which likelihood difference depends on  $\theta$  only; see Section 2, and Fraser (1979, p. 81). In the differential sample-space approach fixing  $(y_2, \dots, y_k)$  is expressed by requiring  $dy_2=0, \dots, dy_k=0$ . This can also be written

$$dU_i(y; \phi) = 0; \quad i = 1, \dots, r \quad (3.2)$$

where  $U_i(y; \phi) = \partial \log f(y; \phi) / \partial \lambda_i$  is the score for  $\lambda_i$ .

Of particular importance for our analysis later, the conditioning just described is invariant under linear reparameterization of the nuisance parameter. For if we write  $\bar{\lambda}$  as a linear rank  $(k - 1)$  transformation of  $\lambda$  and  $\theta$ , (3.1) becomes

$$\exp[\theta\{y_1 - a(y_2, \dots, y_k)\} + \bar{\lambda}_1 \bar{y}_2 + \cdots + \bar{\lambda}_r \bar{y}_k - \bar{K}(\theta, \lambda) - d(y)],$$

and  $(\bar{y}_2, \dots, \bar{y}_k)$  is a one-to-one linear function of  $(y_2, \dots, y_k)$  only. The conditioning thus remains unchanged. This invariance implicitly underlies the differential approach.

#### 3.2 Elimination of nuisance parameters

In this subsection we assume that the dimension of the minimal sufficient statistic is equal to the dimension of  $\phi$ . As the parameter of interest  $\theta$  is one-dimensional, this implies that  $k - 1$  constraints analogous to (3.2) will define a one-dimensional conditional distribution.

From (2.12) it follows that the likelihood from the conditional distribution will not depend on the nuisance parameters if  $d\ell(y; \phi)$  does not depend on  $\lambda$ , i.e.  $(\partial/\partial \lambda_i)d\ell(y; \phi) = 0$ , for the direction  $v(y)$ . Except in special cases such as discussed above, these equations will not have a solution free of  $\phi$ . The simplest possibility is to restrict (3.2) to

$$dU_i(y; \phi)|_{\hat{\phi}} = 0 \quad i = 1, \dots, r, \quad (3.3)$$

where  $\hat{\phi} = (\hat{\theta}, \hat{\lambda})$  is the maximum likelihood estimate at the point  $y$ . This gives the  $r = k - 1$  constraints needed to determine the vector direction  $v(y)$  and thus the conditional procedure.

Another way of motivating (3.3) is to approximate  $f(y; \phi)$  by a tangent exponential model

$$f(y; \phi_0) \exp \{ (\theta - \theta_0) S_1(y; \theta_0) + \sum_{i=1}^r (\lambda_i - \lambda_{0,i}) U_i(y; \phi_0) - \tilde{K}(\theta, \lambda) \} \quad (3.4)$$

where  $\tilde{K}(\theta, \lambda)$  normalizes the density (3.4),  $S_1 = \partial \log f(y; \phi) / \partial \theta$  is the score for the parameter of interest, and  $\phi_0$  is some reference value. This expresses  $f(y; \phi)$ , approximately near  $\phi_0$ , as the simpler model (3.1); then applying the differential version of the conditioning gives  $dU_i(y; \phi_0) = 0$ . Although other possibilities could be envisioned, it seems appropriate to use  $\phi_0 = \hat{\phi} = \hat{\phi}(y)$  as the reference value in examining the choice of  $v(y)$  at  $y$ . The notation  $\phi_0$  has been retained up to this point in order to clarify the sample space differentiation.

Note that in the context of independent, identically-distributed sampling,  $f(y; \phi)$  will itself be a member of a full exponential family, but as the  $k$  canonical parameters will typically each be functions of  $(\theta, \lambda)$  the method for developing similar tests will not usually apply. The approximating model (3.4) essentially expresses the canonical parameters as approximate linear functions of the parameters of interest near  $\phi_0$ . Note also that the approximating model (3.4) is invariant under locally linear reparametrizations of the nuisance parameter. The orthogonal parametrization of  $\lambda$  could be used, but the inference contour specified by (3.3) will not be affected.

*Example 3.1: Ratio of exponential means*

Let  $(y_1, y_2)$  have the joint density  $f(y_1, y_2) = \lambda^2 \theta \exp(-\lambda y_1 - \lambda \theta y_2)$  and assume we have  $n$  independent pairs  $(y_{1i}, y_{2i})$ . The minimal sufficient statistic is  $(y_1, y_2) = \Sigma(y_{1i}, y_{2i})$ , and the nuisance score is  $2n/\lambda - y_1 - \theta y_2$ . Equation (3.3) becomes  $-dy_1 - \hat{\theta} dy_2 = 0$ ; as  $\hat{\theta} = y_1/y_2$ , this gives  $dy_1 / y_1 + dy_2 / y_2 = 0$  as the defining equation for the contour. The solution is  $y_1 y_2 = c$ ; i.e. the contour is defined by fixing the product  $y_1 y_2$ . In fact  $y_1 y_2$  is the maximum likelihood estimate not of  $\lambda$ , but of the version of  $\lambda$  that is orthogonal to  $\theta$ ; see Cox and Reid (1987, Ex. 3.1). Thus the solution of (3.3) is equivalent to conditioning on the maximum likelihood estimate of the orthogonalized nuisance parameter. This can be shown to hold more generally for cases where the observed information has the same nuisance parameter orthogonality as the expected information. The marginal inference obtained using the location structure of this model is compared to the conditional analysis in Fraser and Reid (1987).

The conditions (3.3) can be reexpressed using the gradient vectors

$$u_j(y; \phi) = (u_{j1}, \dots, u_{jk})' = \nabla U_j(y; \phi) \quad (3.6)$$

giving

$$u_j(y; \hat{\phi}) \cdot v(y) = 0 \quad j = 1, \dots, r \quad (3.7)$$

In the present context with  $r = k - 1$  we have that the  $k - 1$  equations (3.7) uniquely determine the unit vector  $v(y)$  except for positive or negative ( $\pm 1$ ) sign corresponding to direction.

We examine in Section 4 some technical details concerning the calculation of  $div \{v(y)\}$  from derivatives of the gradient vectors  $u_j(y; \phi)$ . Then, as indicated in Section 2, we are in a position to calculate the conditional distribution  $g(s; \phi) ds$ .

*3.3 No nuisance parameters*

In this case we have a  $k$ -dimensional sufficient statistic but only one parameter of interest. Two prototype examples are the  $(k, 1)$  curved exponential family and the location family. In the former an exact ancillary in general does not exist; in the latter the residuals or spacings are ancillary as mentioned in the Introduction, and the usual conditional procedure fixes these. Of course from the present point of view we are concerned with constructing a conditional density directly.

The Taylor series expansion of  $\log f(y; \theta)$  about some fixed point  $\theta_0$  gives

$$f(y; \theta) = f(y; \theta_0) \exp \{ (\theta - \theta_0) S_1(y; \theta_0) + \sum_{j=2}^{\infty} (\theta - \theta_0)^j S_j(y; \theta_0) / j! \}$$

where

$$S_j(y; \theta_0) = \{ \partial^j \log f(y; \theta) / \partial \theta^j \} |_{\theta = \theta_0}$$

is the score for the  $j$ th curvature parameter. We will truncate the expression after  $k$  terms and renormalize, giving

$$\tilde{f}(y; \theta) = f(y; \theta_0) \exp \{ \delta_1 S_1 + \sum_{j=2}^k \delta_1^j S_j / j! - \tilde{K}(\delta_1) \}, \quad (3.8)$$

where  $\delta_1 = \theta - \theta_0$ , as an approximation to the original density  $f$  near  $\theta_0$ . The density in (3.8) is a  $(k, 1)$  curved exponential model. If we start with a curved exponential model of the form  $\exp \{ \sum \phi_i(\theta) y_i - K(\theta) \} f(y; \theta_0)$ , approximation (3.8) is obtained from Taylor-series expansions of the parameters  $\phi_i(\theta)$ . The corresponding full exponential model is

$$f(y; \theta_0) \exp \{ \delta_1 S_1 + \sum_{j=2}^k \delta_j S_j - K(\theta_0, \delta) \}. \quad (3.9)$$

In this model the conditional inference contour is defined by  $dS_j(y; \theta_0) = 0, j = 2, \dots, k$ , as at (3.2), which suggests that we might use the  $k - 1$  constraints

$$dS_j(y; \theta) |_{\theta = \hat{\theta}} = 0, \quad j = 2, \dots, k \quad (3.10)$$

with the original model.

A complication with this approach is that the approximating model (3.8) and the embedding model (3.9) depend strongly on the particular parametrization of  $\theta$ . For example if  $\delta_1 = \theta - \theta_0$  is expressed as  $\bar{\delta}_1 + a \bar{\delta}_1^2$ , then the score statistics for the new parametrization are  $\bar{S}_1 = S_1$  and  $\bar{S}_2 = S_2 + 2a_2 S_1$ , and the linear  $\bar{\delta}_1$  term in the new expansion (3.8) corresponds to quadratic variation of  $\delta_1$ . More generally, a  $k$ th degree polynomial transformation  $\delta_1 = \bar{\delta}_1 + a_2 \bar{\delta}_1^2 + \dots + a_k \bar{\delta}_1^k$  gives scores  $(\bar{S}_1, \dots, \bar{S}_k)' = L(S_1, \dots, S_k)'$ , where  $L$  is a  $k \times k$  lower triangular matrix with ones on the diagonal, and the linear  $\bar{\delta}_1$  term now corresponds to a  $k$ th order variation of the original parameter  $\delta_1$ .

We note that there is a  $k$ th degree transformation that makes the scores  $\bar{S}_2, \dots, \bar{S}_k$  uncorrelated with  $\bar{S}_1 = S_1$  at the point  $\theta_0$  and this transformation can be constructed iteratively, as follows. Let  $I_{1j}(\theta_0) = \text{cov}\{S_1(y; \theta_0), S_j(y; \theta_0); \theta_0\}$ ; then with  $\delta_1 = \bar{\delta}_1 + a \bar{\delta}_1^2$ , the choice  $a = I_{12}(\theta_0) I_{11}^{-1}(\theta_0) / 2$  makes  $\bar{S}_2$  uncorrelated with  $S_1$ . Now if  $\delta_1 = \bar{\delta}_1 + b \bar{\delta}_1^3$  in the just obtained parametrization the choice  $b = I_{13}(\theta_0) I_{11}^{-1}(\theta_0) / 3!$  makes  $\bar{S}_3$  uncorrelated with  $S_1$  but does not alter  $\bar{S}_2$ , and so on. We thus replace the constraints (3.10) by

$$d\bar{S}_j(y; \theta)|_{\hat{\theta}} = 0 \quad j = 2, \dots, k$$

Fortunately the  $k$ th degree reparametrization does not need to be explicitly determined. Let

$$\tilde{S}_j = S_j - I_{1j}(\theta_0)I_{11}^{-1}(\theta_0)S_1, \quad j = 2, \dots, k. \quad (3.11)$$

These scores are uncorrelated with  $S_1$  at  $\theta_0$  and the vector  $(S_1, \tilde{S}_2, \dots, \tilde{S}_k)'$  is linearly equivalent to  $(S_1, S_2, \dots, S_k)'$ . As the same is true of  $(S_1, \bar{S}_2, \dots, \bar{S}_k)$  the two sets of transformed scores  $\tilde{S}_2, \dots, \tilde{S}_k$  and  $\bar{S}_2, \dots, \bar{S}_k$  are themselves related by an invertible linear transformation.

We propose then to replace the constraints (3.10) by

$$d\tilde{S}_j(y; \theta)|_{\hat{\theta}} = 0 \quad j = 2, \dots, k; \quad (3.12)$$

i.e. we determine the conditioning using in effect the higher order scores for a parametrization that makes these scores uncorrelated with  $S_1$ . Requiring the higher scores to be uncorrelated with the first in the embedding model (3.9) is the same as requiring the parameters  $\delta_2, \dots, \delta_k$  to be orthogonal to  $\delta_1$ . These parameters measure in some sense the position of the original model in the full exponential model and it seems sensible that they should be as free of  $\theta$  effects as possible. Some further discussion of this point is provided in Section 3.5.

*Example 3.2: Gamma hyperbola*

Assume we have  $n$  independent pairs  $(x_{1i}, x_{2i})$  from the density  $f(x_1, x_2; \theta) = \exp(-\theta x_1 - \theta^{-1} x_2)$ ; the joint log-likelihood is

$$l(\theta) = -(\theta - \theta_0)y_1 - (\theta^{-1} - \theta_0^{-1})y_2$$

where  $y_1 = \sum x_{1i}$  and  $y_2 = \sum x_{2i}$  are the two sufficient statistics. We have  $S_1(y; \theta) = -y_1 + \theta^{-2}y_2$ ,  $S_2(y; \theta) = -2\theta^{-3}y_2$ ,  $I_{11}(\theta) = 2n\theta^{-2}$  and  $I_{12}(\theta) = -2n\theta^{-3}$ . This gives  $\tilde{S}_2(\theta) = -\theta^{-3}y_2 - \theta^{-1}y_1$ , so (3.12) becomes  $\hat{\theta}^{-3}(\hat{\theta}^2 dy_1 + dy_2) = 0$ . Since  $\hat{\theta}^2 = y_2/y_1$ , the curve is defined by  $dy_1/y_1 + dy_2/y_2 = 0$ , i.e. by conditioning on the product  $y_1 y_2$ . In this example  $y_1 y_2$  is exactly ancillary for  $\theta$ . Note that if we had not orthogonalized the scores, but used the original score  $S_2$  directly, we would have conditioned on  $y_2$ .

*Example 3.3: Nonlinear regression*

Let  $y_i = g(\theta, x_i) + e_i$ ;  $i = 1, \dots, n$ , where the  $e_i$  are independent  $N(0, \sigma_0^2)$  variables,  $\theta$  is unknown, and  $g, \sigma_0^2$ , and the  $x_i$  are known. The scores are functions of the  $\theta$ -derivatives of  $g$ :

$$\begin{aligned} S_1 &= \frac{1}{\sigma_0^2} \sum \{y_i - g(\theta_0, x_i)\} \dot{g}(\theta_0, x_i) \\ S_2 &= \frac{1}{\sigma_0^2} \sum \{y_i - g(\theta_0, x_i)\} \{\ddot{g}(\theta_0, x_i) - \dot{g}^2(\theta_0, x_i)\} \\ &\cdot \\ &\cdot \\ &\cdot \end{aligned}$$

with a covariance matrix  $I(\theta_0)$  depending only on  $\sigma_0^2$  and the  $\theta$ -derivatives of the  $g(\theta, x_i)$  at  $\theta_0$ . We wish to determine the vector  $v(y)$  that satisfies  $d\bar{S}_j = 0$  or  $d\tilde{S}_j = 0$  for  $j = 2, \dots, n$ .

The  $y_1, \dots, y_n$  can be expressed linearly in terms of the  $S_1, \dots, S_n$  and are uncorrelated. In terms of the  $y$ 's we can view  $S_1$  as a linear contrast and  $\tilde{S}_2, \dots, \tilde{S}_n$  as linear contrasts orthogonal to  $S_1$ . It follows that the vector  $v(y)$  satisfying  $d\tilde{S}_2 = 0, \dots, d\tilde{S}_n = 0$  is given by the gradient vector of  $S_1$  with respect to the  $y$ 's:

$$v(y) = \{g(\hat{\theta}, x_1), \dots, g(\hat{\theta}, x_n)\} / \{\sum g^2(\hat{\theta}, x_i)\}^{1/2}.$$

The maximum likelihood point is the projection of  $y$  on the mean-vector curve  $\{(g(\theta, x_1), \dots, g(\theta, x_n)) : \theta \in R\}$ , and the equation  $S_1(y, \theta_0) = 0$  determines the orthogonal complement of the tangent vector at the point having  $\theta_0 = \hat{\theta}$ . The curves for conditional inference thus have the following form: each intersects the surface  $S_1(y, \theta_0) = 0$  in a direction parallel to the mean-vector curve at  $\theta_0$ . Equation (2.4) then enables us to use the exact conditional distribution of  $\hat{\theta}$ , rather than a normal approximation to its unconditional distribution.

### 3.4 The general case

In this section we consider inference for  $\theta$  in the presence of nuisance parameters  $\lambda$ , when the dimension of the minimal sufficient statistic  $y$  is greater than the dimension of the parameter  $\phi = (\theta, \lambda)$ . The approach combines the techniques developed for the two special cases in Sections 3.2 and 3.3.

We first assume that the nuisance parameters  $\lambda$  are orthogonal to  $\theta$ , with respect to the Fisher information for the full model  $f(y; \phi)$ . Construction of orthogonal parameters is discussed in Cox and Reid (1987). The exponential tangent model to  $f$ , at  $\phi_0$ , is taken to be

$$\tilde{f}(y; \phi) = f(y; \phi_0) \exp \left\{ \sum_{j=1}^p S_j \delta_1^j / j! + \sum_{j=1}^r U_j (\lambda_j - \lambda_{0,j}) - \tilde{K}(\phi) \right\}, \quad (3.13)$$

where  $r + p = k$ ,  $\delta_1 = \theta - \theta_0$ ,  $S_j = S_j(y; \phi_0)$  is the  $j$ th  $\theta$ -derivative of the log-likelihood as in Section 3.3,  $U_j$  is the score for the  $j$ th component of  $\lambda$ , and  $\tilde{K}(\phi) = \tilde{K}(\phi; \phi_0)$  is the normalizing constant for  $\tilde{f}$ . Although the nuisance parameters were not required to be orthogonal in the pure nuisance parameter case of Section 3.2, the orthogonality seems to be necessary in order to apply the argument of Section 3.3 for the higher order  $\theta$  components. The higher order derivatives with respect to  $\theta$  are then in some sense free of nuisance parameter effects.

The full exponential model corresponding to (3.9) is

$$f(y; \theta_0) \exp \left\{ \sum_{j=1}^p S_j \delta_j + \sum_{j=1}^r U_j (\lambda_j - \lambda_{0,j}) - K(\delta, \lambda) \right\}. \quad (3.14)$$

As in Section 3.3 the linear term  $\delta_1$  can be any  $p$ th order re-expression of  $\theta - \theta_0$ , and we choose the particular parametrization that makes the higher order scores uncorrelated with the first, at  $\phi_0$ . Equivalently, we define

$$\tilde{S}_j(y; \phi_0) = S_j(y, \phi_0) - I_{1j}(\phi_0) I_{11}^{-1}(\phi_0) S_1(y, \phi_0)$$

where  $I_{1j}(\phi_0) = \text{cov}(S_1, S_j; \phi_0)$  and use the conditions

$$d\tilde{S}_j(y; \phi_0)|_{\phi_0 = \hat{\phi}} = 0, \quad j = 2, \dots, p \quad (3.15a)$$

$$d\tilde{U}_j(y; \phi_0)|_{\phi_0 = \hat{\phi}} = 0, \quad j = 1, \dots, r \quad (3.15b)$$

to define the vector field  $\{v(y)\}$ .

These conditions can be rewritten as

$$s_j(\hat{\phi}) \cdot v = 0 \quad j = 2, \dots, p \quad (3.16a)$$

$$u_j(\hat{\phi}) \cdot v = 0 \quad j = 1, \dots, r \quad (3.16b)$$

where  $u_j(\phi) = \nabla U_j(y; \phi)$  and  $s_j(\phi) = \nabla S_j(y; \phi)$  are the  $k \times 1$  gradient vectors for the score functions  $U_j$  and  $S_j$ .

*Example 3.4: Gamma hyperbola (cont.)*

In this case we assume the independent pairs  $(x_{1i}, x_{2i})$  are observations from gamma distributions with rate parameters  $\theta^{-1}$  and  $\theta$ , and shape parameters  $\psi_1$  and  $\psi_2$ . The joint density for a sample of size  $n$  is

$$f(y; \phi) = \Gamma^{-n}(\psi_1)\Gamma^{-n}(\psi_2)\theta^{n\psi_1-n\psi_2} \exp\{(\psi_1-1)y_1 + (\psi_2-1)y_2 - \theta y_3 - \theta^{-1}y_4\}$$

where  $\phi = (\theta, \psi_1, \psi_2)$  and  $y = (y_1, y_2, y_3, y_4) = (\sum \log x_{1i}, \sum \log x_{2i}, \sum x_{1i}, \sum x_{2i})$ . We first orthogonalize  $\psi_1$  and  $\psi_2$  to  $\theta$ ; one solution of the orthogonality equations is obtained from

$$\lambda_1 = \zeta(\psi_1) - \log \theta$$

$$\lambda_2 = \zeta(\psi_2) + \log \theta ,$$

where  $\zeta(\psi) = \Gamma'(\psi)/\Gamma(\psi)$ .

In this orthogonal parameterization the nuisance parameter scores are  $U_1 = \{-n\zeta(\psi_1) + n \log \theta + y_1\} \{\partial \psi_1(\theta, \lambda_1)/\partial \lambda_1\}$  and  $U_2 = \{-n\zeta(\psi_2) - n \log \theta + y_2\} \{\partial \psi_2(\theta, \lambda_2)/\partial \lambda_2\}$ , so the nuisance parameter restrictions (3.15b) specify  $dy_1 = 0$ ,  $dy_2 = 0$ , and are unaffected by the parameter orthogonalization. The two  $\theta$ -scores,  $S_1$  and  $S_2$ , are given by

$$S_1 = n\theta^{-1}(\psi_1 - \psi_2) - y_3 - \theta^{-2}y_4 + (n \log \theta + y_1)\{\theta \zeta'(\psi_1)\}^{-1} \\ + (n \log \theta - y_2)\{\theta \zeta'(\psi_2)\}^{-1} ,$$

and

$$S_2 = -2\theta^{-3}y_4 + y_1[1 + \zeta''(\psi_1)/\{\zeta'(\psi_1)\}^2]\{\theta^2 \zeta'(\psi_1)\}^{-1} \\ + y_2[1 - \zeta''(\psi_2)/\{\zeta'(\psi_2)\}^2]\{\theta^2 \zeta'(\psi_2)\}^{-1} + k ,$$

where  $k = k(\theta, \psi_1, \psi_2)$  does not depend on  $y$ . These scores are computed in the  $(\theta, \lambda_1, \lambda_2)$  parametrization, so  $\psi_1$  and  $\psi_2$  are functions of  $\theta, \lambda_1, \lambda_2$ . The third equation determining the conditional inference curve takes the form

$$\hat{\theta}^2 dy_3 + dy_4(-1 - 2c\hat{\theta}^{-1}) = 0$$

where  $c = c(\hat{\theta}, \hat{\lambda}_1, \hat{\lambda}_2) = I_{12}(\hat{\theta}, \hat{\lambda}_1, \hat{\lambda}_2)/I_{11}(\hat{\theta}, \hat{\lambda}_1, \hat{\lambda}_2)$  is a somewhat cumbersome expression. If  $\hat{\psi}_1 = \hat{\psi}_2 = 1$ , the above expression is the same as that obtained in Example 3.2.

### 3.5 Discussion of parametrization

In this section we discuss in more detail the reason for working with the modified likelihood derivatives  $\bar{S}_2, \dots, \bar{S}_k$ ; i.e. with a version of the primary parameter which is orthogonal to  $(\delta_2, \dots, \delta_k)$  in the embedding exponential model (3.9). For ease of exposition we assume there are no nuisance parameters.

Let  $\Delta$  be the parameter space for the parameters  $\delta_1, \delta_2, \dots, \delta_k$  in the full exponential model (3.9), for the initial parametrization  $\theta$  in which  $\delta_1 = \theta - \theta_0$  and let  $S$  be the sample space for the likelihood derivatives  $S_1(y, \theta_0), \dots, S_k(y, \theta_0)$ . The true model has  $\delta_j = \delta_1^j/j!$  and lies on a one-dimensional curve through  $\Delta$ . This curve is tangent to the  $\delta_1$  axis at the origin; i.e. the tangent to the curve at the origin is defined by  $\delta_2 = 0, \dots, \delta_k = 0$ . In the space  $S$  we are particularly interested in the surface  $S_1 = 0$ , as it corresponds to points  $y$  having  $\hat{\theta}(y) = \theta_0$ . The covariance matrix  $I(\theta_0)$  of  $S_1, \dots, S_k$  at  $\theta_0$  in the true model is the same as the observed or expected information matrix in the embedding exponential model at the origin. The curve defined by  $\{E_\theta S_1(y; \theta_0), \dots, E_\theta S_k(y; \theta_0)\}$  in the true model is tangent to  $I(\theta_0)(\delta_1, 0, \dots, 0)$  at the origin.

Now consider the effect of the  $k$ th order reparametrization  $\bar{\theta} = h(\theta)$  with  $h'(\theta_0) = 1$  that gives the orthogonality described in Section 3.3. The new  $\bar{\delta}_1, \dots, \bar{\delta}_k$  provide alternate coordinates on  $\Delta$  and  $\bar{\delta}_2 = 0, \dots, \bar{\delta}_k = 0$  defines the same tangent to the true model curve at the origin; however  $\bar{\delta}_1 = 0$  now defines a different  $k-1$  plane through the origin. A particular value of  $\theta$  in the true model corresponds to some value of  $\bar{\delta}_1$  in the embedding model together with adjustments provided by the orthogonal parameters  $\bar{\delta}_2, \dots, \bar{\delta}_k$ . On the space  $S$  the new score  $\bar{S}_1 = S_1$  is unchanged and the conditioning defined by  $d\bar{S}_2 = 0, \dots, d\bar{S}_k = 0$  is parallel to the line  $\bar{S}_2 = 0, \dots, \bar{S}_k = 0$ . Because the  $\bar{S}_j$  are defined in the orthogonal parametrization this line is tangent to the curve  $(E_\theta \bar{S}_1, \dots, E_\theta \bar{S}_k)$  at the origin.

For the regression Example (3.3) we observed that the vector  $v(y)$  was parallel to the mean-value curve at the maximum likelihood point. Similar arguments show that this result holds more generally for curved exponential models using the canonical variables.

#### 4. Computational aspects

As indicated in Section 2 the calculation of the conditional distribution  $g(s; \phi)ds$  for inference concerning  $\theta$  can be performed iteratively on the sample space using the vector field  $\{v(y)\}$ . In particular for a point  $y$  we need the vector  $v(y)$  and the expansion rate  $div \{v(y)\}$ .

The methods in Section 3 gave  $k-1$  vectors, say  $a^1(y), \dots, a^{k-1}(y)$ , orthogonal to  $v(y)$ . It is of course straightforward to calculate  $\pm v(y)$  from  $a^1(y), \dots, a^{k-1}(y)$ . In this section we address the calculation of  $div \{v(y)\}$  from derivatives of the vectors  $a^1(y), \dots, a^{k-1}(y)$ .

The orthogonality of the  $a^\alpha(y)$  and  $v(y)$  gives a relation among derivatives. From  $v' a^\alpha = 0$  we obtain

$$a^\alpha{}' V + v' A^\alpha = 0 \tag{4.1}$$

by differentiation where

$$V = \frac{\partial v(y)}{\partial y'} \quad , \quad A^\alpha = \frac{\partial a^\alpha}{\partial y'} \quad . \tag{4.2}$$

The derivative of  $v(y)$  in the direction  $a^\alpha/|a^\alpha|$  is  $Va^\alpha/|a^\alpha|$ . The component of this in the direction  $a^\alpha/|a^\alpha|$  is  $a^\alpha \nabla a^\alpha/|a^\alpha|^2$ . This gives the expansion rate associated with the direction  $a^\alpha/|a^\alpha|$ :

$$d_\alpha(y) \equiv a^\alpha \nabla a^\alpha/|a^\alpha|^2 . \quad (4.3)$$

Alternatively, the derivative of  $a^\alpha(y)$  in the direction  $a^\alpha/|a^\alpha|$  is  $A^\alpha a^\alpha/|a^\alpha|$  and the component of this in the direction  $v(y)$  is  $v \nabla A^\alpha a^\alpha/|a^\alpha|$ . Expressing this as a proportion of  $|a^\alpha|$  and with change of sign gives the separation rate in the direction  $a^\alpha/|a^\alpha|$ :

$$d_\alpha(y) = -v \nabla A^\alpha a^\alpha/|a^\alpha|^2 . \quad (4.4)$$

Formula (4.1) shows the equivalence of the two derivations: the latter is suitable for direct calculation from  $a^\alpha(y)$ . The summation  $D_1(y) = \sum d_\alpha(y)$  gives the expansion rate attributable to divergence in the directions  $a^1, \dots, a^{k-1}$ . It does not however include the effect of change in the relative directions of the vectors  $a^\alpha$ .

For any matrix  $B = (b^1 \dots b^k)$  define

$$skew B = \frac{|B|}{\prod_1^k |b^\alpha|} \quad (4.5)$$

as the ratio of the area of the parallelogram formed by the column vectors of  $B$  to the area of a rectangular parallelogram formed by an orthogonal set of vectors with lengths  $|b^1|, \dots, |b^k|$ .

Now let  $A = (a^1 \dots a^{k-1} a^k)$  where  $a^k$  is  $v$  and consider a parallelogram  $P$  in which the vectors  $a^\alpha$  give the perpendicular separation of opposite faces:

$$P = \{y : 0 < a^\alpha \nabla y < |a^\alpha|^2, \alpha = 1, \dots, k\} . \quad (4.6)$$

For the divergence calculations we need the volume of  $P$  as a proportion of the volume  $\Pi|a^\alpha|$  of a rectangular parallelogram  $Q$  with the same separation of faces: call this proportion  $\rho$ . Under the mapping  $y = (A)^{-1}z$ ,  $z = A \nabla y$  the parallelogram  $P$  is mapped to

$$P^* = \{z : 0 < z_\alpha < |a^\alpha|^2, \alpha = 1, \dots, k\}$$

which has volume  $\Pi|a^\alpha|^2$  and  $Q$  is mapped to some  $Q^*$  which has volume  $\Pi|a^\alpha|/|A|$ . Thus

$$\rho = \frac{\Pi|a^\alpha|^2/|A|}{\Pi|a^\alpha|} = \frac{\Pi|a^\alpha|}{|A|} = \frac{1}{skew A} .$$

It follows that the rate of expansion attributable to change in relative direction of the vectors  $a^\alpha$  is  $D_2(y) = \frac{d}{dv(y)} \log(1/skew A(y)) = -\frac{d}{dv(y)} \log skew A(y)$ . This gives the divergence

$$div \{v(y)\} = D_1(y) + D_2(y) . \quad (4.8)$$

For the numerical calculations however there is no need to differentiate and then numerically integrate the skew component. Accordingly we obtain

$$c(y) = \frac{1}{skew A(y)} \exp\left[\int_0^{s(y)} \sum_{\alpha=1}^{k-1} d_\alpha\{y(t)\} dt\right] .$$

## 5. Discussion

We have approached conditional inference for a real parameter in terms of local likelihood properties of a possible conditional distribution rather than the usual marginal properties of the conditioning variable. For the case where the dimension of the minimal sufficient statistic is equal to the parameter dimension we were able to choose a preferred conditioning direction at each sample point entirely in terms of a local property: that the change in likelihood should have a zero derivative with respect to the nuisance parameters at the maximum likelihood estimate. For the case without nuisance parameters we chose the direction so that the likelihood change has kth-order agreement with the parametrization that is orthogonal to the remaining components parameter in the full exponential tangent model. For the more general case the preceding methods were combined but a preliminary calculation of the orthogonalized nuisance parameters was invoked so that the higher partial derivatives with respect to  $\theta$  were in a sense free of the nuisance parameters.

For the case of a real parameter  $\theta$  without nuisance parameters, (2.7) gives the conditional density for  $\hat{\theta}$  as

$$k(\theta)C(y)f(y; \theta) |j(\hat{\theta})|^{1/2}d\hat{\theta} \quad (5.1)$$

for  $\hat{\theta}$ , conditional on an arbitrary vector field  $\{v(y)\}$ , where

$$C(y) = e^{\int^s \text{div } v(y) ds'} \left| \frac{dS_1(y; \theta)}{dv(y)} \right|_{\hat{\theta}}^{-1} |j(\hat{\theta})|^{1/2} .$$

This also is suitable for iterative computation, once any vector field  $V$  is chosen.

We now compare (5.1) with Barndorff-Nielsen's (1983) formula for the distribution of the maximum likelihood estimate. In the real parameter case this formula is

$$\begin{aligned} &= k(\theta) \frac{f(y; \theta)}{f(y; \hat{\theta})} |j(\hat{\theta})|^{1/2}d\hat{\theta} \\ &= k(\theta) \frac{1}{f(y; \hat{\theta})} f(y; \theta)|j(\hat{\theta})|^{1/2}d\hat{\theta} . \end{aligned} \quad (5.2)$$

This approximates the conditional distribution of  $\hat{\theta}$ , given an exact or approximate ancillary, which needs to be determined in applications.

By making a simple choice of function  $C(y)$  in (5.1) we obtain Barndorff-Nielsen's formula (5.2):  $C(y) = 1/f(y; \hat{\theta})$ . This choice can be interpreted in the following manner: in the

parametrization  $\eta = \int^{\theta} |j(\theta)|^{1/2}d\theta$ , which gives the differential  $d\hat{\eta} = |j(\hat{\theta})|^{1/2}d\hat{\theta}$ , the likelihood function has unit curvature at the maximum likelihood estimate and in this parametrization the choice  $C(y) = 1/f(y; \hat{\theta})$  gives a density function  $g(\hat{\eta}; \eta)$  that has a constant maximum value,  $g(\eta; \eta)$ .

A transformation model has the property of a constant maximum value, in the constant information parametrization  $\eta$ . This property thus determines the function  $C(y)$  in (5.1) and formula (5.2) is thus exact for such models (Barndorff-Nielsen, 1980).

For exponential models it is known that the saddlepoint approximation coincides with Barndorff-Nielsen's formula, and is exact only for the two location cases, normal and log-gamma, and the inverse Gaussian. It follows that the constant maximum density property in the constant information parametrization occurs only for these three models in the exponential family. More generally, the constant maximum-density property is locally a large sample limit, and formula (5.2) is then an asymptotic result. One advantage of (5.2) is that it can be used for multidimensional  $\theta$ , if an exact or approximate ancillary can be found.

## Acknowledgement

We gratefully express our appreciation to Prof. Dipak K. Sen for valuable advice and suggestions concerning differential geometry issues.

## References

- Amari, S.-I. (1982). Geometric theory of asymptotic ancillarity. *Biometrika* 69, 1-17.
- Barndorff-Nielsen, O.E. (1980). Conditionality resolutions. *Biometrika* 67, 293-310.
- Barndorff-Nielsen, O.E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika* 70, 343-65.
- Berger, J. and Wolpert, R. (1985). *The Likelihood Principle*. Institute of Math. Statist.: Hayward.
- Birnbaum, A. (1962). On the foundations of statistical inference (with discussion). *J. Amer. Statist. Assoc.* 57, 269-306.
- Cox, D.R. (1980). Local ancillarity. *Biometrika* 67, 273-8.
- Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics*. London: Chapman & Hall.
- Cox, D.R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *J. Royal Statist. Soc. B* 49, 1-39.
- Evans, M., Fraser, D.A.S. and Monette, G. (1986). On principles and arguments to likelihood. *Canad. J. Statist.* 14, 181-200.
- Fisher, R.A. (1934). Two new properties of mathematical likelihood. *Proc. R. Soc. A* 144, 285-307.
- Fraser, D.A.S. (1979). *Inference in Linear Models*. McGraw-Hill: New York.
- Fraser, D.A.S. and Massam, H. (1985). Conical tests: observed level of significance and confidence regions. *Statistical Papers* 26, 1-18.
- Fraser, D.A.S. and Reid, N. (1987). Fibre analysis and conditional inference. *Proceedings of the 2nd Pacific Area Statistical Conference*. ,to appear.
- Hinkley, D.V. (1980). Likelihood as approximate pivotal. *Biometrika* 67, 287-92.
- Lehmann, E.L. (1986). *Testing Statistical Hypotheses*. New York: Wiley.
- Skovgaard, I. (1987). Saddlepoint expansions for directional test probabilities. *J. Royal Statist. Soc. B* ,to appear.