FIBRE ANALYSIS AND CONDITIONAL INFERENCE

D.A.S. Fraser and N. Reid

This paper examines conditional procedures from a viewpoint that focuses on distributional and inference properties of the conditional distribution itself. We address inference for a real parameter in the presence of nuisance parameters, and one-manifolds or fibres are examined generally as a basis for conditional inference. For these, local or differential conditions can be used to achieve a reduction in nuisance-parameter and model-curvature effects. Computer implementation requires only one-dimensional iteration routines and thus lends itself to graphical display. Other approaches to conditional inference are briefly discussed in the introduction.

1. INTRODUCTION

Conditioning as part of the inference process is usually justified in terms of properties of the conditioning variable. Fisher [13], in the context of the location and location-scale models, argued that a configuration statistic defines the appropriate reference set for conditional probability calculations. The ancillarity or fixed-distributional property of this conditioning variable was then taken rather generally as the operating property for such conditioning, and led to the formulation of the conditionality principle (Basu [5],[6], Birnbaum [7],[8]). Recently Evans, Fraser and Monette [12] showed that the conditionality principle implies the strong likelihood principle, which in turn suggests that the conditionality principle is not particularly compelling, at least in the common formulations.

Fisher [13] was probably also implicitly recognizing certain physical aspects of the conditioning variable. This type of physical or experimental support for conditional procedures was expanded in various directions in Basu [6], Fraser [15] and Kalbfleisch [23], and in terms of necessary conditioning of error models and data in Fraser [14],[16].

Recent approaches to conditioning have reemphasized the ancillarity or near-ancillarity of the conditioning variable; see for example, Amari [1],[2], Barndorff-Nielsen [3],[4], Cox [9], Hinkley [22].

In Fraser and Massam [18] a simple hypothesis is tested by conditioning on the data direction as representing a direction for parameter departure and then examining the magnitude of the departure. This is developed under the assumption that the sample-space dimension is equal to the parameter-space dimension; an extension to a more general case involving nonlinear regression may be found in Fraser and Massam [19].

The approach outlined in this paper focusses directly on properties of the conditional distribution. In this respect it is closer in spirit to McCullagh [24]. The objective is to consider to what extent it is possible to construct a one-dimensional conditional
distribution that is sensitive to the parameter of interest and insensitive to the nuisance parameters. In Section 2 we show that choosing a variable to condition on is equivalent to choosing a vector field on the sample space. The vector field defines a one-dimensional sample contour, or fibre. In Section 3, we show that the vector field may be chosen so that likelihood change along the fibre is free of the nuisance parameter, in a neighbourhood of the maximum likelihood estimate. Inference for the ratio of two exponential means is discussed as an illustrative example. We assume that the dimension of the full parameter \( \phi \) and the minimal sufficient statistic are the same. In the context of independent sampling, this means that attention is restricted to full exponential families. However the parameter of interest need not be a component of the canonical parameter. Extensions to more general settings are briefly considered in Section 4.

2. ONE-MANIFOLDS AND CONDITIONAL INFERENCE

Consider a variable \( y \) with sample space \( \mathbb{R}^n \), a parameter \( \phi \) with parameter space \( \mathbb{R}^k \) and a continuous statistical model \( f(y|\phi)dy \). We write \( \phi = (\theta, \lambda) \) where \( \theta \) is a real parameter of interest and \( \lambda \) is the related nuisance parameter. We will assume that \( n = k \) and \( y \) is the minimal sufficient statistic. The case \( n > k \) is briefly discussed in Section 4.

A one-dimensional conditional procedure can be obtained from a one-to-one change of variable \( (T(y), s(y)) \), where \( T \) has dimension \( k-1 \), by examining the conditional distribution of \( s \) given \( T \). The variable \( T \) generates a partition \( P \) of \( \mathbb{R}^k \) into one dimensional fibres or one-manifolds and the variable \( s \) locates a point \( y \) on its one-manifold. In general, if we have a partition \( P \) of \( \mathbb{R}^k \) into one-manifolds and a real function \( s(y) \) that locates \( y \) on its one-manifold, we will call \( P \) a one-dimensional conditional procedure. By this we mean that the conditional distribution of \( s \), given its one-manifold, will be used for inference concerning the parameter \( \theta \).

A conditional procedure \( P \) may not contain all the information concerning the parameter \( \theta \); just as of course an unconditional procedure may not contain all the information. We accept the loss of information in the conditioning variable from a somewhat pragmatic viewpoint: the one-dimensional conditional distribution may be easy to use for direct inference about \( \theta \); and conditioning removes the need for high dimensional integration that is typically needed if an appropriate marginal model is to be examined. A further point, however, arises, as to whether the loss of information in the marginal model is the only concern or whether there are additional reasons why the conditional model cannot be interpreted at face value. Our viewpoint in this paper ignores this latter concern.

Our approach in effect is to consider all one dimensional conditional distributions and seek a preferred choice on the basis of inference properties of the conditional distribution itself.

In terms of the coordinates \( (T, s) \) the conditional model can be written

\[
g(s|T;\phi)ds = k(T;\phi)f(y;\phi)c(y)ds
\]

where \( y = y(T,s) \), \( k(T;\phi) \) is the reciprocal of the marginal density for \( T \) and \( c(y) = |\partial y/\partial (T,s)| \) is the Jacobian. We now describe how the conditional density can be computed without actually specifying the function \( T \).

The variable \( s(y) \) given \( T \) records the position of \( y \) on any one-manifold in the partition \( P \). Differentiating with respect to \( s \) (along the one-manifold) gives the tangent vector
\[ v(y) = \frac{\partial y}{\partial s}; \quad (2.2) \]

If \( s(y) \) is chosen to measure arc length, then \( v(y) \) has unit length. Thus a conditional procedure \( P \) generates a unit vector field \( \{v(y)\} \) on \( \mathbb{R}^k \). Conversely, a smooth unit vector field \( \{v(y)\} \) integrates to give a conditional procedure \( P \). Thus we can represent a conditional procedure either by its partition \( P \) or by its unit vector field \( V = \{v(y)\} \).

In terms of the vector field \( V \) the conditional model can now be written

\[ g(s|T;\phi)ds = k(T;\phi)f(y;\phi)J(r(\cdot);s)ds \quad (2.3) \]

where

\[ r(s) = \text{div} v(y) = \sum_i \frac{\partial v_i(y)}{\partial y_i} \quad (2.4) \]

is the divergence of the vector field, and

\[ J(r(\cdot),s) = \exp\left\{ \int_0^s r(u)du \right\} \quad (2.5) \]

is the accumulated expansion determined by the logarithmic rate \( r(s) \).

The alternative formula (2.3) has substantial computational advantages. From any initial point \( y^0 \) which in applications would reasonably be the observed data point, the equation (2.2) can be solved iteratively giving \( y_{i+1} = y_i + \delta v(y_i) \) where \( y_i = y_i(T,s_i); s_{i+1} = s_i + \delta; \) the solution would be obtained for both positive and negative steps from the data point labelled \( s = 0 \). Also as part of the iteration, the Jacobian factor \( J_{i+1} = J_i(1 + \delta \text{div} v(y_i)) \) would be calculated, and for various \( \phi \) values of interest the product \( J_i f(y_i;\phi) \) would give the density \( g(s|T;\phi) \), except for the norming constant \( k(T,\phi) \), which can be determined from the cumulated sum.

We note that inference can be relatively straightforward from the one-dimensional model \( g(s|\theta,\lambda)ds \). For this we assume that the model as obtained is stochastically monotone in \( \theta \) and relatively insensitive to \( \lambda \). The observed level of significance for a test of the null hypothesis \( \theta = \theta_0 \) is

\[ p(s^0,\theta_0) = 2 \cdot \int_{s^0}^{s^*} \tilde{g}(s|\theta_0)ds \]

where \( \int_{\tilde{s}^0}^{s^*} \) gives the smaller of \( \int_{\tilde{s}^0}^{s^*} \) and \( \int_{s^*}^{\infty} \); this is the conditional probability of being farther from the median of \( g(s|\theta_0) \) than \( s_0 \), given the direction from the median, and is a special case of the conical test of Fraser and Massam [19]. From this testing procedure, a \( 1-\alpha \) confidence interval \((\theta^L, \theta^U)\) is obtained from the solutions \( \theta^L, \theta^U \) to the equation \( p(s^0,\theta) = \alpha \).

Obtaining a \( g(s|\theta,\lambda) \) that is approximately insensitive to \( \lambda \) is discussed in the next section. In practice the sensitivity of the procedure to \( \lambda \) would need to be checked directly, by perturbing \( \lambda \) slightly from the maximum likelihood value \( \lambda \) that might reasonably be used for initial calculations.
3. CHOICE OF VECTOR FIELD $V$

We now consider the choice of a vector field $V$ to achieve some desirable inference properties for the conditional density $g(s\mid T;\phi)$. Equation (2.3) provides the means for linking these properties to related properties of the original density $f(y;\phi)$.

We assume that log-density is differentiable on both the sample and parameter spaces and examine

$$\ln g(s \mid T;\phi) = \ln k(T;\phi) + \ln f(y;\phi) + \ln c(y). \tag{3.1}$$

Let

$$\tilde{l}(s \mid T;\phi) = \ln \frac{g(s \mid T;\phi)}{g(s \mid T;\phi_0)}, \quad K(T;\phi) = \ln \frac{k(T;\phi)}{k(T;\phi_0)},$$

$$l(y;\phi) = \ln \frac{f(y;\phi)}{f(y;\phi_0)} \tag{3.2}$$

give likelihood relative to some $\phi_0$; then

$$\tilde{l}(s \mid T;\phi) = K(T;\phi) + l(y;\phi) \tag{3.3}$$

Of particular interest will be the changes in the log-likelihood along the one-manifold being considered for inference. For this we take differentials in the sample space,

$$dl(y;\phi) = \frac{\partial l(y;\phi)}{\partial y_1} dy_1 + \ldots + \frac{\partial l(y;\phi)}{\partial y_n} dy_n \tag{3.4}$$

$$\tilde{d}l(s;T;\phi) = \frac{\partial l(s;T;\phi)}{\partial s} ds,$$

and from (3.3) obtain

$$dl(s,T;\phi) = dl(y;\phi)_{|_{dy = v(y)}} ds \tag{3.5}$$

where

$$dl(y;\phi)_{|_{dy = v(y)}} ds = \frac{\partial l(y;\phi)}{\partial y_1} v_1(y) ds + \ldots + \frac{\partial l(y;\phi)}{\partial y_n} v_n(y) ds$$

is the differential $dl$ in (3.4), restricted to the direction $v(y)$ being considered for the conditional one-manifold. Equation (3.5) shows that the sample space change in log-likelihood is the same for the conditional and the overall density.

We note in general that the likelihood differential $dl(y;\phi)$ in (3.4) is the minimal sufficient statistic at the point $y$; thus the functions $\partial l(y;\cdot)/\partial y_1, \ldots, \partial l(y;\cdot)/\partial y_n$ span a space of dimension equal to the local dimension of the minimal sufficient statistic. However, by assumption here $y$ is minimal sufficient so that these $n=k$ functions of $\phi$ are linearly independent.

If the conditional density is free of $\lambda$, then (3.4) is free of $\lambda$ and conversely. Thus the vector field will satisfy

$$dl(y;\theta,\lambda)_{|_{dy = v(y)}} ds = \lambda - \text{free} \tag{3.6}$$

Typically a conditional density fully free of the nuisance parameter will not be attainable. (An exception is if $\theta$ is a component of the canonical parameter for the exponential family.) It is reasonable then to require first order freedom in the neighbourhood of the maximum likelihood estimate:

$$dl_y(y;\theta,\lambda)_{|_{\hat{\theta},\hat{\lambda}}} = 0 \quad i = 1,\ldots,r \tag{3.7}$$
for \( \text{d}y = v(y) \text{d}s \) where

\[
I_i(y; \theta, \lambda) = \frac{\partial}{\partial \lambda_i} l(y; \theta, \lambda) \tag{3.8}
\]

is the score for the \( i \)th nuisance parameter, and \( r = k-1 \). The equations (3.7) provide \( k-1 \) linear constraints in the tangent direction \( v(y) \text{d}s \) at the point \( y \).

**Example 3.1: Full exponential family.**

Suppose that the density of \( y \) is a full exponential model, and that the parameter of interest is a component of the canonical parameter. Then

\[
f(y; \phi) = \exp\{\theta y_1 + \lambda^T y_2 - \psi(\theta, \lambda) - d(y)\},
\]

and the conditional distribution of \( y_1 \) given \( y_2 = (y_2, \ldots, y_k) \) would generally be used for inference about \( \theta \). In the notation outlined above, \( l(y, \theta, \lambda) = (\theta - \theta_0) y_1 + (\lambda - \lambda_0)^T y_2 - \psi(\theta, \lambda) + \psi(\theta_0, \lambda_0) \), \( I(\theta, \lambda) = y_1 - \psi_1(\theta, \lambda) \), and the vector field \( V \) is determined by solving the \( k-1 \) equations \( \text{d}y_i = 0, \ i = 2, \ldots, k \).

That is we simply recover the usual conditional procedure. This result motivates the extensions discussed in Section 4.

**Example 3.2: Ratio of exponential means.**

Let \( \{x_{1i}, x_{2i}, i = 1, \ldots, n\} \) be a sample of size \( n \), where \( x_{1i} \) and \( x_{2i} \) are independent exponentials with means \( \lambda^{-1} \) and \( (\theta \lambda)^{-1} \), respectively. The joint density of \( (y_1, y_2) \), where \( y_1 = \sum x_{1i}, \ y_2 = \sum x_{2i} \) is

\[
\Gamma^{-2}(n)^{\theta^2} \lambda^{2n} y_1^{\lambda^T y_2 - \lambda} \exp\left\{ - (\lambda y_1 + \theta \lambda y_2) \right\}. \tag{3.9}
\]

Note that the statistic \( y_1/y_2 \) provides exact inference for \( \theta \); a special case of the general theory for location models. In fact the marginal distribution of \( y_1/(\theta y_2) \) is \( F \) with \( 2n \) degrees of freedom. This would no doubt be preferred in practice for this example, but the purpose here is to see how the general conditional procedure compares to the prescribed marginal inference. From (3.9) we have

\[
l = n \ln \theta - n \ln \theta_0 + 2n \ln \lambda - 2n \ln \lambda_0 - y_1(\lambda - \lambda_0) - y_2(\theta \lambda - \theta_0 \lambda_0),
\]

\[
\frac{dl}{\theta} = -dy_1(\lambda - \lambda_0) - dy_2(\theta \lambda - \theta_0 \lambda_0), \quad \text{and}
\]

\[
\frac{dl}{\lambda} \big|_{\theta = \theta_0, \lambda = \lambda_0} = -dy_1 - \hat{\lambda} \text{d}y_2.
\]

Since \( \hat{\theta} = y_1/y_2 \), the differential equation to be solved is \( \text{d}y_1 + \text{d}y_2(y_1/y_2) = 0 \), giving \( y_1 y_2 = c \) as the equation for the one-manifold in the space of \( (y_1, y_2) \). The function \( (y_1 y_2)^{-1} \) is in fact the maximum likelihood estimate of \( \theta^2 \lambda \), an orthogonalized expression of the nuisance parameter, orthogonalized with respect to the observed (or expected) information matrix (Cox and Reid, [10]).

To compare the conditional inference derived from fixing \( y_1 y_2 \) to the marginal inference discussed above, we write \( \eta = \log \theta, \ z = \log(y_1/y_2) - \eta \), and \( u = y_1 y_2 \). Then

\[
g(z \mid u) \text{d}z = k \exp\left\{ -(e^{z/2} + e^{-z/2}) \kappa \right\} \text{d}z,
\]
where $\kappa = (\lambda^2 \partial_0)^{1/2}$. The location parameter for the conditional density is $\eta = \log \theta$ and is free of the nuisance parameter $\lambda$, whereas the precision index $\kappa$ depends on the orthogonalized nuisance parameter and the corresponding maximum likelihood estimate $u^{-1} = (y_1y_2)^{-1}$.

4. MORE GENERAL FAMILIES

To construct a one-manifold by conditioning when the dimension of the sufficient statistic is greater than that of the parameter requires adding further constraints to the system of differential equations (3.7). (We continue to assume that the parameter of interest is scalar.) Suppose the dimensions of $\gamma$ and $\lambda$ are, respectively, $n$ and $r$, where $n > r+1$. If $n \geq r + r(r+1)/2 + 1$, second order freedom from the nuisance parameter can be obtained by adding the constraints $\partial^2_0(\theta, \lambda)|_{\theta, \lambda} = 0$ to the equations (3.7), where $t_{ij} = \partial^2_0/\partial \gamma_i \partial \lambda_j$.

An alternative possibility is to approximate the density $f$ by a curved exponential family using a Taylor series expansion of $\log f$ around a fixed point $\theta_0$ (Efron, [11]). Construction of a conditional inference model from this approximating family is investigated in Fraser and Reid [21]. Local conditions used to supplement equations (3.7) are derived, in order that the conditional density is as free from the curvature effects as possible.

REFERENCES


University of Toronto
University of Toronto