211:1999may22

Strong matching of frequentist and Bayesian inference

BY D.A.S. Fraser and N. Reid
Department of Statistics
University of Toronto
Toronto, Canada, M5S 3G3

SUMMARY

We define strong matching to be the equality of frequentist and Bayesian tail probabilities for the testing of scalar interest parameters; for the special case of a location model strong matching is obtained for any interest parameter linear in the location parameters. A brief survey of methods for choosing a prior, of principles relating to the Bayesian paradigm, and of confidence and related procedures leads to the development of a general location reparameterization. This is followed by a brief survey of recent likelihood asymptotics which provides a basis for examining strong matching with general continuous statistical methods. It is then shown in a general context that a flat prior with respect to the general location parameterization gives strong matching for linear parameters. For nonlinear parameters strong matching requires an adjustment to the flat prior which is a nuisance information determinant standardized by the corresponding nuisance information determinant for a related linear parameterization. Pivotal quantities are then used to generate second order preferred priors for linear parameters. A concluding section describes a confidence, fiducial, or default Bayesian inversion relative to the location parameterization. This gives a means to adjust confidence or related intervals by means of a personal prior taken relative to the flat prior in the location parameterization.

1. INTRODUCTION

We examine the agreement between frequentist and Bayesian methods using recent methods from higher order likelihood asymptotics. As part of this we present a definition of strong matching of the two methods; this is recorded at the end of this section. Theory then developed produces the appropriate prior for assessing the full parameter and also the appropriate adjustment to this for assessing a component scalar parameter of interest

In Section 2 we examine location models and show that a flat prior in terms of the location parameter provides strong matching for the family of parameters that are linear in the location parameterization. We also show that this natural flat prior does not in general give strong matching for other parameters that can be described as curved in the location parameterization.

In Section 3 we review the familiar choices for a default prior density. In most cases these do not provide strong matching for linear parameters: this is an issue beyond the need to specially target the prior in the case of non linear component parameters of interest.

In Section 4 we examine the Bayesian paradigm and the closely related Strong Likelihood Principle. This leads to the definition of a less restrictive and perhaps more realistic Local Inference Principle.

In Section 5 we examine confidence and fiducial methods and compare these with the Bayesian inversion method. We note that confidence and fiducial differ with respect to a minor procedural technicality and that the distinction in the present context is of negligible importance. We also note that in the presence of an appropriate default prior, the Bayesian method permits the modification of confidence or fiducial presentations to accommodate relative reemphasis based on a personal or communal prior defined relative to the default prior.

In Section 6 we show that a location reparameterization exists for continuous models under wide generality. The reparameterization is straightforward for a scalar full parameter, but is computationally more difficult for vector parameters.

Section 7 gives a brief summary of recent likelihood asymptotics. This builds on a long succession of asymptotic results, from the simple case of a scalar parameter exponential model to a general asymptotic model with fixed dimension parameter. The important recent extension for this latter case is then from a fixed dimension variable to the truly asymptotic case with an increasing dimension variable. These results lead to a general p-value for testing a scalar parameter: the mechanics of this involves the reduction to the finite dimension case by conditioning; a marginalization from this to obtain a scalar measure of departure from a hypothesized value for the interest parameter; and finally a third order accurate approximation for the resulting p-value. The accuracy for these is typically high even for very small samples. A notion then of how a variable measures a parameter leads to the uniqueness of the results as based on the full data. Somewhat related asymptotic results lead to a posterior right tail value for the Bayesian context.

Section 8 examines strong matching for a scalar parameter in the context of a general continuous model. It is seen that this matching is available immediately for parameters that are linear in the location reparameterization of Section 7. Then for nonlinear parameters a modification of the prior is typically needed, as anticipated from Section 2. The modification is by a weight function or relative prior given as the ratio of the nuisance information determinant for the curved parameter to the corresponding nuisance information determinant for the linear parameter.

Section 9 examines to what degree these results can be obtained from pivotal quantities.

Section 10 presents a third order location parameter pivotal quantity and discusses how this can be used to determine a confidence, fiducial, and default Bayesian posterior. This posterior has direct confidence type interpretation and yet can be adjusted by relative prior to give subjective or communal posteriors.

If the frequentist context gives a p-value $p(\psi)$ for assessing a value ψ for an interest parameter component $\psi(\theta)$ and if the Bayesian posterior analysis gives a survivor type assessment $s(\psi)$ of the same interest parameter value $\psi(\theta) = \psi$, then the equality of $p(\psi)$ and $s(\psi)$ is called strong matching.

2. DEFAULT PRIOR FOR A LOCATION MODEL: AN OBVIOUS CHOICE

Consider a scalar y that directly measures a scalar θ with error density $f(e)$; the model is $f(y-\theta)$. To test a value $\theta = \theta^0$ with data y^0, the basic frequency calculation gives the probability position of the data under the hypothesis as the p value

$$p(\theta^0) = \int_{-\infty}^{y^0} f(y - \theta^0) dy \; ; \tag{2.1}$$

this can be adjusted to present significance in one or other direction or in either direction from what is expected under θ^0. Alternatively in the Bayesian framework a natural choice for default prior is the flat or uniform prior $\pi(\theta) = c$. The corresponding posterior density is then $\pi(\theta|y^0) = f(y^0 - \theta)$, and the corresponding survivor function recording posterior probability greater than θ^0 is

$$s(\theta^0) = \int_{\theta^0}^{\infty} f(y^0 - \theta) d\theta \; . \tag{2.2}$$

Clearly, we have strong matching $p(\theta^0) = s(\theta^0)$, and each records $P(e \le e^0) = \int_{-\infty}^{e^0} f(e)de$ where $e^0 = y^0 - \theta^0$; for this, the frequentist variable is $e = y - \theta^0$ and the Bayesian variable is $e = y^0 - \theta$ but the distribution for the e in each case is just the given $f(e)$.

Now consider a vector variable y where each coordinate measures θ with joint error density $f(e)$; the model is $f(y - \theta 1)$ and the error density can describe either independency or dependence, without complication. To test a value $\theta = \theta^0$ the usual frequentist calculation derived from Fisher is conditional along the line $y^0 + \mathcal{L}(1)$ parallel to the span $\mathcal{L}(1)$ of the one vector 1; the p value is

$$p(\theta^0) = \int_{-\infty}^{\hat{\theta}^0} L(\theta^0 - \hat{\theta} + \hat{\theta}^0) d\hat{\theta}$$

where $L(\theta) = cf(y^0 - \theta 1)$ is the observed likelihood and the c in the integral is taken to be the normalizing constant. For the Bayesian case the flat default prior $\pi(\theta) = c$ is again a natural choice, giving the posterior density $\pi(\theta|y^0) = cL(\theta)$ with survivor function $s(\theta^0)$ at $\theta = \theta^0$,

$$s(\theta^0) = \int_{\theta^0}^{\infty} L(\theta) d\theta \ .$$

Again we have strong matching $p(\theta^0) = s(\theta^0)$, and note that the two probabilities each record $P(\tilde{e} \leq \tilde{e}^0)$ where $\tilde{e}^0 = \hat{\theta}^0 - \theta^0$; for this the frequentist variable is $\tilde{e} = \hat{\theta} - \theta^0$ and the Bayesian variable is $\tilde{e} = \hat{\theta}^0 - \theta$ but the distribution for $\tilde{e}$ in each case is the same, as given by $L(-\tilde{e} + \hat{\theta}^0)$.

Now consider a vector parameter θ in the special context of a location model $f(y - \theta)$. For a scalar parameter component say θ_1, a corresponding variable is y_1 and it has distribution free of $\theta_2, \ldots, \theta_p$. The corresponding p value is

$$\begin{aligned} p(\theta_1^0) &= \int_{-\infty}^{y_1} f_1(y_1 - \theta_1) dy_1 \\ &= \int_{-\infty}^{\hat{\theta}_1^0} \left\{ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} L(\theta^0 - \hat{\theta} + \hat{\theta}^0) d\hat{\theta}_2 \ldots d\hat{\theta}_p \right\} d\hat{\theta}_1 \\ &= \int_{\hat{\theta}_1 \leq \hat{\theta}_1^0} L(\theta^0 - \hat{\theta} + \hat{\theta}^0) d\hat{\theta} \\ &= \int_{e_1 \leq e_1^0} L(-e + \hat{\theta}^0) de \end{aligned} \tag{2.3}$$

with $e = \hat{\theta} - \theta^0$, $\theta^0 = (\theta_1^0, \theta_2, \ldots, \theta_p)'$, and first coordinate value $e_1^0 = \hat{\theta}_1^0 - \theta_1^0$; the constant c is taken to be the normalizing constant. For the Bayesian case the flat or uniform prior $\pi(\theta) = c$ is a natural choice giving the posterior $\pi(\hat{\theta}|y^0) = L(\theta)$. This leads to the posterior survivor value for θ_1 at θ_1^0:

$$\begin{aligned} s(\theta_1^0) &= \int_{\theta_1 \geq \theta_1^0} L(\theta) \\ &= \int_{e_1 \leq e_1^0} L(-e + \hat{\theta}^0) de \end{aligned} \tag{2.4}$$

with derived variable $e = \hat{\theta}^0 - \theta$ and first coordinate value $e_1^0 = \hat{\theta}_1^0 - \theta_1^0$. Again we have $p(\theta_1^0) = s(\theta_1^0)$ with strong matching of frequentist and Bayesian results.

The more general location model case with $\dim y > \dim \theta$ can be put in the familiar regression form $f(y - X\theta)$. The frequentist analysis conditions on the residuals, $y - X\hat{\theta} = y^0 - X\hat{\theta}^0$, and derives the following conditional density for $\hat{\theta}$,

$$L(\theta - \hat{\theta} + \hat{\theta}^0) = L(-e + \hat{\theta}^0) \ ,$$

where $L(\theta 0 = cf(y - X\theta)$, $e = \hat{\theta} - \theta$, and the constant c is chosen to norm the distribution. First consider a linear parameter $\psi = \Sigma a_i \theta_i = a'\theta$. The frequentist p-value for testing $\psi = \psi^0$ has the form

$$p(\psi^0) = \int_{\tilde{e} \le \tilde{e}^0} L(-e + \hat{\theta}^0) de$$

with derived variable $\tilde{e} = \hat{\psi} - \psi^0$ and value $\tilde{e}^0 = \hat{\psi}^0 - \psi^0$. The Bayesian survivor value for $\psi = \psi^0$ is

$$s(\psi^0) = \int_{\tilde{e} \le \tilde{e}^0} L(-e + \hat{\theta}^0) de$$

with variable $\tilde{e} = \hat{\psi}^0 - \psi$ and value $\tilde{e}^0 = \hat{\psi}^0 - \psi^0$. Again we have agreement, $p(\psi) = s(\psi)$, and thus strong matching of the frequentist and Bayesian calculations.

In summary: For any scalar-parameter location location model, the flat prior $\pi(\theta) = 1$ gives strong matching for all parameters; and for any vector-parameter location model, the flat prior gives strong matching for any linear parameter, linear in the location parameters. This leaves open the possibilities for nonlinear parameters. However a simple concluding example shows that the flat prior does not in general give strong matching for nonlinear parameters.

Consider (y_1, y_2) with mean (θ_1, θ_2) and a standard normal error distribution. With flat prior the posterior distribution for (θ_1, θ_2) has mean (y_1^0, y_2^0) and standard normal error. The scalar parameter $\psi(\theta) = \{(\theta_1 + R)^2 + \theta_2^2\}^{1/2}$ defines circles centered at $(-R, 0)$. Consider the hypothesis $\psi(\theta) = \psi$ which defines a circle through $(\psi - R, 0)$; then under the hypothesis, the distribution has mean located on this circle. Suppose the data value is $(y_1^0, 0)$, with say $y_1^0 > (\psi - R)$ for easier visualization. The common frequentist p-value would record probability for a standard normal centered on the hypothesized circle and calculated interior to the circle through the data; this is given as

$$p(\psi) = G_\psi(R + y_1^0),$$

where $G_\delta(\chi)$ is the distribution function of the noncentral chi with 2 degrees of freedom and noncentrality parameter δ. From the Bayesian viewpoint the posterior is a standard normal centered at the data $(y_1^0, 0)$; the resulting survivor function at $\psi(\theta) = \psi$ records probability for a standard normal centered at the data and calculated outside the circle $\psi(\theta) = \psi$; this is given as

$$s(\psi) = 1 - G_{R+y_1^0}(\psi) \ .$$

It is easy to see that $p(\psi) \neq s(\psi)$. The geometry is more transparent: a normal density is centered at a distance $|y_1^0 - \psi + R|$ from a circle and the calculation gives probability bounded by the circle. In the frequentist case the point is inside the circle; in the Bayesian case the point is outside: the first probability is less than the second.

As a partial converse for the scalar variable scalar parameter case suppose we have strong matching for all y and θ relative to a flat prior in the parameterization θ. Then

$$\int_{-\infty}^{y} f(y;\theta)dy = \int_{\theta}^{\infty} f(y;\theta)d\theta \ ,$$

which gives $f_{;\theta}(y;\theta) + f_y(y;\theta) = 0$, where the subscripts denote differentiation with respect to the indicated variable; this then in turn determines location model structure $f(y;\theta) = f(y-\theta;0)$.

In conclusion: With location models, the commonly preferred flat prior gives strong probability matching for the broad class of linear parameter, but typically does not give strong matching for nonlinear parameters. The pragmatic solution then for nonlinear parameters is to target the choice of prior on the particular nonlinear parameter of interest.

3. ON CHOOSING THE PRIOR

Consider a scalar or vector continuous parameter θ. Perhaps the oldest choice for a default prior is the uniform prior $\pi(\theta)d\theta = cd\theta$ dating from Bayes (1763) and Laplace (1814), and subsequently referred to as the prior that expresses *insufficient reason* to prefer one θ value over another. This can have particular appeal if the parameterization has some natural physical interpretation. However, in the restricted context of a model and data only it lacks parameterization invariance: a uniform prior for θ differs from a uniform prior for say $\varphi = \theta^3$.

In part to address this nonuniqueness Jeffreys (1946) proposed a constant information prior

$$\pi(\theta)d\theta = c|i(\theta)|^{1/2}d\theta \tag{3.1}$$

where $i(\theta) = E\{-\ell_{\theta\theta}(\theta; y) : \theta\}$ is the information matrix and $\ell_{\theta\theta}(\theta) = (\partial/\partial\theta)(\partial/\partial\theta')\ell(\theta)$ is the Hessian of the likelihood function. This prior is invariant under reparameterization and as such has some special properties. In particular in the scalar parameter case the reparameterization

$$\beta(\theta) = \int^{\theta} i^{1/2}(\theta)d\theta \tag{3.2}$$

yields an information function $i(\varphi)$ that is constant in value. This leads to the prior $d\beta(\theta)$, which is the uniform prior in the parameterization β.

Consider the special case of a general location model say $f\{y - X\beta(\theta)\}$: the Jeffreys' prior $|i(\theta)|^{1/2}d\theta = cd\beta$ extracts the uniform prior calculated in the location parameterizations $\beta(\theta)$; this is the preferred prior of the discussions in Section 2.

Now consider a general model $f(y;\theta)$ but restricted initially to say the scalar parameter case. Then as explained in Cakmak et al (1995, 1998) the model is location to the second order in the information parameterization (3.2) and as noted above the Jeffrey's prior (3.1) gives a flat prior for that parameterization. Thus for the scalar parameter case the Jeffreys' prior has to the second order the preferred properties discussed in Section 2.

Now consider the general model $f(y;\theta)$ but with vector parameter θ. It is widely noted that the Jeffreys' prior (3.1) does not treat component parameters in a satisfactory manner. For example, with the location model $y_i = \mu + \sigma z_i$ using standard normal error, the information determinant $|i(\mu, \sigma^2)| = n/2\sigma^6$ give the Jeffreys' prior $d\mu d\sigma^2/\sigma^3 = d\mu d\sigma/\sigma^2$. This leads to a posterior for σ^2 that in effect derives from a $\sigma^2\chi^2$ distribution for $\Sigma(y_i - \bar{y})^2$, on n degrees of freedom rather that on the natural $n-1$ degrees; and for μ it gives a relocated and rescaled Student distribution again with an inappropriate n degree of freedom. For this example an alternative preferred prior is $d\mu d\sigma/\sigma$. These two priors correspond in this transformation model context to left and right invariant measures on the location scale group. What is not widely noted, however, is that the right prior is invariant under change of origin on the group parameter space and also has some other natural properties (Fraser, 1972).

For the special case of location models we have seen in Section 2 that strong matching is available for the wide class of linear parameters but is typically not available more generally. This phenomenon has stimulated the development of default priors that are specific to component parameters of interest. Thus Peers (1965) and Tibshirani (1993) recommend a prior for $\theta = (\lambda, \psi)$ with interest parameter ψ of the form

$$i_{\psi\psi}^{1/2}(\theta) g(\lambda) d\psi d\lambda$$

where λ is chosen orthogonal to the parameter ψ with $i_{\psi\lambda}(\theta) = 0$. The arbitrariness in the choice of $g(\lambda)$ for the nuisance parameter can cause anomalies but the use of $i_{\psi\psi}^{1/2}$ for the interest parameter is a natural extension from the scalar Jeffreys' case based on (3.2).

As noted above the appropriate handling of component parameters of interest seems to require that the prior be targeted on the parameter of interest. Reference priors initiated by Bernardo (1979) and then generalized (see Bernardo & Smith, 1994) can target a succession of scalar parameter components. The prior is chosen to maximize the possible information that can be contributed by the statistical model; it can be organized in terms of utility and requires various limiting operations. For some cases the reference prior avoids difficulties commonly associated with the Jeffreys prior.

4. BAYESIAN PARADIGM AND RELATED PRINCIPLES

The analysis of a statistical model $f(y;\theta)$ with data y^0 in the context of an assumed prior density $\pi(\theta)$ is a standard conditional calculation yielding the posterior density $\pi(\theta|y^0)$,

$$\pi(\theta|y^0) \propto \pi(\theta) f(y^0;\theta) \ . \tag{4.1}$$

If there are doubts concerning the prior or if it is excluded on legal or other grounds, the analysis would default to other statistical methodologies. Our interest in this paper centers on default procedures that derive from the Bayesian paradigm (4.1).

In (4.1) the model information enters as the y^0 section of the full model, recorded as

$$cf(y^0;\theta) = L(\theta; y^0) = L^0(\theta) \ . \tag{4.2}$$

As such the analysis is said to conform to the Strong Likelihood Principle: that inference should use only model information that is available from the observed likelihood function.

The Jeffreys prior and the reference priors discussed in Section 4 both use sample space averages for each θ examined and thus do not conform to the Strong Likelihood Principle.

At issue in a general sense is whether we should care about model information other than at or near the observed data. A strong argument that we shouldn't care has been given by John Pratt (1962) in his discussion of Birnbaum's (1962) analysis of sufficiency, conditionality, and likelihood. Two instruments are available to measure a scalar θ: the first has a full range while the second has an upper limit on the reading. A measurement is obtained which is within the reporting range of the second instrument; does it matter which instrument made the measurement? To most Bayesians and a few frequentists the answer is that it doesn't matter. Variations on the example argue that only model information in a neighbourhood of the data point is relevant for inference.

Pratt (ibid) views these arguments as "a justification of the (strong) likelihood principle". This may be an overstatement as it seems to require an instrument or sequence of instruments that ultimately register only for the mathematical point, the observed data point. Accordingly we focus on a more moderate principle that focuses on a neighbourhood of the data point.

Consider an instrument that has an interval range, producing the measurement when it is in range and producing effectively the relevant end point when out of range: the distribution function on the range fully records the behaviour of the instrument. Now consider this restrained instrument in comparison with a regular instrument whose distribution function coincides on the particular interval. For a data value that falls within the interval of the restrained instrument, the information available concerning the parameter would seemingly be equivalent to that available from the same data with the unrestrained instrument. We summarize this as a principle:

Local Inference Principle: Inference from a statistical model and data should use only the distribution function at and near the data value together with the data value.

For a vector variable with independent coordinates, this would extend to the vector of distribution functions. For the case with dependent coordinates some further framework is needed that specifies how the coordinates measure the parameters involved; pivotal quantities can provide this extra framework.

The Local Inference Principle provides background for a notion of the sensitivity of a measurement y to a change in the parameter θ: the sensitivity concerning θ at the data point y_0^0 is defined as

$$v^0(\theta) = \frac{dy}{d\theta}\Big|_{y^0} = -\frac{F_{;\theta}(y^0;\theta)}{F_y(y^0;\theta)} , \tag{4.3}$$

where the $dy/d\theta$ is calculated for fixed pivotal F and the subscripts in the third expression denote differentiation. The sensitivity $v^0(\theta)$ can be viewed as the velocity of y at y^0 under θ change calculated for fixed p-value, that is, for fixed probability position $F(y;\theta)$. This velocity $v^0(\theta)$ at y^0 can be viewed as describing how y at y^0 measures the parameter θ.

With a vector $y = (y_1, \ldots, y_n)'$ of independent coordinates the sensitivity becomes a velocity vector $v^0(\theta)$ that records the velocity of y at y^0 under change in the parameter value θ:

$$v^0(\theta) = (v_1(\theta), \ldots, v_n(\theta))' = \frac{dy}{d\theta}\Big|_{y^0} \tag{4.4}$$

where $v_i(\theta)$ is given by (4.3) applied to the coordinate y_i and the third expression is calculated for fixed coordinate by coordinate p-values. Again in this vector context the velocity $v^0(\theta)$ describes how y near y^0 measures the parameter θ; a generalized definition is available in Fraser & Reid(1999). Further discussion of this is presented in Section 7.

The Local Inference Principle allows the use of the observed likelihood function $L^0(\theta)$ of course. It also allows the use of the sensitivity vector $v^0(\theta)$. We will see in Section 6 how the sensitivity vector provides an important calibration in the calculation of measures of departure.

5. CONFIDENCE AND OTHER INVERSIONS

Consider a 95% confidence procedure $C(y)$. A pivotal type quantity $z(y;\theta)$ is given by the indicator function

$$z(y;\theta) = 1_{C(y)}(\theta)$$

and the corresponding survivor function has a lower bound 0.95 at $z = 1$,

$$S(z;\theta) = P(z(y;\theta) \geq 1\ ;\ \theta) \geq .95\ .$$

As our concerns in this paper focus on continuous variables we find it convenient to restrict attention to exact confidence regions at arbitrary confidence levels: specifically we assume that confidence procedures are based on a pivotal quantity $z(y,\theta)$ that has a fixed distribution, is continuously differentiable, and has one-one mappings between any pair of z_i, y_i, θ for each i,

$$z_i(y_i,\theta) \leftrightarrow y_i(z_i,\theta) \leftrightarrow \theta\ . \tag{5.1}$$

The last condition indicates that each coordinate variable y_i can be viewed as measuring θ, as discussed in Section 4. For a vector θ each y_i would be a vector of corresponding dimension.

Now consider a set A on the pivot space with probability content β. Then

$$C(y) = \{\theta : z(y^0;\theta) \text{ in } A\} \tag{5.2}$$

is a β level confidence region. For example with $(y_1,\ldots,y_n)$ from the normal (μ,σ^2) and $z_i = (y_i-\mu)/\sigma$, the set $A = \{z : \sqrt{n}\bar{z}/s_z < t_\alpha\}$ using the right tail α point gives the $1-\alpha$ confidence lower bound $\bar{y}^0 - t_\alpha s_y^0/\sqrt{n}$. In this case the pivotal quantity as just defined above would require the nominal coordinate to be a pair of coordinates with a sign or order condition, but this is a minor technical point of no real consequence here.

A somewhat different way of obtaining an assessment on the parameter space is provided by the fiducial method. This again requires a pivotal quantity and we assume the continuity and other properties as above. Fiducial also requires the same dimension for variable and parameter or a reduction to this by conditioning on an ancillary variable: accordingly we assume that an ancillary $a(y) = a(z)$ is available so that given $a(y) = a(y^0) = a^0$ the pairwise links $y \leftrightarrow z \leftrightarrow \theta$ are one-one.

The fiducial distribution for θ is obtained by mapping the pivotal distribution for given $a(z) = a^0$ onto the parameter space using the one-one mapping $z \leftrightarrow \theta$ for fixed $y = y^0$. A β level fiducial region $D(y^0)$ then has a proportion β of the fiducial distribution. We note in passing that the inverse of $D(y^0)$ using the one-one mapping for fixed y^0 gives a set A on pivot space with probability content β.

We now compare the two inversion methods and assume a model-data instance $\{f(y;\theta), y^0\}$ together with the pivotal structure described above. First we note that the fiducial is more restrictive in that it requires the set A to have conditional probability β given the ancillary in addition to the marginal probability β. We could of course add this conditioning requirement as a positive feature for a confidence procedure.

For β confidence we choose a β region A and then invert, while for fiducial we choose a β region $D(y^0)$ and can note that there is a corresponding β region A. If indeed we choose the regions after the model-data instance $\{f(y;\theta), y^0\}$ is available, then there is a one-one correspondence between confidence and fiducial procedures; indeed, there is no mathematical difference, just a procedural difference: choose and invert or invert and choose. There can of course be differences in assessment particularly in the frame of repeated sampling from the same θ.

As a third method of inverting consider the use of the Bayesian paradigm (4.1). A prior density is a density with respect to some specified support measure. Suppose we have a preferred default prior. We could then combine it with the specified support measure to give a new support measure; the default prior density then becomes the uniform or flat prior with respect to the support measure and other possibilities say $\pi(\theta)$ for the prior become modifications of the default prior. This is part of the background for the Bernardo (1979) reference prior approach (Bernardo & Smith, 1994).

In the location model case examined in Section 2 the location reparameterization gave a natural Euclidean support measure. The flat prior relative to this measure gives a Bayesian inversion that agrees with the confidence inversion and the fiducial inversion. The option then of using a prior $\pi(\theta)$ relative to the chosen parameterization can be viewed as a way of supplementing confidence or fiducial intervals to account for the modifying information $\pi(\theta)$. We pursue this link more generally in Section 10.

6. THE LOCATION PARAMETERIZATION

Consider a statistical model with variable and parameter of the same dimension. With an observed data point y^0 we would of course be primarily interested in the observed log likelihood $\ell^0(\theta) = \ell(\theta; y^0)$. Then, following the discussion in Section 4, we could also quite naturally be interested in the gradient of the log likelihood taken with respect to y at y^0:

$$\varphi^0(\theta) = \nabla \ell(\theta; y)\big|_{y^0} = (\partial/\partial y)\ell(\theta, y)\big|_{y^0} \ . \tag{6.1}$$

As likelihood is typically viewed as $a + \log f(y; \theta)$ with arbitrary a we find it necessary for the uses of (6.1) to work from likelihood that has been standardized to have value 0 at the observed maximum likelihood value, that is, to take likelihood to be

$$\ell(\theta; y) = \log f(y;\theta) - \log f(y;\hat{\theta}) \ . \tag{6.2}$$

In this case $\varphi^0(\hat{\theta}^0) = 0$. If we work more loosely with $\varphi(\theta) = (\partial/\partial y) \log f(y;\theta)$ then the $\varphi(\theta)$ that we use in the various results to follow will be replaced by $\varphi(\theta) - \varphi(\hat{\theta})$ and in effect will be given as

$$\varphi^0(\theta) = \frac{\partial}{\partial y}\ell(\theta;y)\big|_{y^0} - \frac{\partial}{\partial y}\ell(\theta;y)\big|_{(\hat{\theta}^0, y^0)} \ , \tag{6.3}$$

Now consider a more general model $f(y;\theta)$ where the dimension n of the variable is larger than the dimension p for the parameter. The familiar reduction is by means of sufficiency but this is only available for quite special model structure. An examination of general asymptotic models shows that quite generally an ancillary say $a(y)$ of dimension $n-p$ is available thus permitting the conditional analysis as found for example with location models; see Section 2. Of course the ancillary is an approximate ancillary of appropriate order that suffices for third order inference calculations; see Section 7 and Fraser & Reid (1999).

For this general model context the gradient of likelihood would be calculated within the conditional model or equivalently for computation calculated from the full model but taken tangent to the conditioning variable. Let $V = (v_1, \ldots, v_p)$ be p vectors tangent to the observed ancillary surface at the data point. Then we write $\varphi^0(\theta) = \ell_{;V}(\theta; y^0)$ if the likelihood function has been standardized at the maximum likelihood value; this uses

$$\ell_{;V}(\theta;y) = (\partial/\partial V') \ \ell(\theta;y) = \{\ell_{;v_1}(\theta, y), \ldots, \ell_{;v_p}(\theta;y)\} \tag{6.4}$$

where $\ell_{;v}(\theta;y) = (d/dt)\ell(\theta; y+tv)\big|_{t=0}$ defines the directional derivative in the direction v. More generally with $\ell(\theta;y) = \log f(y;\theta)$ we write

$$\varphi^0(\theta) = \ell_{;V}(\theta;y^0) - \ell_{;V}(\hat{\theta}^0; y^0) \tag{6.5}$$

which incorporates the likelihood standardization.

For notation we now use just $\ell(\theta)$ and $\varphi(\theta)$ but emphasize that these depend on the observed data y^0 and also in the general case on the tangent directions V to the ancillary at the data point.

In the context of the Strong Likelihood Principle or in the context of the standard Bayesian paradigm we can view the effective model to be any model so long as the likelihood at y^0 agrees with the observed $\ell(\theta)$. In the present context with the Local Inference Principle we can view the effective model to be any model in the much smaller class that has both likelihood and likelihood gradient equal to the observed $\ell(\theta)$ and $\varphi(\theta)$. On the basis of the discussion concerning the Local Inference Principle, we view this smaller class of models as the more appropriate background for the inference context.

Now consider the possible models that have the given characteristics $\ell(\theta)$ and $\varphi(\theta)$ at the data y^0. An exponential model with given $\ell(\theta)$ and $\varphi(\theta)$ has the form

$$f_E(y;\theta) = \frac{c}{(2\pi)^{p/2}} \exp\{\ell(\theta) + \varphi'(\theta)(y - y^0)\} |\hat{\jmath}_{\varphi\varphi}|^{-1/2} \tag{6.6}$$

where $\hat{\jmath} = -\ell_{\varphi\varphi}(\theta)|_{\hat{\varphi}}$ is the negative Hessian calculated with respect to φ. This arises (Fraser & Reid, 1993; Cakmak et al, 1995, 1998; Cakmak et al, 1994; Andrews, Fraser & Wong, 1999) as a third order exponential model approximation and is referred to as the tangent exponential model at the data y^0. This model is shown (Fraser & Reid, 1993) to provide third order inference at y^0 for any model with given $\ell(\theta)$ and $\varphi(\theta)$.

In a somewhat related manner it is shown (Cakmak et al, 1995, 1998; Cakmak et al, 1994) that the location model with given $\ell(\theta)$ and $\varphi(\theta)$ has the form

$$f_L(y;\theta) = \frac{c}{(2\pi)^{p/2}} \exp\left[\ell\{\theta(\beta - y + y^0)\}\right] |\hat{\jmath}^0_{\beta\beta}|^{-1/2} \tag{6.7}$$

where $\hat{\jmath}^0_{\beta\beta}$ is the observed information on the $\beta = \beta(\theta)$ scale and $\beta(\theta)$ is an essentially unique location reparameterization. This arises as a third order location model approximation and provides third order inference at y^0. The parameter $\beta(\theta)$ has uniqueness

(Fraser & Yi, 1999) subject to expandability in a Taylor's series. We refer to $\beta(\theta)$ as the location reparameterization for the statistical model with given $\ell(\theta)$ and $\varphi(\theta)$ at the data y^0.

For the case of a scalar variable and scalar parameter an explicit expression is available for $\beta(\theta)$:

$$\beta(\theta) = \int_{\hat{\theta}^0}^{\theta} -\frac{\ell_\theta(\theta)}{\varphi(\theta)} d\theta \tag{6.8}$$

where $\ell_\theta(\theta) = (\partial/\partial\theta)\ell(\theta)$ is the score function for the given model (Fraser, 1996).

For the vector parameter case the definition of $\beta(\theta)$ gives the differential equation

$$\ell_{\varphi'}(\theta) = -\varphi'(\theta)\frac{\partial\beta(\theta)}{\partial\varphi'} \tag{6.9}$$

where θ is viewed as a function of φ. This has a unique solution subject to expandability in a power series. A simple expression for $\beta(\theta)$ as in the scalar case does not seem accessible (Fraser & Yi, 1999).

The notion of a variable measuring a parameter in a general sense has been discussed previously and viewed as an important part of statistical modelling. A coordinate by coordinate pivotal quantity can provide a definition for this. Let $z = z(y, \theta)$ be a pivotal quantity as defined in Sections 4 and 5; this can describe the manner in which a variable and parameter are interrelated, or how the variable in a general sense measures the parameter. The directions $V = (v_1, \ldots, v_p)$ that are tangent to an essentially unique second order ancillary are available from the pivotal quantity:

$$V = \frac{\partial y}{\partial \theta}\Big|_{(y^0,\hat{\theta}^0)} = -z_{;y'}^{-1} z_{\theta'}\Big|_{(y^0,\hat{\theta}^0)} , \tag{6.10}$$

where $z_{\theta'} = (\partial/\partial\theta')\ z(y;\theta)$, $z_{;y'} = (\partial/\partial y')\ z(y;\theta)$ and the middle expression involves differentiation for fixed pivotal value. This provides quite generally the $n \times p$ matrix V for the definition (6.5) of $\varphi(\theta)$. In the scalar parameter case the matrix V becomes a vector v that is equal (4.3) to the velocity vector $v^0(\theta)$ at $\theta = \hat{\theta}^0$. For background, see Fraser & Reid (1999).

7. RECENT LIKELIHOOD ASYMPTOTICS

Recent likelihood asymptotics has produced the conditioning procedure described in Section 6 that in effect reduces the dimension of the variable from n for the primary variable to p for the conditional variable which is then the essential measurement variable for the parameter which is also of dimension p (Fraser & Reid, 1999). This then builds on earlier theory that permits the reduction of this variable to a scalar pivotal quantity that gives an essentially unique third order measure of departure from a value say ψ for a scalar interest parameter $\psi(\theta)$. (Barndorff-Nielsen, 1986; Fraser & Reid, 1993, 1995). And this in turn builds on earlier likelihood and saddlepoint approximation theory that gives quite accurate p-values (Daniels, 1954; Lugannani & Rice, 1980).

In almost all cases, frequentist or Bayesian, the resulting approximation is obtained through one or other of the combining formulas

$$\begin{aligned} \Phi_1(r,Q) &= \Phi(r) + \phi(r)\left\{\frac{1}{r} - \frac{1}{Q}\right\} \\ \Phi_2(r,Q) &= \Phi\{r - r^{-1}\log(r/Q)\} \end{aligned} \tag{7.1}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal density and distribution functions and r and Q are very specially chosen measures of departure. These formulas were developed in specific contexts by Lugannani & Rice (1980) and Barndorff-Nielsen (1986, 1991).

For the case of testing a scalar interest parameter ψ, the r quite generally is the signed likelihood ratio

$$r = \mathrm{sgn}(\hat{\psi} - \psi) \cdot \left[2\big\{(\hat{\theta}; y) - \ell(\hat{\theta}_\psi; y)\big\}\right]^{1/2} \tag{7.2}$$

where $\hat{\theta}_\psi$ is the constrained maximum likelihood value given the fixed tested value for ψ. The other ingredient Q is quite problem specific in the recent development of asymptotic likelihood theory and a general definition has been a primary goal.

First consider a scalar full parameter θ with interest parameter $\psi = \theta$. In this case the reparameterization $\varphi(\theta)$ from (6.5) with (6.10) is a scalar parameter and the Q is a corresponding standardized maximum likelihood departure

$$q_f = \mathrm{sgn}(\hat{\theta} - \theta) \cdot |\hat{\varphi} - \varphi| |\hat{\jmath}_{\varphi\varphi}|^{1/2} \ . \tag{7.3}$$

With (7.1) and (7.2) this gives the p-value $p(\theta)$. For the Bayesian approach with θ as the integration variable and $\pi(\theta)$ as the prior, the survivor posterior probability $s(\theta)$ is given by (7.1) and (7.2) with Q taken to be a standardized score departure

$$q_B = \ell_\theta(\theta) |\hat{\jmath}_{\theta\theta}|^{-1/2} \cdot \frac{\pi(\hat{\theta})}{\pi(\theta)} \ . \tag{7.4}$$

These given third order accuracy (Fraser 1990; Fraser & Reid, 1993).

Now consider the general case with interest parameter ψ and an explicit nuisance parameter λ; for this it is convenient to take $\theta' = (\lambda', \psi)$. Let $\varphi(\theta)$ be the reparameterization given by (6.5) with (6.10). The frequentist calculation needs a scalar parameter linear in $\varphi(\theta)$ that stands in for the interest parameter ψ, and the local form of $\psi(\theta)$ at $\hat{\theta}_\psi$ provides the coefficients:

$$\chi(\theta) = \frac{\psi_{\varphi'}(\hat{\theta}_\psi)}{|\psi_{\varphi'}(\hat{\theta}_\psi)|} \varphi(\theta) \ , \tag{7.5}$$

where $\psi_{\varphi'}(\theta) = \partial\psi(\theta)/\partial\varphi' = (\partial\psi(\theta)/\partial\theta') \cdot (\partial\varphi(\theta)/\partial\theta')^{-1} = \psi_{\theta'}(\theta)\varphi_{\theta'}^{-1}(\theta)$. The calculations also require an information determinant for λ at the tested $\psi(\theta) = \psi$ but recalibrated in the φ parameterization:

$$|\jmath_{(\lambda\lambda)}(\hat{\theta}_\psi)| = |j_{\lambda\lambda}(\hat{\theta}_\psi)| \cdot |\varphi_{\lambda'}(\hat{\theta}_\psi)|^{-2} \tag{7.6}$$

where the $r \times (r-1)$ determinant is evaluated as with a design matrix X, $|X| = |X'X|^{1/2}$.

For the frequentist p-value $p(\psi)$ the formulas (7.1) use

$$q_f = \mathrm{sgn}(\hat{\psi} - \psi) \cdot (\hat{\chi} - \hat{\chi}_\psi) \left\{ \frac{|\hat{\jmath}_{\varphi\varphi}|}{|\jmath_{(\lambda\lambda)}(\hat{\theta}_\psi)|} \right\}^{1/2} \tag{7.7}$$

and for the Bayesian survivor probability $s(\psi)$ use

$$q_B = \ell_\psi(\hat{\theta}_\psi) \left\{ \frac{|\hat{\jmath}_{\theta\theta}|}{|j_{\lambda\lambda}(\hat{\theta}_\psi)|} \right\}^{-1/2} \frac{\pi(\hat{\theta})}{\pi(\hat{\theta}_\psi)} \ . \tag{7.8}$$

Examples may be found in Fraser & Reid (1995, 1999) and Fraser, Reid & Wu (1999), Fraser, Wong, Wu (1999). These formulas have third order accuracy (Fraser & Reid, 1995, 1999; Fraser, Reid, & Wu, 1999). A general formula version without explicit nuisance parameterization is available (Fraser, Reid and Wu, 1999).

8. STRONG MATCHING

For a scalar parameter θ and a location model we saw in Section 2 that a flat prior in the location parameterization gives strong matching of frequentist and Bayesian methods. We now use likelihood asymptotics to examine a converse: if strong matching is available then what are the constraints on the model and the prior.

Consider a data point y^0 and suppose that strong matching occurs for all values of θ, that is, $r(\theta) = s(\theta)$. The expressions for $r(\theta)$ and $s(\theta)$ using (7.3) and (7.4) both involve the same signed likelihood ratio r but have different expressions (7.3) and (7.4) for the needed Q. The equality of $r(\theta)$ and $s(\theta)$ thus gives the equality of q_f and q_B, which with the $\varphi(\theta)$ as standardized $\hat{\varphi} = 0$ from (6.3) gives

$$\begin{aligned} \frac{\pi(\theta)}{\pi(\hat{\theta})} &= \frac{\ell_\theta(\theta)}{-\varphi(\theta)} \cdot \frac{\varphi_\theta(\hat{\theta})}{\hat{\jmath}_{\theta\theta}} \\ &= c \left| \frac{d\beta(\theta)}{d\theta} \right| , \end{aligned} \tag{8.1}$$

using (6.8); the first expression on the right is for the case that $\varphi(\theta)$ is an increasing function of θ and has been centered with $\hat{\varphi} = 0$. It follows that strong matching is obtained with a flat prior in the location parameterization (6.8) or equivalently with the prior

$$\pi(\theta) \propto \left| \frac{\ell_\theta(\theta)}{\varphi(\hat{\theta}) - \varphi(\theta)} \right| \tag{8.2}$$

based on the initial θ parameterization. The constant in (8.1) compensates for the possibly different scaling for $\beta(\theta)$ and θ at $\hat{\theta}$; thus $c\beta_\theta(\hat{\theta}) = 1$.

It is of interest that the change of parameter defined by $d\beta(\theta)/d\theta$ is closely related to the velocity $v(\theta)$ of y with respect to θ as recorded in (6.10) based on a pivotal quantity.

We have from (6.8) that

$$\frac{d\beta(\theta)}{d\theta} = -\frac{\ell_\theta(\theta); y^0}{\ell_{;y}(\theta; y^0) - \ell_{;y}(\hat{\theta}^0; y^0)} = \frac{dy}{d\theta}\bigg|_{y^0} . \tag{8.3}$$

In the third expression the differentiation is taken for fixed $\ell(\theta; y) - \ell(\hat{\theta}; y)$, thus treating this standardized likelihood as a pivotal quantity near y^0; for some related views on likelihood as pivotal quantity, see Hinkley (1980). We can thus view (8.3) as a velocity $v(\theta)$ based on an approximate pivotal rather than on the exact pivotal used in (6.10).

Now consider a statistical model $f(y; \theta)$ with vector parameter θ. We saw in Section 2 that a location model with a flat prior in the location parameterization has strong matching for parameters that are linear in the location parameter. We now examine inference for an interest parameter $\psi(\theta)$ that is possibly nonlinear in the present general model context.

For a data point y^0 let $\ell(\theta)$ and $\varphi(\theta)$ be the corresponding likelihood and likelihood gradient. We have noted in Section 6 that there is a corresponding essentially unique location parameterization; let $\beta(\theta)$ be such a parameterization. For statistical and inference properties we note in passing that both $\varphi(\theta)$ and $\beta(\theta)$ are unique up to affine transformations; they can then be standardized to coincide with $\theta - \hat{\theta}^0$ to first derivative at $\hat{\theta}^0$.

Suppose that we have strong matching $p(\psi) = s(\psi)$ for inference concerning ψ. It follows then that q_f and q_B from (7.7) and (7.8) are equal giving

$$\frac{\pi(\hat{\theta}_\psi)}{\pi(\hat{\theta})} = \frac{\ell_\psi(\hat{\theta}_\psi)}{-\hat{\chi}_\psi} \frac{|\varphi_\theta(\hat{\theta})|}{|\hat{\jmath}_{\theta\theta}|} \frac{|\jmath_{\lambda\lambda}(\hat{\theta}_\psi)|}{|\varphi_\lambda(\hat{\theta}_\psi)|} , \tag{8.4}$$

where we assume that $\varphi(\theta)$ is centered so that $\varphi(\hat{\theta}) = \hat{\varphi} = 0$.

To better examine the structure of the prior and to avoid the constant that appears in the scalar case (8.1) we standardize the parameterizations. First we take the definition of $\varphi(\theta)$ with respect to coordinates at y^0 to be such that $\hat{\jmath}_{\varphi\varphi} = I$; next we rescale $\theta - \hat{\theta}^0$ so that φ and $\theta - \hat{\theta}^0$ agree to first derivative at $\theta = \hat{\theta}^0$. We then choose a linear transformation of the linear parameterization $\beta(\theta)$ so that it too coincides with $\theta - \hat{\theta}^0$ at $\theta = \hat{\theta}^0$. In these new parameterizations we then have $\hat{\jmath}_{\theta\theta} = \hat{\jmath}_{\beta\beta} = \hat{\jmath}_{\varphi\varphi} = \varphi_\theta(\hat{\theta}) = \beta_\theta(\hat{\theta}) = I$. This eliminates the middle factor in (8.4).

The derivation of the Bayesian survivor function $s(\psi)$ assumes that the integration coordinates are (λ', ψ) rather than the more general θ used here. To handle the general case here and yet avoid the use of the more general formula in Fraser, Reid & Wu (1999), we recalibrate $\psi(\theta)$ in a one-one manner so that $|\partial\psi/\partial\theta| = 1$ along the curve $\theta = \hat{\theta}_\psi$ generated by varying ψ; for this we note that the recalibration of ψ does not affect the Bayesian survivor function derived from the integration parameter θ, as the essential Bayesian inputs are just the variable of integration and the prior density. With this redefinition we then obtain an interpretation for the first factor in (8.4):

$$\frac{\ell_\psi(\hat{\psi})}{-\hat{\chi}_\psi} = \left| \frac{\partial\psi(\theta)}{\partial\beta'(\theta)} \right|^{-1}_{\hat{\theta}_\psi} . \tag{8.5}$$

We can view this as making a component type adjustment to the prior that in effect attributes a flat prior to change in β along $\theta = \hat{\theta}_\psi$. This aspect then is in accord with the results from the scalar parameter case in (8.1) and (8.2); it also has an observed information correspondence with the Peers-Tibshirani prior mentioned in Section 3.

The results for the scalar parameter case (8.1) and the calculation just given for the first factor in (8.4) suggest that the location parameterization $\beta(\theta)$ is the natural reference parameterization for Bayesian integration. Accordingly we now take the integration variable θ to be $\beta(\theta)$ and then examine the prior when taken with respect to β. In particular the first factor in (8.4) becomes unity and we then obtain

$$\frac{\pi(\hat{\beta}_\psi)}{\pi(\hat{\beta})} = \frac{|\jmath_{[\lambda\lambda]}(\hat{\theta}_\psi)|}{|\varphi_{[\lambda]}(\hat{\theta}_\psi)|} \tag{8.6}$$

where $|j_{[\lambda\lambda]}(\hat{\theta}_\psi)$ is the information determinant recalibrated in the β scale and $|\varphi_{[\lambda]}(\theta_\psi)|$ is the Jacobian determinant with λ rescaled in β coordinates, all at $\hat{\theta}_\psi$.

Now consider a rotation $(\gamma_1, \ldots, \gamma_{p-1}, \alpha)$ of the revised θ coordinates such that $\alpha =$ constant is tangent to $\psi(\theta)$ at $\hat{\theta}_\psi$. If as a special case we have that $\psi(\theta)$ is a linear parameter in terms of the location parameterization β, then we have that strong matching is obtained with a flat prior that has $\pi(\hat{\beta}_\psi)/\pi(\hat{\beta}) = 1$. In this case, α is equivalent to the special $\psi(\theta)$ at $\hat{\theta}_\psi$ and $\gamma = (\gamma_1, \ldots, \gamma_{p-1})$ is the nuisance parameter. Now if $\psi(\theta)$ is nonlinear at $\hat{\theta}_\psi$ we still have $|\varphi_{[\lambda]}(\hat{\theta}_\psi) = |\varphi_{[\gamma]}(\hat{\theta}_\psi)|$ and thus have that

$$\frac{\pi(\hat{\beta}_\psi)}{\pi(\hat{\beta})} = \frac{|\jmath_{[\lambda\lambda]}(\hat{\theta}_\psi)|}{|\jmath_{[\gamma\gamma]}(\hat{\theta}_\psi)|} = \frac{|\jmath_{\lambda\lambda}(\hat{\theta}_\psi)|}{|\jmath_{\gamma\gamma}(\hat{\theta}_\psi)|} \ , \tag{8.7}$$

which is the ratio of the Hessian determinant of likelihood at $\hat{\theta}_\psi$ calculated for the curved nuisance parameter $\lambda(\theta)$ to the Hessian determinant calculated for the linear parameter $\gamma(\theta)$, both treated as nuisance parameters at $\hat{\theta}_\psi$ and both calibrated in the same parameterization. The final expression in (8.7) follows by noting that the ratio is free of the coordinate scaling provided that γ is obtained from the integration coordinates $\theta' = (\lambda', \psi)$.

In conclusion, for the vector parameter case we have strong matching if the interest parameter is linear (in the latent location parameterization) and other wise have strong matching if the general flat prior is adjusted by the nuisance information ratio (8.7).

For an example consider the normal circle problem at the end of Section 2. For the full parameter this is a location model and we have $(\theta_1, \theta_2) = (\varphi_1, \varphi_2) = (\beta_1, \beta_2)$ with observed information determinants equal to one at all points. For a curved component parameter we examined the distance $\psi = \{(\theta_1 + R)^2 + \theta_2^2\}$ of (θ_1, θ_2) from $(-R, 0)$; let $r = \{(y_1 + R)^2 + y_2^2\}$ be the analogous distance of (y_1, y_2) from $(-R, 0)$. Certainly r is a natural variable measuring ψ. Also let α and a be the related polar angles for (θ_1, θ_2) and (y_1, y_2) relative to the positive axis from the point $(-R, 0)$: We can view α as the nuisance parameter and note the $a - \alpha$ has the von Moses distribution with shape parameter ψr conditional on r.

The first factor in (8.4) has the value 1 for this example, as ψ directly records Euclidean distance. The second also has the value 1 as the informations are already standardized. The third factor recorded in (8.7) takes the value

$$\frac{j_{\lambda\lambda}(\hat{\theta}_\psi)}{j_{\gamma\gamma}(\hat{\theta}_\psi)} = \frac{r}{\psi} = \frac{\pi(\hat{\theta}_\psi)}{\pi(\hat{\theta})} \ .$$

Thus the prior ψ^{-1} adjusts the general flat prior $\psi d\alpha d\psi = d\theta_1 d\theta_2$ to give strong matching for $\psi(\theta)$ to the third order.

We can give a geometrical overview of this by examining (θ_1, θ_2) and $(y_1, y_2) = (\hat{\theta}_1, \hat{\theta}_2)$ on the same 2-dimensional plane. For given ψ we have (θ_1, θ_2) on the circle $\psi(\theta) = \psi$; for given r we have (y_1, y_2) on the circle $r(y_1, y_2) = r$. The vector from (θ_1, θ_2) to (y_1, y_2) is standard normal from the frequency viewpoint and also from the Bayesian flat prior viewpoint. ¿From the frequentist viewpoint this vector is integrated on a region having endpoint (y_1, y_2) on the circle $r = r^0$; from the Bayesian viewpoint this vector is integrated on a region having origin point (θ_1, θ_2) on the circle $\psi(\theta) = \psi$; recall the comments at the end of Section 2 on the probability on the inside or outside of a circle at a distance from the datapoint. This shows clearly the need for the Bayesian adjustment for a curved parameter component; and as indicated above the adjusted prior is uniform $d\alpha d\psi$ in the polar coordinates.

9. LOCATION PRIORS FROM PIVOTAL QUANTITIES

In the preceeding section we showed that strong matching to third order was obtained by the use of a flat prior with respect to a location parameterization $\beta(\theta)$. An explicit expression (8.3) for $\beta(\theta)$ was obtained in the scalar parameter case and an existence result for $\beta(\theta)$ was presented for the vector parameter case. These results were based on an observed likelihood $\ell(\theta) = \ell(\theta; y^0)$ and an observed likelihood gradient $\varphi(\theta) = \ell_{;V}(\theta; y^0)$. Calculation of the gradient $\varphi(\theta)$ required an approximate ancillary with vectors $V = (v_1, \ldots, v_p)$ tangent to the ancillary at the data point; fortunately for applications the tangents V can be derived from a pivotal quantity (5.1) without explicit construction of the approximate ancillary. In this section we show that the flat prior $d\beta(\theta)$ itself can be developed to second order directly from the pivotal quantity.

The approximate ancillary for these calculations was derived in Fraser & Reid (1995, 1999). This used a location model (Fraser, 1964) that coincides with the given model at

$\theta_0 = \hat{\theta}^0$ to first derivative. The orbits on the full sample space for this location model were then bent to give second derivative, second order ancillarity as calculated in terms of the given model; this order of ancillarity then provides third order inference. While the constructed ancillary would seemingly depend on the data point, it can be shown to be free of that choice to the requisite order for third order inference.

The bending of the orbits was to eliminate a marginal effect of second order magnitude and thus to produce ancillarity to the second order. Our interest here however centers on the conditional distribution on the orbits and how it is affected by the bending. For this we follow Fraser & Reid (ibid) and restrict attention initially to the scalar parameter case.

First consider the location model orbits (Fraser, 1964) derived from properties of the given model to first derivative at θ_0. The velocity vector $v(\theta_0)$ from (4.4) gives the direction of the orbit at the data point and also the magnitude of y-change corresponding to θ-change at θ_0. Does the bending of the orbits affect this?

Consider the conditional distribution along the location orbit through y^0 but using the given model rather than the tangent location model. The distribution will typically not be location; however a reexpressions of the variable and the parameter can make it location to second order (Cakmak et al, 1998), with standardized form say

$$(2\pi)^{-1/2} \exp\left\{ -\frac{(y-\theta)^2}{2} + \frac{a}{\sqrt{n}} \frac{(y-\theta)^3}{6} + \frac{k}{\sqrt{n}} \right\} . \tag{9.1}$$

In terms of the original parameter and variable the non location characteristics will to second order depend on some variable say x which by general theory (Fraser & Reid, 1995) can be examined in terms of a one dimensional conditional distribution, with standardized form say

$$\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

to first order. Now consider bending in the context of the two dimensional conditional distribution for (x, y). For the conditional distribution of y suppose we bend the orbit to the right say and condition on $X = x - cy^2/2n^{1/2}$ with $c > 0$; the new conditional distribution for y has location $\theta(1-cX/n^{1/2})$ and scale $(1-cX/2n^{1/2})$. At the point $y=0$ we then have that $dy/d\theta = 1+cX/n^{1/2}$, written as say $\exp\{k_1/n^{1/2}\}$, which is a constant free of θ. Thus the bending changes the velocity $v^0(\theta_0)$ to $\exp\{k_1/n^{1/2}\}v^0(\theta_0)$.

Now consider the velocity vector $v(\theta)$. At the data y^0 this is tangent to an orbit generated by the location model derived from first derivative change in the given model at the value θ. Such orbits are typically at an $O(n^{-1/2})$ angle to the bent orbit just described except of course at the point having maximum likelihood value θ, where the orbit is tangent to the bent orbit. The conditional distribution on the bent orbit as opposed to this θ orbit distribution will then have a factor $\exp\{k_1/n^{1/2}\}$ coming from the curvature in the manner described above for the value θ_0. In that bending result the standardized variable y recorded distance from the maximum likelihood surface $(y=0)$. Now the reference maximum likelihood surface corresponds to the value θ and a contour with fixed y is parallel to this surface. To transfer the velocity vector $v(\theta)$ to the bent orbit with tangent space $\mathcal{L}\{v(\theta_0)\}$ we should thus project parallel to this θ surface. The observed maximum likelihood surface differs from this by an $O(n^{-1/2})$ angle; and the projection of $v(\theta)$ to $\mathcal{L}\{v(\theta_0)\}$ is through an $O(n^{-1/2})$ angle. Thus it suffices to project parallel to the observed maximum likelihood surface and still retain $O(n^{-1})$ accuracy.

Now let $Hv(\theta)$ be this projection. We then have that the velocity vector on the curved orbit is $\exp\{k_1/n^{1/2}\}Hv(\theta)$. It follows that the location prior satisfies

$$d\beta(\theta) = \exp\{k_1/n^{1/2}\}|Hv(\theta)|d\theta \tag{9.2}$$

when calibrated by unit change at y^0 or satisfies

$$d\beta(\theta) = \frac{|Hv(\theta)|}{|v(\theta_0)|}d\theta \tag{9.3}$$

when calibrated by unit change in θ at θ_0.

To simplify these expressions we now examine the process of projecting parallel to the observed maximum likelihood surface. The observed maximum likelihood surface satisfies

$$\ell_\theta(\hat{\theta}^0; y) = 0$$

and the gradient vector nominally perpendicular to the surface is given by

$$k(y;\hat{\theta}^0) = \ell_{\theta';y}(\hat{\theta}^0;y) \tag{9.4}$$

which is the vector $w = k(y^0;\hat{\theta}^0)$ at the data point y^0. The length of the vectors in (9.3) can then be compared by projecting them to $\mathcal{L}(w)$, that is, by projection parallel to the maximum likelihood surface. Accordingly we can rewrite (9.3) as

$$d\beta(\theta) = \frac{w'v(\theta)}{w'v(\theta_0)} d\theta \ . \tag{9.5}$$

This expression for the prior was calculated from a distribution function viewpoint whereas (8.3) was derived from a likelihood viewpoint. A small detail remains to reconcile the different approaches. Consider the asymptotic distribution given the approximate ancillary. Using the pivotal quantity $F(y;\theta)$ we obtain

$$d\beta(\theta) = -\frac{F_{;\theta}(y^0;\theta)}{F_y(y^0;\theta)}\ d\theta \ ,$$

as in (4.3); while from the likelihood analysis we obtain

$$d\beta(\theta) = -\frac{\ell_\theta(\theta;y^0)}{\ell_{;y}(\theta;y^0) - \ell_{;y}(\theta_0;y^0)}\ d\theta \ .$$

The integration results in Andrews et al (1999) show that these differ to third order by a constant factor $\exp\{k_2/n\}$ and thus provide the same location reparameterization to that order.

We now record a preliminary examination of the vector parameter case. A first derivative change at a parameter value θ generates (4.4, 6.10) a vector $v(\theta)$ for each direction of change from the value θ; these can be assembled as an $n \times p$ array

$$V(\theta) = \big(v_1(\theta), \ldots, v_p(\theta)\big)$$

of p vectors corresponding to the p coordinates for θ. The vectors $V = V(\hat{\theta}^0)$ provide the tangent vectors (6.10) to the second order ancillary.

First suppose that θ is the location reparameterization whose existence is established in Fraser & Yi (1999). It follows that first derivative change at a value θ generates on the ancillary surface the location orbits for the tangent location model. Within this location model we seek the Jacobian determinant recording the ratio of volume change at y^0 to volume change at θ.

The bending of the conditional distribution in the vector parameter case was examined in Fraser & Reid (1995, 1999). Then following the pattern earlier in this section for the scalar case, we find that the standardized coordinates are rescaled by factors $\exp\{k_3/n^{1/2}\}$ free of θ and that projection can be taken parallel to the observed maximum likelihood surface with retention of second order accuracy.

The gradient vectors nominally perpendicular to the maximum likelihood surface are given by (9.4) which at the data point y^0 form the $n \times p$ array

$$W = \ell_{\theta';y}(\hat{\theta}^0, y^0) \ . \tag{9.6}$$

We can then compare $V(\theta)$ to $V(\theta_0)$, projected parallel to the observed maximum likelihood surface, by taking the inner product array with W giving the location prior

$$d\beta(\theta) = \frac{|W'V(\theta)|}{|W'V(\theta_0)|} d\theta \tag{9.7}$$

when calibrated by unit change at the observed $\theta_0 = \hat{\theta}^0$.

10. LOCATION PIVOTAL QUANTITIES

In Section 5 we discussed confidence and other inversion procedures and found an equivalence among them in the context of a location statistical model: a flat prior relative to the location parameterization produces a Bayesian inversion that coincides with confidence and other inversions. In addition this then allows that a personal or communal prior expressed relataive to the location parameterization can be used to adjust confidence or fiducial priors to include the personal or communal input.

We note also that a location parameterization for the presentation of likelihood has been strongly promoted by Professor David Sprott. For some discussion and related attractive properties see Fraser & Reid (1998). The present use of the location parameterization in the inversion context has close ties to reference priors; see for example, Bernardo and Smith (1994).

Consider first the case of a scalar parameter θ and suppose the corresponding variable is scalar as isolated say by sufficiency or conditionality. If the parameter θ itself is location then $z = \hat{\theta} - \theta$ is pivotal. More generally, the location parameterization (6.8) gives the approximate pivotal quantity

$$z = \hat{j}_{\beta\beta}^{1/2}(\hat{\beta} - \beta) = \hat{j}_{\beta\beta}^{1/2} \int_{\hat{\theta}}^{\theta} \frac{-\ell_\theta(\theta}{\varphi(\theta)} d\theta \tag{6.1}$$

To the first order this is standard normal; to the second order it has a fixed distribution which is available from likelihood (6.7) in terms of β; and to the third order it has pivotal properties available from Section 7. These pivotal quantities can be inverted to give confidence, fiducial or flat prior intervals, and these in turn can be adjusted by prior information expressed relative to the location parameterization.

Now briefly consider the vector parameter case. The location prameterization $\beta(\theta)$ exists as noted in Section 6. Let $\hat{j}_{\beta\beta}^{1/2}$ be a square root of the observed information expressed in terms of β; then

$$z = \hat{j}_{\beta\beta}^{1/2}(\hat{\beta} - \beta)$$

is an approximate vector pivotal quantity with fixed distribution properties as indicated by the scalar case above. This can be inverted to give confidence, fiducial or flat prior Bayesian regions following the methods in Sections 2 and 5 using approximation theory from Section 6.

REFERENCES

Andrews, D.F., Fraser, D.A.S. and Wong, A (1999). Some useful integrals for asymptotic densities and the mystery of hyper-accuracy, submitted *J. Statist. Planning and Inference.*

Barndorff-Nielsen, O.E. (1986). Inference on full or partial parameters based on the standardized, signed log likelihood ratio. *Biometrika* **73** 307-22.

Barndorff-Nielsen, O.E. (1991).

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances.

Birnbaum, A. (1962). On the foundations of statistical inference. *Amer. Statist. Assoc.* **57** 269-306.

Bernardo, J.M. (1979). Reference posterior distributions for Bayesian inference. *J. Royal. Statist. Soc. B* **41**, 113-147.

Bernardo, J.M. and Smith, A.F.M. (1994). Bayesian Theory. Chichester: J. Wiley & Sons.

Cakmak, S., Fraser, D.A.S., McDunnough, P., Reid, N. and Yuan, X. (1995). Likelihood centered asymptotic model : exponential and location model versions. In the *Journal of Mathematical and Statistical Science* **4** 1-12.

Cakmak, S., Fraser, D.A.S., McDunnough, P., Reid, N. and Yuan, X. (1998). Likelihood centered asymptotic model: exponential and location model versions. *J. Statist. Planning and Inference* **66**, 211-222; this is a republication in error of Cakmak et al. (1995).

Cakmak, S., Fraser, D.A.S., and Reid, N. (1994). Multivariate asymptotic model: expoonential and location model approximation, *Utilitas Mathematica* **76**, 604-608.

Daniels, H.E. (1954). Saddlepoint approximations in statistics. *Annals Math. Statist.* **46**, 21-31.

Fraser, D.A.S. (1964). Local conditional sufficiency. *Jour. Royal. Statist. Soc. B* **26**, 52-62.

Fraser, D.A.S. (1972) Bayes, likelihood or structural. *Annals Math. Statist.* **43**, 777-790.

Fraser, D.A.S. (1990), Tail probability from observed likelihood, *Biometrika* **77**, 65–76.

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. (1995). Ancillaries and third order significance. *Utilitas Mathematica*, 47, 33-53.

Fraser, D.A.S. and Reid, N. (1998). On the informative presentation of likelihood. *Applied Statistical Science*, Commack, New York: Nova Science Publishers.

Fraser, D.A.S. and Reid, N. (1999). Ancillary information for statistical inference, Proceeding of the confidence on Emperical Bayes and likelihood. Springer, to appear.

Fraser, D.A.S., Reid, N., and Wu, J. (1999). A simple general formula for tail probabilities for frequentist and Bayesian inference. *Biometrika*, to appear.

Fraser, D.A.S., Wong, A., and Wu, J. (1999). Regression analysis, Nonlinear on nonnormal: simple and accurate p-values from likelihood analysis, to appear *J. Amer. Stat. Assoc.*

Fraser, D.A.S. and Yi, Y. (1999). Uniqueness of the location parameterization in likelihood asymptotics, submitted to *Annals Statistics.*

Hinkley, D.V. (1980). Likelihood as approximate pivotal distribution. *Biometrika* **67**, 287-292.

Laplace, P.S. (1814). Essai Philsophique sur les Probabilités, Paris: Courcier, *Phil. Trans. Roy. Soc. London* **53**, 370-418.

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.

Peers, H.W. (1965). On confidence points and Bayesian probability points in the case of several parameters. *J. Royal Statist. Soc. B* **27**, 9-16.

Pratt, J.W. (1962). Discussion of Birnbaum (1962). *Jour. Amer. Statist. Assoc.* **57**, 314-315.

Tibshirani, R. (1993). Noninformative priors for one parameter of many. *Biometrika* **76**, 604-608.