Pak. J. Statist.
200x, Vol. xx(x), xx-xx

SOME CORRECTIONS FOR BAYES CURVATURE

Fraser, D.A.S.[1] and **Sun, Y.**[2]

[1] Department of Statistics, University of Toronto, Toronto, Canada
Email: dfraser@utstat.utoronto.ca
[2] Department of Statistics, University of Toronto, Toronto, Canada
Email: ysun@utstat.utoronto.ca

ABSTRACT

Bayesian and frequentist methodologies when applied to the same model–data information can lead to different statistical inference results. A prominent example involves a rotationally symmetric normal error distribution located at an arbitrary point (θ_1, θ_2) on the plane. The radial distance $\rho = (\theta_1^2 + \theta_2^2)^{1/2}$ from the origin has a Bayes posterior survival value $s(\rho)$ that is uniformly greater than the frequentist p-value $p(\rho)$, can be expressed in terms of the noncentral chi-square distribution function with 2 degrees of freedom, and can attain 8 percentage points when $\hat{\rho} = 5$. We use this Bayes–frequentist difference as a reference to explore the Bayesian bias attributable to parameter curvature.

For this, we consider a two parameter regular statistical model and define a curvature measure for an interest parameter; the curvature measure is a mofication of the Efron measure and targets Bayesian adjustment rather than departure from the information lower bound as considered by Efron. Examples are given and simulations are provided.

KEYWORDS

Bayes; Asymptotics; Inference; Curvature corrections; Conditioning; Default priors.

2000 Mathematics Subject Classification: 62F15; 62F12.

1 Introduction

Bayesian and frequentist methodologies applied to the same model-data information can lead to quite different inference statements. This has been discussed intermittently in the

literature: for example David, Stone & Zidek (1973) examined Bayesian and confidence distributions and found that marginalization to a component parameter can quite widely give results in conflict with a direct analysis of the component parameter; Stainforth, Allen, Tredger & Smith (2007) discussed two comprehensive weather models that give quite different results with comparable analyses; Fraser (2009) examined a model asymptotically close to normality and found that no Bayesian analysis could reproduce the usual confidence results to third order accuracy.

The prominent example for such discrepancies involves two independent normal variables y_1, y_2 with means θ_1, θ_2 and say common variance equal to 1. A parameter $\rho = (\theta_1^2+\theta_2^2)^{1/2}$ whose contours have apparent curvature has a direct p-value $p(\rho) = H_2(r^2;\, \rho^2)$ using $r^2 = y_1^2+y_2^2$ where $H_2(r^2;\, \rho^2)$ is the distribution function of the noncentral chi-square distribution with 2 degrees of freedom and noncentrality ρ^2. This is the elementary p-value calculation that records the percentage position of the data value with respect to the distribution with parameter value ρ^2; in repetitions, the corresponding confidence lower bound $\hat{\theta}_\beta(y_1, y_2)$ has repetition accuracy β.

The variable (y_1, y_2) has of course the obvious location invariance assumed by Bayes (1763) and leads to the posterior that (θ_1, θ_2) is Normal $\{(y_1, y_2); I\}$; the marginal distribution for ρ^2 is the noncentral chi-square with degrees of freedom 2 and noncentrality $r^2 = y_1^2 + y_2^2$, and the resulting Bayes posterior survival value is $s(\rho) = 1 - H_2(\rho^2; r^2)$; this is strictly larger than $p(\rho)$ and the difference $s(\rho) - p(\rho)$ can attain the value 8% with $\hat{\rho} = r = 5$; for some details see Fraser & Reid (2002) and Fraser (2009); and a plot of the discrepancy

$$D(\rho, \hat{\rho}) = s(\rho) - p(\rho) = 1 - H_2(\rho^2, \hat{\rho}^2) - H_2(\hat{\rho}^2, \rho^2)$$

with $\hat{\rho} = 5$ is recorded in Figure 1. We view the parameter curvature as illustrated in this simple example to be the central source of the discrepancies coming from the use of the Bayesian approach with default priors; for further discussion see Fraser (2009).

In Section 2, we give a brief overview of Efron's definition of parameter curvature, which is defined primarily for exponential models. We then discuss how recent likelihood asympotics shows that to the second order exponential and location models can be viewed as differing in just their parameterization. And we argue that for inference curvature should be defined using the location model framework.

In Section 3, we discuss the parameter linearity examined in Fraser, Reid, Marras & Yi (2009) and define a measure of curvature for a scalar interest parameter ψ in the presence of a scalar nuisance parameter λ; the full parameter can then be presented as $\theta = (\psi, \lambda)$. The measure of curvature γ depends on the data say y and is recorded as $\gamma = \hat{\gamma} = \gamma(\hat{\psi})$, and its

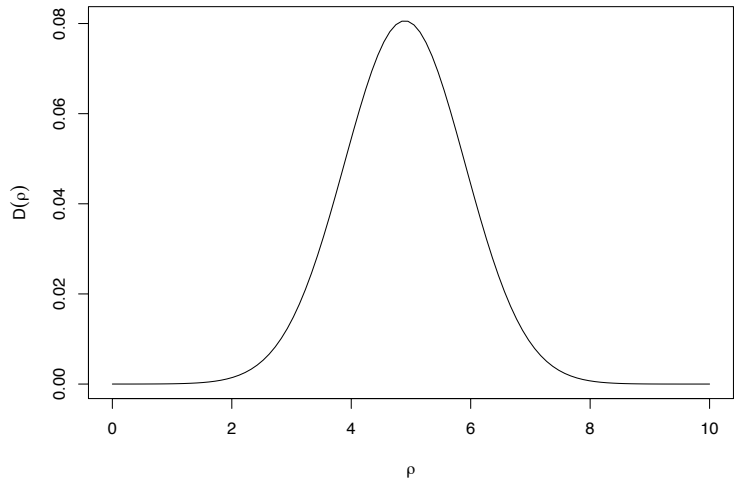

Figure 1: Bayesian discrepancy $D(\rho, \hat{\rho} = 5)$

reciprocal is the corresponding radius of curvature $\rho = 1/\gamma$. For the example we have $\hat{\gamma} = 0.2$, with reciprocal the radius of curvature $\hat{\rho} = 5$. This measure of curvature is calculated in relation to how the parameter affects the sample space near the observed data, and it agrees with the curvature measure of Efron (1975) in the case of a normal location model. More generally it relates to a linearity or additivity on the sample space rather than linearity or additivity for the log-density as in Efron (1975). The present definition addresses the role of Bayesian analysis while the Efron definition focuses on the availability of uniformly most powerful test. Section 5 and 6 discuss a familiar example in substantial detail. For an exponential model a corresponding canonical parameter is linear in the Efron sense, while for a location model a corresponding parameter is linear in the present modified sense.

2 Defining parameter curvature

For the simple location normal on the plane discussed in Section 1 we obtained a curvature $\gamma = 0.2$ for the parameter ρ. This curvature measure however does not in general agree with a prominent definition of curvature in the literature. Efron (1975) defined parameter curvature with primary reference to an exponential model. In this paper, we propose a

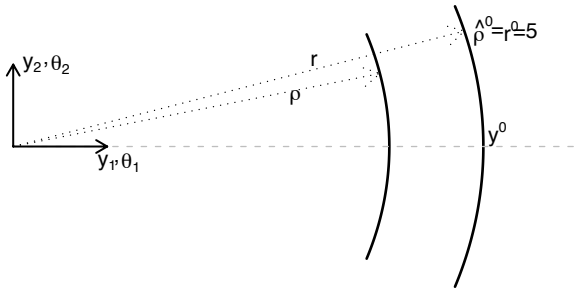

Figure 2: Normal$\{(y_1,y_2);(\theta_1,\theta_2),I)\}$ on the plane. Data point $y^0=(y_1^0,y_2^0)=(r^0,0)$ for convenience on the first axis. Parameter contour $\rho=\hat{\rho}=r^0$ through the data point. Increasing values of ρ are to the right with center of curvature to the left and curvature $\gamma=1/\rho$ where ρ is the radius of curvature. As pictured, the curvature is $0.2=1/\hat{\rho}^0=1/5$. Positive curvature reads as inflated posterior probability. If we evaluate at the mle $\hat{\rho}^0=5$, we obtain $s(5)=0.54009839$ and $p(5)=0.45990161$ with discrepancy 0.08019677.

modified measure of curvature having primary reference to a location model.

First consider a curve $y=y(x)$ on the (x,y) plane. If $y_x(x)=(d/dx)y(x)=0$, the curvature $\gamma(x)=y_{xx}=(d^2/dx^2)y(x)$ is the second derivative at the point x and its reciprocal $\rho(x)=1/\gamma(x)$ is the radius of curvature of a circle that has first and second degree agreement with the given curve, that is, fits the given curve to second order at x where $y_x(x)=0$. Standard geometry then shows that the curvature at a general point x is

$$\gamma(x)=\frac{y_{xx}}{(1+y_x^2)^{3/2}}$$

with a sign for $\gamma(x)$ possibly based on curvature to the left or to the right relative to a chosen direction on the curve. Alternatively, if the curve of interest is given implicitly as $g(x,y)=c$, then the curvature can be written as

$$\gamma(x,y)=\frac{g_{xx}g_y^2-2g_{xy}g_xg_y+g_{yy}g_x^2}{(g_x^2+g_y^2)^{3/2}}.$$

For some discussion see Efron (1975).

If the scaling for x and y is changed then the curvature will change in a corresponding way. Thus when examining the curvature of a parameter the local scaling will typically be derived from the model based on an expected or observed information matrix. The expected information for θ is

$$i_{\theta\theta}(\theta) = E\left\{ \begin{pmatrix} -\ell_{\theta_1\theta_1}(\theta;y) & -\ell_{\theta_1\theta_2}(\theta;y) \\ -\ell_{\theta_1\theta_2}(\theta;y) & -\ell_{\theta_2\theta_2}(\theta;y) \end{pmatrix} ;\, \theta \right\}$$

and the parametric distance ds from (θ_1,θ_2) to $(\theta_1+d\theta_1,\theta_2+d\theta_2)$ is obtained from

$$(ds)^2 = \begin{pmatrix} d\theta_1 \\ d\theta_2 \end{pmatrix}' i_{\theta\theta}(\theta) \begin{pmatrix} d\theta_1 \\ d\theta_2 \end{pmatrix} \tag{2.1}$$

If observed information is used then recent likelihood theory suggests the use of the data-defined canonical parameter $\varphi(\theta)$ for a reference exponential model:

$$\varphi(\theta) = \frac{d}{dV}\ell(\theta,y)|_{y^0} \ ,$$

where $V=(v_1,v_2)$ records two vectors tangent to a second order ancillary, usually available as $V = dy/d\theta|_{(y^0;\hat\theta^0)}$ where $y = y(x,\theta)$ is the full vector quantile function derived from independent coordinate distribution functions and x has a null or reference distribution corresponding to y at say some θ_0. See Fraser, Reid & Wu (1999).

Efron (1975) works from an exponential model

$$g(s;\varphi) = \exp\{s'\varphi(\theta) - \kappa(\varphi)\}g(s)$$

where reduction from original data to the canonical variable (s_1,s_2) involves a natural use of sufficiency. For such a model, an alternative parametrization is given by

$$\begin{pmatrix} \tau_1 \\ \tau_2 \end{pmatrix} = E\left\{ \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} ;\varphi \right\} = \frac{\partial}{\partial\varphi}\kappa(\varphi) = \kappa_\varphi(\varphi)$$

with the final expressions available from the familiar Mean-score-equal-zero relationship. The corresponding variance matrix is

$$var\left\{ \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} ;\varphi \right\} = \kappa_{\varphi\varphi}(\varphi) = i_{\varphi\varphi}(\theta)$$

which is the Hessian matrix obtained by twice differentiating the underlying cumulant-generating function, which is the function $\kappa(\varphi)$. This is the information matrix for the parameter $\varphi(\theta)$.

Now consider a scalar parameter sub-model obtained by having $\varphi(\alpha)$ expressed as a function of the scalar parameter α. It follows then that the mean parameter is also a function of α

$$\begin{pmatrix} \tau_1 \\ \tau_2 \end{pmatrix} = \kappa_\varphi(\varphi(\alpha))$$

The Efron (1975) curvature of this scalar parameter model is defined to be the curvature of the expectation curve $\{\tau(\alpha)\}$ calculated with respect to information scaling; it follows that

$$\gamma(\alpha) = \frac{|M(\alpha)|^{1/2}}{\nu_{11}^{3/2}(\alpha)} = \frac{\begin{vmatrix} \nu_{11}(\alpha) & \nu_{12}(\alpha) \\ \nu_{12}(\alpha) & \nu_{22}(\alpha) \end{vmatrix}^{1/2}}{\nu_{11}^{3/2}(\alpha)}$$

where

$$\begin{aligned} \nu_{22}(\alpha) &= \varphi'_{\alpha\alpha}(\alpha) i_{\varphi\varphi}(\alpha) \varphi_{\alpha\alpha}(\alpha) \\ \nu_{11}(\alpha) &= \varphi'_{\alpha}(\alpha) i_{\varphi\varphi}(\alpha) \varphi_{\alpha}(\alpha) \\ \nu_{12}(\alpha) &= \varphi'_{\alpha}(\alpha) i_{\varphi\varphi}(\alpha) \varphi_{\alpha\alpha}(\alpha); \end{aligned}$$

the determinant can be viewed as the variance of an acceleration vector orthogonalized to a tangent velocity.

Example 2.1. Bivariate normal (Efron). Let y be bivariate normal with mean (φ_1, φ_2) and variance I and consider the sub model with mean vector

$$\varphi(\alpha) = \begin{pmatrix} \varphi_1(\alpha) \\ \varphi_2(\alpha) \end{pmatrix} = \begin{pmatrix} \alpha \\ (\gamma_0/2)\alpha^2 \end{pmatrix}$$

dependent on the scalar parameter α. Then

$$\varphi_\alpha = \begin{pmatrix} 1 \\ \gamma_0 \alpha \end{pmatrix}, \quad \varphi_{\alpha\alpha} = \begin{pmatrix} 0 \\ \gamma_0 \end{pmatrix}$$

$$M(\alpha) = \begin{pmatrix} 1+\gamma_0^2\alpha^2 & \gamma_0^2\alpha \\ \gamma_0^2\alpha & \gamma_0^2 \end{pmatrix}$$

$$\gamma(\alpha) = \frac{\gamma_0}{(1+\gamma_0^2\alpha^2)^{3/2}}$$

giving $\gamma(0) = \gamma_0$ which is just the ordinary curvature of the mean vector at the origin, a consequence of location normality with identity variance.

Recent likelihood asymptotics examines the form of the log density of a statistical model and uses the data accretion rate $O(n)$ that is applicable to the log density itself and also to the related coefficients of its Taylor expansion about an observed data point; see for example, Andrews, Fraser &Wong (2005) and Cakmak, Fraser, McDunnough, Reid & Yuan (1998). From this, it is found that the model to second order can be written as an exponential model or as a location model and that an $O(n^{-1})$ adjustment can convert from one to the other; the adjustment can be shown to correspond to reexpression of the variable and reexpression of the parameter.

We use recent results from default priors theory concerning location parameterizations to develop a curvature measure that references location model properties. We do not obtain this measure in an explicit form but rather as a computational algorithm. An explicit formula would be appealing but the algorithm has some advantages in explaining more clearly the basis of the definition and its advantages. We address this in the next section.

3 Parameter effect: Linearity and curvature

We have noted in the preceding section that a model can be written in exponential form or written in location form to second order by just reexpression of the variable, and reexpression of the parameter; and to measure curvature, we have from the literature, the well defined Efron curvature measure which is appropriate with the exponential type form. We now address the need for a curvature measure that is based on the location type context.

First consider the location model $f(y_1-\psi, y_2-\lambda)$ on the (y_1, y_2) plane with data (y_1^0, y_2^0). If we view the parameter ψ as being of primary interest, we might examine the marginal density for y_1 and then find that it depended only on ψ and was thus free of λ

$$f(y_1;\psi) = f_1(y_1-\psi)$$

where $f_1(t_1) = \int f(t_1,t_2)dt_2$. We might then speak of ψ as being linear as opposed to curved. How can this be generalized?

For this we draw on results from default priors in Fraser, Reid, Marras & Yi (2009) and ancillary statistics in Fraser, Fraser & Staicu(2009). We assume a model with asymptotic

properties and with continuous differentiability with respect to the variable and the parameter. Then to go beyond the pattern indicated by large sample likelihood theory we would seem to need a parameter–variable connection of the type provided by a pivotal quantity. We generalize this and with independent coordinate distribution functions use the quantile functions to express the parameter–variable relationships. We can then write (Fraser, Reid, Marras & Yi, 2009)

$$d\hat{\theta} = W(\theta)d\theta, \;\; d\theta = M(\theta)d\hat{\theta} \tag{3.1}$$

where $W(\theta)$ and $M(\theta) = W^{-1}(\theta)$ are $p \times p$ matrices that show how change $d\hat{\theta}$ at the observed data y^0 relates to change $d\theta$ at an arbitrary θ value, as calculated to second order; for this we have

$$W(\theta) = \hat{j}^{-1}H'V(\theta)$$

where $\hat{j}$ is the observed information matrix,

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

is the gradient of the score function at the observed data, and

$$V(\theta) = \frac{dy}{d\theta} = \frac{d}{d\theta}y(x;\theta)\bigg|_{y^0} \tag{3.2}$$

is the derivative of the quantile function at the observed data; the reference variable x could be the corresponding estimated p-value vector or an equivalent variable.

Example 3.1. Normal linear regression. Consider a sample from the standard linear regression model $y = X\beta + \sigma z$ where z is a sample from the standard normal and X has full column rank and $\theta = (\beta', \sigma^2)$. Then

$$V(\theta) = \frac{dy}{d\theta}\bigg|_{y^0} = \{X, z^0(\theta)/2\sigma\}$$

where $z^0(\theta) = (y^0 - X\beta)/\sigma$ is the standardized residual; and the likelihood gradient $\ell_{;y} = (X\beta - y)/\sigma^2$ gives the score gradient

$$H = (X/\hat{\sigma}^2, \hat{z}^0/\hat{\sigma}^3)$$

where $\hat{\sigma}^0$ is written just $\hat{\sigma}$ for convenience; and $\hat{j} = diag\{X'X/\hat{\sigma}^2, n/2\hat{\sigma}^4\}$. Together these give

$$W(\theta) = \begin{pmatrix} I & (\beta - \hat{\beta}^0)/2\sigma^2 \\ 0 & \hat{\sigma}^2/\sigma^2 \end{pmatrix} \tag{3.3}$$

with inverse

$$M(\theta) = \begin{pmatrix} I & -(\beta - \hat{\beta}^0)/2\hat{\sigma}^2 \\ 0 & \sigma^2/\hat{\sigma}^2 \end{pmatrix} \tag{3.4}$$

Now suppose we have $\theta = (\psi, \lambda)$ and wish to examine the linearity or curvature of ψ at the data point y^0. For this we use more than likelihood: we use the direct parameter effect at the data y^0 as described by the Jacobian matrixes $W(\theta)$ and $M(\theta)$. The maximum likelihood value $\hat{\theta}^0 = (\hat{\psi}^0, \hat{\lambda}^0)$ provides a reference value and for change with $\psi = \hat{\psi}^0$ held fixed we examine how $d\lambda$ at $\hat{\lambda}^0$ affects the data point; we obtain

$$\begin{aligned} d\hat{\theta} &= \begin{pmatrix} w_{11}(\hat{\theta}^0) & w_{12}(\hat{\theta}^0) \\ w_{12}(\hat{\theta}^0) & w_{22}(\hat{\theta}^0) \end{pmatrix} \begin{pmatrix} 0 \\ d\lambda \end{pmatrix} \\ &= \begin{pmatrix} w_{12}(\hat{\theta}^0) \\ w_{22}(\hat{\theta}^0) \end{pmatrix} d\lambda. \end{aligned}$$

We then examine how this data change $d\hat{\theta}$ from $(\hat{\psi}^0, \hat{\lambda}^0)$ to $(\hat{\psi}^0, \hat{\lambda}^0) + d\hat{\theta}$ affects the parameter at various θ, by using the inverse relationship $d\theta = M(\theta)d\hat{\theta}$. If in moderate deviation, we have that this $d\theta$ produces no ψ change, then we have linearity. If however there is ψ change then we can use the generalized curvature measures from the proceeding section and obtain our new sample space–based measure of curvature.

4 An example illustrating curvature

The Normal $\{(\theta_1, \theta_2); I\}$ example exemplifies the location invariance used in Bayes (1763) proposal for statistical analysis. To go beyond global location invariance and yet still, for ease of calculation, have familiar normality we examine the case of sampling from the Normal (μ, σ^2) distribution and explore several scalar parameters of interest.

The curvature criteria proposed in Section 3 show that the interest parameters $\psi = \mu$, $\psi = \sigma$ and $\psi = \mu + k\sigma$ have linearity; accordingly we use coordinate axes μ and σ and thus have that a linear parameter is a straight line in the (μ, σ) coordinates. We then show that the parameters $\psi = \mu/\sigma^2$ and $\psi = \mu + k\sigma^2$ have nonlinearity. Now consider this in more detail.

Just consider linearity properties of the parameter $\psi = \mu + k\sigma$ for given k. To simplify notation we take $(\hat{\mu}^0, \hat{\sigma}^0) = (0, 1)$ without loss of generality due to location-scale properties

of the model. We also for convenience take the parameter $\theta = (\mu, \sigma)$ rather than the earlier (μ, σ^2); this requires a modification of $W(\theta)$ and $M(\theta)$ in (3.3) and (3.4) giving

$$\begin{pmatrix} d\hat{\mu} \\ d\hat{\sigma} \end{pmatrix} = \begin{pmatrix} 1 & -\mu/\sigma \\ 0 & 1/\sigma \end{pmatrix} \begin{pmatrix} d\mu \\ d\sigma \end{pmatrix} \tag{4.1}$$

$$\begin{pmatrix} d\mu \\ d\sigma \end{pmatrix} = \begin{pmatrix} 1 & \mu \\ 0 & \sigma \end{pmatrix} \begin{pmatrix} d\hat{\mu} \\ d\hat{\sigma} \end{pmatrix} \tag{4.2}$$

Following Fraser, Reid, Marras & Yi (2009) we then consider a change $(d\mu, d\sigma)$ at $\theta = \hat{\theta}^0$ that leaves the parameter ψ unchanged; thus $d(\mu + k\sigma) = 0$ or $d\mu = -kd\sigma$ at $(0,1)$. From (4.1) the corresponding increasement at $\hat{\theta} = (0,1)$ is

$$\begin{pmatrix} d\hat{\mu} \\ d\hat{\sigma} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} d\mu \\ d\sigma \end{pmatrix} = \begin{pmatrix} -k \\ 1 \end{pmatrix} d\sigma$$

giving $d\hat{\mu} = -kd\hat{\sigma}$. We then determine what this sample space change $d\hat{\mu} = -kd\hat{\sigma}$ implies at a general point (μ, σ); from (4.2) we obtain

$$\begin{pmatrix} d\mu \\ d\sigma \end{pmatrix} = \begin{pmatrix} 1 & \mu \\ 0 & \sigma \end{pmatrix} \begin{pmatrix} -k \\ 1 \end{pmatrix} d\hat{\sigma}$$

or

$$\frac{d\mu}{d\sigma} = \frac{\mu - k}{\sigma}$$

with initial condition $(\mu, \sigma) = (0,1)$. Integration of the equation gives $\mu - k = -k\sigma$ or $\mu + k\sigma = k$. This shows that the quantile $\mu + k\sigma$ is linear, and in particular that μ and σ are linear.

Now we consider the parameter $\psi = \mu/\sigma^2$. Figure 3 records some contours for this parameter together with an observed data point. The contours do look curved with respect to μ and σ and they are still curved according to our current criteria; of course $\psi = 0$ is special and corresponds to $\mu = 0$ and does have linearity.

For illustration we consider a data point y^0, with $\hat{\mu} = \bar{y}^0 = 0.975442, \hat{\sigma} = \sqrt{(n-1)/n} s_y^o = 1.226137$ and $n = 3$. Figure 3 records the contour of $\psi = \hat{\psi} = 0.6488188$ where $\hat{\psi}$ is the maximum likelihood value of $\hat{\mu}/\hat{\sigma}^2$; it also records the observed $(\hat{\mu}, \hat{\sigma})$. The Bayes posterior survival value $s(\hat{\psi})$ is obtained by integrating the observed likelihood $L^o(\mu, \sigma)$ with an

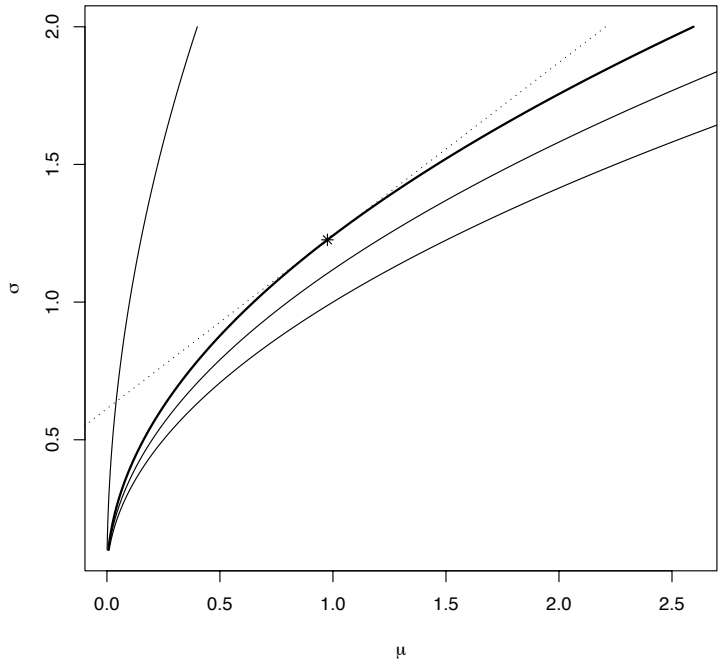

Figure 3: Some contours or level lines for the parameter $\psi = \mu/\sigma^2$ when $\hat{\mu} = 0.975442$, $\hat{\sigma} = 1.226137$ and $\hat{\psi} = 0.6488188$. The contour of $\hat{\psi}^0$ goes through the data $(\hat{\mu}^0, \hat{\sigma}^0)$; The dotted line records the linear parameter contour through $(\hat{\mu}^0, \hat{\sigma}^0)$.

appropriate weight function or prior $\pi(\mu, \sigma)$. This location scale normal example does not fit directly with Bayes original location or translation analysis but an extended version can be found in Jeffreys (1946). The original Jeffreys (1939) involved the prior $\sigma^{-2} d\mu d\sigma$ but was modified in the later paper to the prior $\sigma^{-1} d\mu d\sigma = d\mu d \log\sigma$. For some current views on these two priors see Fraser (2009). The modified Jeffreys prior has been found to have acceptable Bayesian properties, provided the parameter of interest has the linearity defined in Fraser, Reid, Marras & Yi (2009) and developed further here.

Accordingly, for Bayesian inference, we have the posterior distribution.

$$\begin{aligned}\pi(\mu,\sigma|y^0)d\mu d\sigma &= cL^o(\mu,\sigma)\sigma^{-1}d\mu d\sigma \\ &= \frac{(s^2)^a}{\Gamma(a)2^{a-1}}\sigma^{-n}e^{-s^2/2\sigma^2}\frac{\sqrt{n}}{\sqrt{2\pi}\sigma}e^{n(\mu-\bar{y}^0)^2/2\sigma^2}d\mu d\sigma,\end{aligned}$$

where $a=(n-1)/2=1$ and $s^2=\sum_{i=1}^n(y_i-\bar{y}^0)^2=4.510235$. This posterior distribution can be presented in terms of generic variables, a standard normal variable z and an independent chi variable χ with $n-1$ degree of freedom:

$$\sigma=\sqrt{n}\hat{\sigma}^o\chi^{-1},\quad \mu=\hat{\mu}-\hat{\sigma}z\chi^{-1}.$$

For the parameter $\psi=\mu/\sigma^2$ we do the integration numerically using software **R** over the region of the parameter space to the right of various ψ contours,

$$s(\psi)=\int_{\mu/\sigma^2>\psi}cL^o(\mu,\sigma)\sigma^{-1}d\mu d\sigma$$

and obtain the posterior survival plot for $s(\psi)$ recorded in Figure 4. The preceding calculation represents standard Bayesian calculations; these are routine for the present parameter $\psi=\mu/\sigma^2$ and are equally routine for other interest parameters. A typical modification would use Markov chain Monte Carlo in place of the numerical integration, but $N=100$ million simulations can typically give just second decimal accuracy; for such McMC simulations see Bédard, Fraser & Wong (2008) and for some related discussion see Fraser, Wong, & Sun (2009).

There does not seem to be an obvious and immediate frequentist calculation to give a p-value. This familiar common lack of an available frequentist procedure is a common complaint from the Bayesian approach.

We do have however the higher order procedures available from recent likelihood asymptotics. For the present interest parameter ψ we can calculate the signed likelihood root r for assessing ψ and a special maximum likelihood departure q with nuisance information adjustment.

$$\begin{aligned}r &= sign(\hat{\psi}-\psi)[2\{\ell(\hat{\theta}^o)-\ell(\hat{\theta}_\psi)\}]^{1/2} \\ q &= sign(\hat{\psi}-\psi)|\hat{\chi}-\hat{\chi}_\psi|\left(\frac{|j_{\varphi\varphi}(\hat{\theta})|}{|j_{(\lambda\lambda)}(\hat{\theta}_\psi)|}\right)^{1/2}\end{aligned}$$

These involve the overall mle $\hat{\theta}$ and the constrained mle $\hat{\theta}_\psi$ as with the likelihood ratio quantity, but also the full information $j_{\varphi\varphi}$ for the canonical parameter $(\mu/\sigma^2,1/\sigma^2)$, the

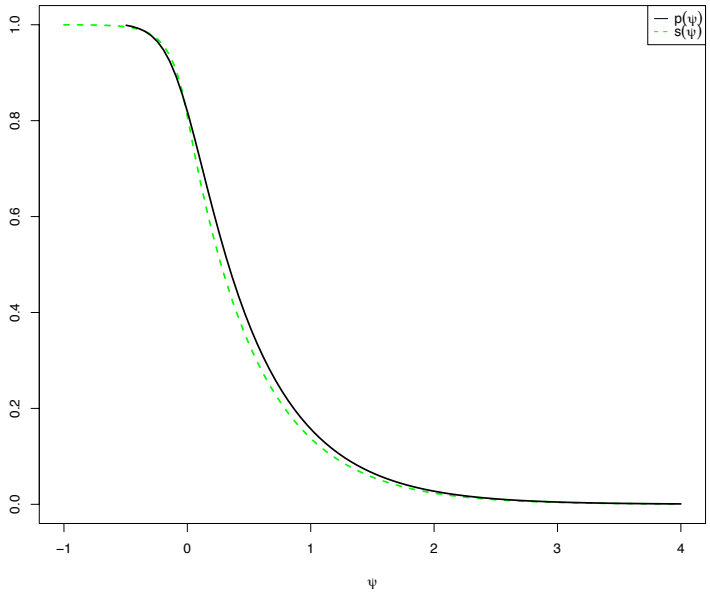

Figure 4: The posterior survival function $s(\psi)$ for $\psi = \mu/\sigma^2$ for the given data; and the third order $p(\psi)$ for the same data.

nuisance information $j_{(\lambda\lambda)}$ rescaled to the canonical parameter, and a canonical parameter departure $\hat{\chi} - \hat{\chi}_\psi$ from the constrained maximum likelihood value. The third order p-value for ψ is then

$$p(\psi) = \Phi\left(r - \frac{1}{r}\log\frac{r}{q}\right).$$

We do not verify here these third order p-values but refer to extensive simulations in the literature that record high accuracy. Bédard, Fraser, & Wong (2008), Fraser, Wong & Sun (2009), Fraser (1990), Fraser, Reid & Wu (1999) and Fraser, Reid & Wu (1999). The third order p-value function is also plotted in Figure 4.

For our present data set and parameter $\psi = \mu/\sigma^2$ we see that $p(\psi) > s(\psi)$ uniformly. This equality is in the opposite direction from our introductory example in Section 1; and not surprisingly we see that our parameter is concave to the right, the opposite from that in Section 2.

5 The example continued

We continue with our Normal (μ, σ^2) example, with interest parameter $\psi = \mu/\sigma^2$ and data $\hat{\mu} = 0.975442,\ \hat{\sigma} = 1.226137,\ \hat{\psi} = 0.6488188$. The Bayes survival function $s(\psi)$ and the frequentist p-value $p(\psi)$ are plotted in Figure 4; and they are different; and they are different oppositely from what was found for the initial example in Section 1. And we have attributed this opposite effect to the curvature; convex leading with increasing interest ψ in the initial case and then concave leading with increasing ψ in the present example. We now discuss this further and explore a global curvature measure and its consequences.

Our approach is to examine properties on the parameter space and to focus there on the parameter of interest ψ. For this we have a 2 dimensional full parameter and a 1 dimensional interest parameter; we restrict attention to this case following a similar restriction in Efron (1975).

We can write the full parameter as $\theta = (\mu, \sigma)$ in terms of familiar components, or as $\varphi(\theta) = (\varphi_1, \varphi_2) = (\mu/\sigma^2, 1/\sigma^2)$ in terms of the exponential canonical parametrization. For our present purpose, it would be helpful to be able to write it in a location–type parameterizations, say as (β_1, β_2) but what would be the basis for such a parameterization? Large sample analysis (Cakmak, Fraser & Reid 1994) suggests that there is a location model as in Section 1 but such does not address the regression type structure found with our present Normal (μ, σ^2) example; even with this change there are consequences that do not follow the pattern seen with $y_1 - \psi$, $y_2 - \lambda$ in Section 3; for some differential consequences indicated by $W(\theta)$ and $M(\theta)$ see (3.1, 3.3, 3.4).

For a curvature measure we need to standardize the coordinates locally, as described in Section 2. For our present example, we have an exponential model and the expected information $i(\theta)$ is available as for the Efron curvature. More generally if we were to work in moderate deviations, we would have a local canonical parametrization $\varphi(\theta) = (d/dV)\ell(\theta; y^0)$ obtained from the sensitivity array $V(\theta)$ in (3.2). This gives an exponential canonical parameterization if the model is exponential and otherwise gives an approximate version of such. For some background, see Fraser, Wong & Sun (2009), and Fraser, Reid, Marras & Yi (2009). In the more general context, we would use $j(\theta) = j_{\varphi\varphi}(\theta)$ as the information function with special construction details in Fraser, Reid, Marras, and Yi (2009).

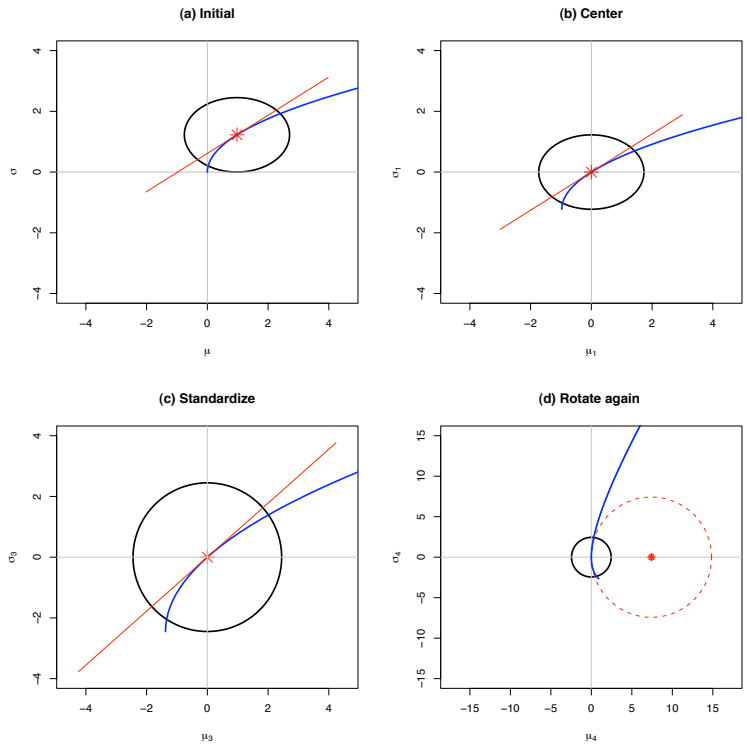

Figure 5: In each component plot we have recorded the linear contour that locally coincides with the parametric ψ at the observed maximum likelihood value $\hat{\theta}^0$. Going from figure to figure straight lines remain straight lines so we are now in a position to assess the numerical curvature of the contour $\psi = \hat{\psi} = 0.6488188$. In Figure 5 d, we have plotted the circle that coincides with the given $\psi = \hat{\psi}^0$ to the second order at the data point. The curvature is $\hat{\gamma}^0 = 0.1346722$ with radius of curvature $\hat{\rho}^0 = 7.425437$.

From the method given by (2.1) we have

$$\begin{aligned} & \begin{pmatrix} d\theta_1 \\ d\theta_2 \end{pmatrix}' i_{\varphi\varphi} \begin{pmatrix} d\theta_1 \\ d\theta_2 \end{pmatrix} \\ = & \begin{pmatrix} d\mu \\ d\sigma \end{pmatrix}' \begin{pmatrix} n/\sigma^2 & 0 \\ 0 & 2n/\sigma^2 \end{pmatrix} \begin{pmatrix} d\mu \\ d\sigma \end{pmatrix} \end{aligned}$$

and see that the parametric coordinates as given are orthogonal; see Figure 5. We then center the parameter as in (b); use a root information to rescale as in (c); then rotate as in (d) so that the ψ parameterization increases to the right. In each case, we show the data point, the likelihood contour at 2 standard units and the contour $\psi(\theta) = \hat{\psi}^0$.

6 The example: Linearity and curvature

For the discussion of the Normal (μ, σ^2) example, we have used $\theta = (\mu, \sigma)$ and presented plots in terms of coordinates μ and σ. We could have used other coordinates such as $(\mu, \log\sigma)$ or (μ, σ^2) but the present choice has the advantage of linearity. The development of linearity in Fraser, Reid, Marras & Yi (2009) was mentioned as general background for the Bayesian–frequentist difference discussed in Fraser (2009), and the Normal (μ, σ^2) was the next choice for exhibiting departure from the simplicity of the Normal–on–the–plane. Then straight from the definition of linearity we showed that μ, σ and even the general quantile $\mu + k\sigma$ for given k are linear to second order. This then gives us the background to assess the curvature of $\psi = \mu/\sigma^2$.

In each component plot in Figure 5, we have recorded the observed maximum likelihood value, $\hat{\theta}^0$, the maximum likelihood contour $\psi(\theta) = \hat{\psi}^0$ for the interest parameter ψ and the tangent linear parameter at $\theta = \hat{\theta}^0$. The changes from (a) to (b) to (c) to (d) leave straight lines straight. From (a) to (b), we center at the maximum likelihood value; from (b) to (c) we standardize using root observed information; and from (c) to (d) we rotate so that the tangent to the interest parameter is vertical. In each case we record the 2 standard deviation likelihood ellipse about $\hat{\theta}^0$. And then finally in plot (d) we present the fitted circle at the maximum likelihood value. This gives the observed curvature $\hat{\gamma}^0 = 0.1346722$ and radius $\hat{\rho}^0 = 7.425437$.

7 Discussion and future directions

Statistics as a discipline has two prominent methodologies that can, in common areas of application, give significantly different conclusions from the same model-data input. The methodologies are the Bayesian (1763) that makes use of what we might now call default or invariant priors, and the frequentist that is perhaps older but was given modern formalism (Fisher, 1930; Neyman, 1937) with the proposal for what is now called confidence. How can a discipline hope for serious respect from the larger scientific community when it has two logics that extract different conclusions from the same input information; if the conclusions are contradictory then one or other or both of the so-called logics must be defective! Some recent thoughts on this conflict may be found in Fraser (2009) where it is argued that the Bayes posterior in suitable circumstances can be viewed as an approximation to confidence. In this paper we have addressed this Bayesian frequentist difference by considering the Normal on the plane where the Bayesian frequentist difference clearly focuses on curvature in the parameter of interest; and we have raised the question whether the quick-and-easy of the default Bayes approach can be upgraded so as to reduce the misrepresention concerning repetition properties of posterior quantiles. For this we have developed an algorithm for measuring curvature, an algorithm that leads to a curvature measure that at higher order can differ from the familiar Efron curvature. Part of defining curvature is to first close on what linearity of parameters means when the canonical parameters of a full exponential model are linear in the Efron sense whereas the canonical parameters of a location model are linear in the proposed sense (Fraser, Reid, Marras & Yi, 2009); the reliability of the location model used with the Bayes approach argues for the merits of linearity in the location sense.

An initial hope was that the discrepancy function in Section 1 for the normal circle could be applied directly to upgrade the Bayesian, and improve the repetition reliability of resulting summary inference statements. For this, the null model has location normal properties to first order when standardized, and the curvature and the nonnormality are each of second order; but there seem to be subtle interactions between these second order effects.

The linearity in the present location sense however allows some removal of the Bayes bias. In Figure 3 we have recorded the linear parameter that coincides with the interest parameter at the constrained maximum. By taking the linear parameter as a first step towards repetition reliability we can partially correct the Bayes bias. In Figure 6 we have included the Bayes posterior survivor function for the linear parameter and then pondered whether

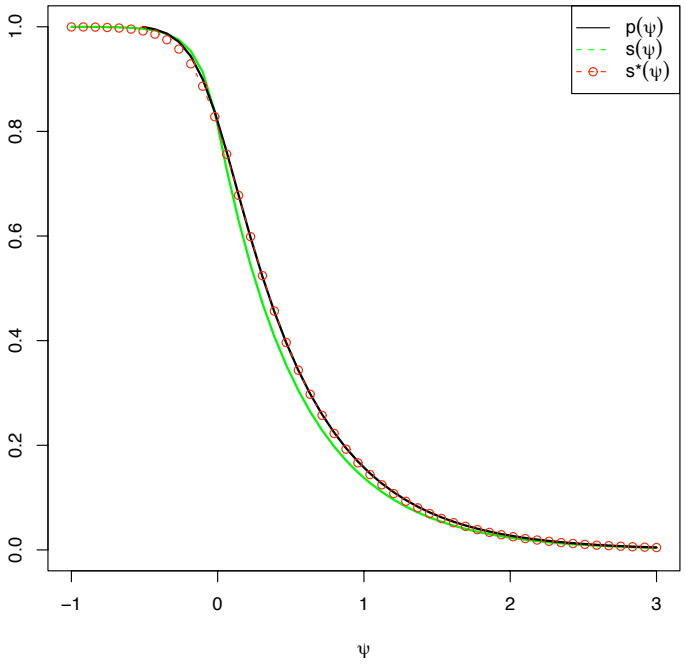

Figure 6: p-value for ψ, $p(\psi)$; Bayes survivor for ψ, $s(\psi)$; and linearly corrected survivor for ψ, $s^*(\psi)$

doubling the correction to get a nominal p-value would be effective; when compared with the third order p-values this seems to be an over correction. Thus we would recommend at this stage just the correction provided by the step to the linear parameter survivor value.

The Bayesian frequentist difference represents a huge challenge for our discipline. and to seek ways to improve the repetition reliability should take prominence. We have attempted a recognition of this urgency.

In this paper we have approached the Bayesian frequentist difference by focussing on parameter curvature as the prime source for breakdown in the Bayes methodology, by proposing a measure of curvature, and by developing a simple correction procedure for an elementary statistical problem. In doing this we feel we are pointing a direction for research and that only the preliminaries have been touched.

The directions suggest an explicit formula for the curvature, an explicit formula for the linear correction, simulation studies to evaluate the corresponding correction accuracy, and hopefully a full second order correction for Bayes curvature bias.

Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada, and we view this as a preliminary report of location based curvature. Deep gratitude and appreciation go to the participants in a research seminar that focussed on the role of curved parameters in Bayesian and frequentist inference: T. Cai, F. Chang, K. Ji, W. Lin, M. Mallo, R. Thinniyam, A. Wong, Y.Y. Wu. And very special thanks to K. Ji for support with the development of the manuscript.

REFERENCES

[1] Andrews, D.F., Fraser, D.A.S., and Wong, A. (2005). Computation of distribution functions from likelihood information near observed data. *Journal Statist Plann. Inference* **134**, 180-193.

[2] Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. Phil. *Trans. Roy. Soc. London* **53**, 370-418; **54**, 296-325. Reprinted in *Biometrika* **45** (1958), 293-315.

[3] Bédard, M., Fraser, D.A.S. and Wong, A. (2008). Higher accuracy for Bayesian and frequentist inference: Large sample theory for small sample likelihood. *Statistical Science* **22** , 301-321.

[4] Cakmak, S., Fraser, D.A.S., McDunnough, P., Reid, N., and Yuan, X. (1998). Likelihood centered asymptotic model: exponential and location model versions. *Int. J. Math. & Stat. Sci.* **4**, 211-222.

[5] Cakmak, S., Fraser, D.A.S., and Reid, N. (1994). Multivariate asymptotic model: exponential and location approximations. *Utilitas Mathematica* **46**, 21-31.

[6] David, A. P., Stone, M. and Zidek, J. V. (1973). Marginalization paradoxes in Bayesian and structural inference. *J. Roy. Statist. Soc. B* **35**, 189-233.

[7] Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). *Ann. Statist.* **3**, 1189-1242.

[8] Fisher, R. A. (1930). Inverse probability. *Proc. Camb. Phil. Soc.* **26**, 528-535.

[9] Fraser, D.A.S. (1990). Tail probabilities from observed likelihoods. *Biometrika* **77**, 65-76.

[10] Fraser, D.A.S. (2009). Is Bayes posterior just quick and dirty confidence? *Statistical Science*, in review.

[11] Fraser, A.M, Fraser, D.A.S & Staicu, A.-M. (2009). The second order ancillary. *Bernoulli*, in revision.

[12] Fraser, D.A.S., and Reid, N. (2002) Strong matching of frequentist and Bayesian parametric inference. *Journal of Statistical Planning and Inference* **103**, 263-285.

[13] Fraser, D.A.S., Reid, N. and Wu, J. (1999). A simple general formula for tail probabilities for frequentist and Bayesian inference. *Biometrika* **86**, 249-264.

[14] Fraser, D.A.S., Reid, N., Marras, E., and Yi, G. Y. (2009). Default prior for Bayesian and frequentist inference. *J. Roy. Statist. Soc. B*, In revision.

[15] Fraser, D.A.S., Wong, A. and Sun, Y. (2009). Three enigmatic examples and inference from the likelihood. *Canadian Journal of Statistics* **37**, 161-181.

[16] Jeffreys, H. (1939). Theory of Probability. Oxford: Oxford University Press. Third Edition.

[17] Jeffreys, H. (1946). An invariant form for the prior probability in estimation problem. *Proc. Roy. Soc. A.* **186**, 453-461.

[18] Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Phil.Trans. Roy. Soc. A* **237**, 333-380.

[19] Stainforth, D. A., Allen, M. R., Tredger, E. R. and Smith, L. A. (2007). Confidence, uncertainty and decision-support relevance in climate predictions. *Phil. Trans. Roy. Soc*. A, **365**, 2145-2162. See also: Gambling on tomorrow. Modelling the Earth's climate mathematically is hard already. Now a new difficulty is emerging. *Economist*. August 18, 2007, p 69.