Bernoulli **10**(4), 2004, 639–649

On the global geometry of parametric models and information recovery

PAUL MARRIOTT[1] and PAUL VOS[2]

[1]*Dept. of Statistics and Applied Probability, National University of Singapore, 3 Science Drive 2, Singapore 117543. E-mail: paul@stat.duke.edu*
[2]*Biostatistics Department, SAHS, East Carolina University, Greenville NC 27858-4353, USA. E-mail: VOSP@MAIL.ECU.EDU*

We examine the question of which statistic or statistics should be used in order to recover information important for inference. We take a global geometric viewpoint, developing the local geometry of Amari. By examining the behaviour of simple geometric models, we show how not only the local curvature properties of parametric families but also the global geometric structure can be of crucial importance in finite-sample analysis. The tool we use to explore this global geometry is the Karhunen–Loève decomposition. Using global geometry, we show that the maximum likelihood estimate is the most important one-dimensional summary of information, but that traditional methods of information recovery beyond the maximum likelihood estimate can perform poorly. We also use the global geometry to construct better information summaries to be used with the maximum likelihood estimate.

Keywords: ancillarity; asymptotic analysis; geometry; global geometry; information recovery; Karhunen-Loève decomposition; likelihood

1. Introduction

In this paper we examine the question of which statistic or statistics should be used in order to recover information important for inference. A geometric approach is taken throughout, and a new technique for selecting highly informative statistics is proposed. In contrast to Amari's (1990) work, this selection is not done using only local information; rather a global approach is taken, and the consequences of this are explored. The tool used to understand the global structure is a functional version of principal component analysis, with the spectrum giving important information about the existence, or otherwise, of good low-dimensional summary statistics.

All examples used in this paper have the structure of curved exponential families; for regularity conditions, see Kass and Vos (1997, p. 27). The examples are chosen to illustrate the geometric aspects of information recovery rather than to be statistically realistic. Our aim in this short paper is only to introduce broad ideas; further issues for consideration are discussed in the final section.

Example 1 *Helix model.* This example is considered in Amari (1990), where it is called the

1350–7265

'spiral model'. The helix model comprises trivariate normal densities denoted by $MN_3(\mu, I_{3\times 3})$, for which μ is restricted to the helix $(r\cos\theta, r\sin\theta, \theta d)$. We parametrized by $\theta \in \Theta = \mathbb{R}$, with r, d being fixed and known and the data being denoted by $X = (X_1, X_2, X_3)$.

***Example 2** tanh-link model.* The hyperbolic link is useful for modelling data that approaches an asymptote; see, for example, Vos (1991). The model is given by the nonlinear regression

$$Y_i = \tanh(\beta X_i) + \epsilon_i,$$

where the error term has an independent Gaussian distribution with a fixed, assumed known, variance.

2. Spectral decomposition of the likelihood

In this paper the phrase *information in the statistic* S is used as an abbreviation for the phrase 'information in the statistic S for making inferences about the parameter θ'. That is, the concept of information depends both on the data and on the structure of the model. All examples in this paper are curved exponential families where the maximum likelihood estimate (MLE), $\hat{\theta}$, alone is not sufficient; we therefore look for a statistic of the form $(\hat{\theta}, A)$ which recovers more information. A general principle which this paper follows is that if the statistic $(\hat{\theta}, A)$ is highly informative about θ, then knowing the values of this statistic determines all the important parts of the log-likelihood function $\ell(\theta; x)$. One of the issues that the paper therefore considers is what 'important' means in this context. Since log-likelihoods are only defined up to an additive term, independent of θ but possibly dependent on the data, one form of variation of $\ell(\theta; x)$ which is unimportant is addition of such terms. Furthermore, since we are interested in inference, variation of parts of the log-likelihood which are negligible in comparison to its maximum value will also be considered unimportant. To formalize these ideas we first normalize the log-likelihood, either by dropping additive constants, or by considering $\ell(\theta; x) - \ell(\phi; x)$, where $\phi \in \Theta$ and $p(x, \phi)$ is the data-generation process. Secondly, define the truncated log-likelihood by

$$\ell_D(\theta; x) = \begin{cases} \ell(\theta; x) - \ell(\phi; x), & \text{if } \theta \in D, \\ L, & \text{otherwise,} \end{cases} \tag{1}$$

where L is a negative number and D is the region where $\ell_D(\theta; x)$ is 'large'.

In order to more formally capture the variation of $\ell_D(\theta; x)$, and hence information loss, consider $\ell_D(\theta; x)$ as a function-valued random variable. It is then natural to consider its variance–covariance structure,

$$G_D^{\phi}(\theta_1, \theta_2) = \mathrm{cov}_{\phi}[\ell_D(\theta_1; x), \ell_D(\theta_2; x)]. \tag{2}$$

For example, for a full exponential family, with θ the natural parameter, a simple calculation gives

$$G_D^{\phi}(\theta_1, \theta_2) = \begin{cases} (\phi - \theta_1)' I_\theta(\phi)(\phi - \theta_2), & \theta_1, \theta_2 \in D, \\ 0 & \text{otherwise.} \end{cases}$$

where

$$I_\theta(\phi) := \mathrm{cov}_\phi\left(\frac{\partial}{\partial\theta}\ell(\phi; x)\right).$$

Example 1 (*continued*). In this example the covariance is

$$G_D^{\phi}(\theta_1, \theta_2|\phi = 0) = -r^2\cos(\theta_1) - r^2\cos(\theta_2) + r^2\cos(\theta_1 - \theta_2) + r^2 + d^2\theta_1\theta_2.$$

Since this paper is concerned with understanding the sources of variation in the log-likelihood, a functional version of principal component analysis is constructed using the Karhunen–Loève (KL) decomposition; see Papoulis (1984). This allows a 'diagonalization' of variance–covariance functions such as (2), and finds sources of large variability. Consider, therefore, the eigenvalue equation

$$\int_D G_D^{\phi}(\theta_1, \theta_2)\psi^{\phi}(\theta_2)\,\mathrm{d}\mu(\theta_2) = \lambda^{\phi}\psi^{\phi}(\theta_1) \tag{3}$$

over the region D, which here is assumed compact. To ensure invariance to reparametrization the integration is done with respect to a measure such as $\mathrm{d}\mu(\theta) = |I(\theta)|^{-1/2}\mathrm{d}\theta$, where $I(\theta)$ is the expected Fisher information, although other choices are possible. When there is no chance of ambiguity we suppress the dependence of $\psi^{\phi}(\theta)$ and λ^{ϕ} on ϕ.

By spectral theory (Rudin 1973, pp. 305–311), if D is compact and G continuous there exists a countable set of eigenfunctions $\{\psi_i\}$ with eigenvalues $\{\lambda_i\}$ ordered such that $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant 0$. These eigenfunctions can be chosen to form an orthonormal basis for the set of smooth functions from $D \subset \Theta$ to $\mathbb{R}$ with respect to the inner product defined by

$$\langle f, g\rangle = \int_D f(\theta)g(\theta)\,\mathrm{d}\mu(\theta).$$

Writing the log-likelihood with respect to this basis gives (Noting that in our examples the number of non-zero eigenvalues, n, is finite due to the finite-dimensional sufficient statistics) a representation

$$\ell_D(\theta; x) = \sum_{i=1}^{n} s_i(x)\psi_i^{\phi}(\theta) + C^{\phi}(\theta), \tag{4}$$

with the first terms in the sum contributing most to the data variability of $\ell_D(\theta; x)$ and the last term being independent of the data. Using this analysis, low-dimensional affine spaces, spanned by the first few eigenfunctions, can be found that provide a good approximation to the log-likelihood.

3. Examples

Example 1 (*continued*). We apply the above theory to the helix model when $r = 0.1$ and $d = 0.1$, having defined the region D to be $[-30, 30]$. This choice of r and d gives a high statistical curvature and high torsion; the curved model is in fact uniformly close to a one-dimensional model. Figure 1(a) shows several realizations of the log-likelihood function for data generated when $\phi = 0$. For $G_D^{\phi}(\theta_1, \theta_2|\phi = 0)$ there are exactly three non-zero eigenvalues, corresponding to the three-dimensional sufficient statistic for this model. The eigenvalues have been calculated numerically and in this example are in the ratio 612:1:1. Thus the important issue is that there is one dominant eigenfunction which is responsible for almost all of the variation. This eigenfunction is also calculated numerically and shown in

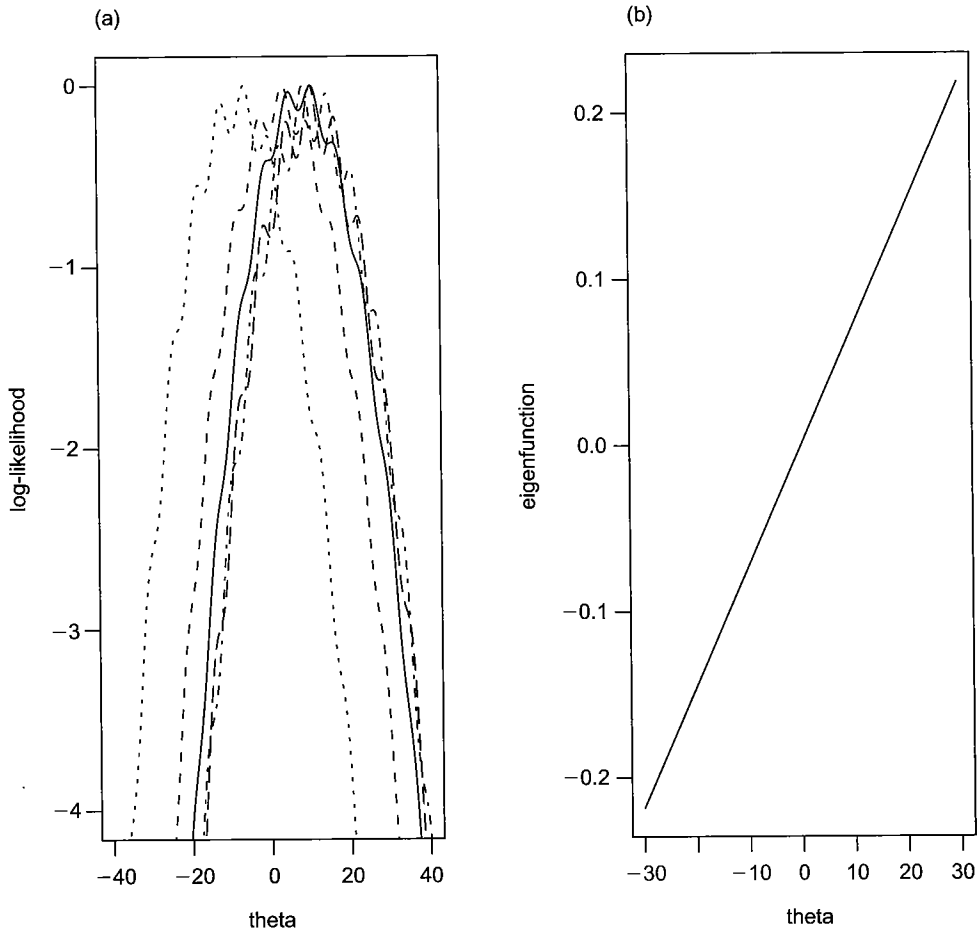

Figure 1. The helix model for $r = 0.1$, $d = 0.1$. (a) Some realizations of the log-likelihood. (b) The dominant eigenfunction.

Figure 1(b). This approximately linear function corresponds to variations in the log-likelihood of the form

$$s_1(x)\theta + \mathrm{C}(\theta),$$

where $C(\theta)$ is data-independent. Figure 1(a) shows visually that for this model the log-likelihood is approximately quadratic, at least in the sense that a quadratic function is uniformally close. Furthermore, by observing the vertical scale it is clear that all perturbations from a quadratic function are inferentially unimportant.

The dominant variation in this case has the effect of translating the log-likelihood while leaving the shape of the log-likelihood essentially unaltered. That is in statistical terms simply a translation of the MLE. Thus for this model there seems to be an approximately sufficient one-dimensional statistic which is one-to-one with the MLE.

This result is remarkable because it shows how the global geometry can completely dominate the local geometry. In terms of Efron's (1975) statistical curvature this model has a curvature of 5, which would normally be considered large enough to have an important effect on inference. In particular, it is normally interpreted as meaning that the MLE alone is not a good approximately sufficient statistic. However, by considering its global geometry it is clear that the model lies uniformly very close to a one-dimensional full exponential family given by $(\mu_1, \mu_2, \mu_3) = (0, 0, \theta)$. The curved model is wrapped around a very narrow cylinder which encloses this model. Despite the high local curvature, the KL decomposition automatically picks up this global geometric structure and gives the correct information summary.

In the literature three possible principles for recovering information beyond the MLE are frequently considered:

(i) Expected information loss: choose a statistic A that makes the expected information loss of $T = (\hat{\theta}, A)$ small, where the expected information loss in a statistic T is defined to be $\Delta I^T = I^X - I^T$, in which I^W is the expected Fisher information for the statistic W, that is,

$$I^W(\theta) = -\mathrm{E}_\theta\left(\frac{\partial^2}{\partial\theta^2} l_W(\theta;\, W)\right), \tag{5}$$

and $l_W(\theta;\, W) = \log p_W(W, \theta)$.

(ii) Observed information: record the observed information and use $T = (\hat{\theta}, I_{\mathrm{obs}})$.

(iii) Ancillarity principle: record a statistic A whose distribution is functionally independent of θ so that $T = (\hat{\theta}, A)$.

Example 1 (*continued*). We now consider the three information recovery principles in turn. In this example there is no two-dimensional statistic which has zero expected information loss. Amari's (1990, p. 229) local curvature statistic is used which minimizes the expected information loss. In this example, if $\hat{\theta} = 0$, it is the projection of the sufficient statistic in the direction $(\cos\hat{\theta}, \sin\hat{\theta}, 0)$, hence the principle of minimizing expected information loss selects $A_1 = x_1$. The observed information for this model is $rx_1 \cos\hat{\theta} + rx_2 \sin\hat{\theta} + d$; when

$\hat{\theta} = 0$ this reduces to $I_{obs} = rx_1 + d$. Thus the principle of recording the observed information means we should again record $A_1 = X_1$. Finally, the ancillarity principle tells us to use a statistic whose distribution is independent of θ, for example, $A_2 = X_3 - \hat{\theta}d$.

For particular values of r and d, the poor performance of the statistics A_1 and A_2 for finite sample sizes can be shown by plotting a series of plausibly observed log-likelihood functions, each with the same fixed value of A_i. Poor performance then shows up as large variability in these plots.

Consider first the case where $r = 1$, $d = 1$. Figure 2(a) shows plots of the normalized log-likelihood function, where the statistic $(\hat{\theta}, A_1)$ is held constant, while plausible values for the remaining data are chosen. The observed information in this example has done a good job of describing the log-likelihood in a small neighbourhood of the MLE; however, it has failed to detect the important global properties of the log-likelihood, such as the existence of secondary modes. Figure 2(b) records four plausibly observed likelihoods, all having $(\hat{\theta}, A_2) = (0, -1)$. This time the statistic A_2 cannot distinguish between the case of a

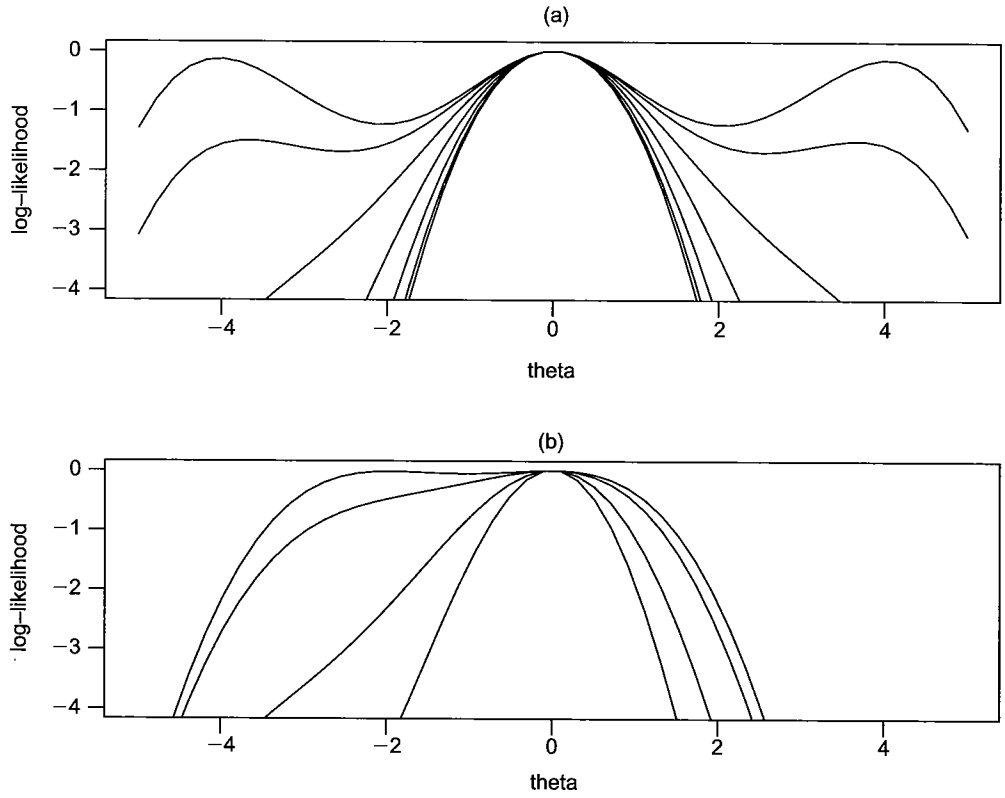

Figure 2. Observed log-likelihoods for helix model for $r = 1$, $d = 1$, showing cases where (a) the statistic A_1 is held constant at 1, (b) A_2 is held constant at -1.

well-defined maximum which has a high value for I_{obs} and the case where the log-likelihood is almost flat across a region of the parameter space.

We now consider the KL decomposition of the information in this example. We use the relative sizes of the three non-zero eigenvalues either to find an informative two-dimensional summary statistic, or to demonstrate its non-existence.

Case 1: $r = 0.5$ $d = 3$. An observed log-likelihood function is plotted in Figure 3(a). This is used to choose the region D, selected to be all θ values whose log-likelihood is within 10 of the maximum. In this case the eigenvalues are in the ratio 180:1:0.03, so there is a dominant eigenfunction shown in Figure 3(b). This function is approximately linear and

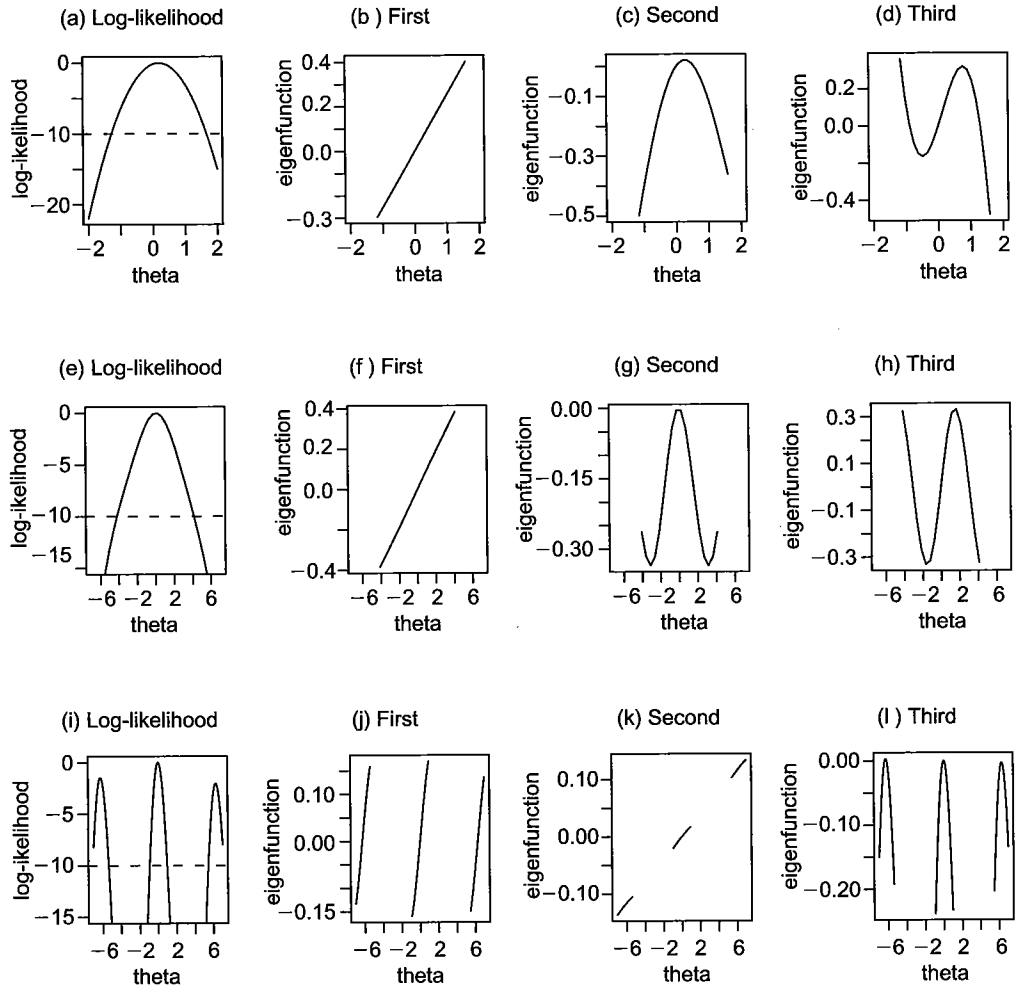

Figure 3. The observed log-likelihoods and the region D, the dominant and secondary eigenfunction for the helix model: (a–d) $r = 0.5$, $d = 3.0$; (e–h) $r = 1$, $d = 1$; (i–l) $r = 5$, $d = 0.3$.

adding multiples of it to the approximately quadratic log-likelihood simply moves the mode. Thus, over 99% of the variation is, as expected, explained by $\hat{\theta}$. Of the remaining 1% of the variation over 97% is explained by the second eigenfunction (Figure 3(c)), and is approximately of the form $s_2(x)(\theta - \hat{\theta})^2 + C(\theta)$. It is immediate that the Hessian term, A_1, is more informative than A_2 with $\hat{\theta}$. The third eigenfunction (Figure 3(d)) changes the symmetry of the log-likelihood and in this case can be neglected.

Case 2: $r = 1$ $d = 1$. In Figure 3(e) the observed log-likelihood is again used to select D. In this case the three eigenvalues are in the ratio 113:35:8, with the eigenfunctions shown in Figure 3(f–h). The MLE by itself has lost a large amount of information, but neither of the remaining eigenfunctions is completely dominant. Thus the KL decomposition indicates that there will not be a good one-dimensional summary statistic, and only a poor two-dimensional one. This confirms the analysis above, which shows that for $r = 1$, $d = 1$ both A_1 and A_2 perform badly.

Case 3: $r = 5$, $d = 0.3$. For this model the region D selected is a disjoint union of intervals (see Figure 3(i)). However, this is not a problem for the KL analysis. The eigenvalues are in the ratio 600:240:75, with the eigenfunctions shown in Figure 3(j–l). The dominant eigenfunction (j) has the effect of translating the local maxima within each local mode. The secondary source of variation (k), however, moves the heights of the modes relative to each other, hence can change the global maximum. The smallest source of variation (l) changes the Hessian term for each mode. Thus the KL decomposition has again used the global geometry to decompose the local and global information.

Example 2 (*continued*). Suppose that $(x_1, x_2, x_3) = (0.4, 1.0, 5.0)$ and we have observed $(y_1, y_2, y_3) = (0.133, 0.510, 0.991)$. Figure 4 shows the result of simulating 50 samples which have the same $\hat{\beta}$. It is clear that the Hessian is almost constant for each simulation and that all the variation is in the global geometry away from a small open set around $\hat{\beta}$. Calculating the corresponding eigenfunctions after fixing the MLE gives a single dominant eigenfunction whose first and second derivatives are zero at the MLE. Hence adding linear multiples of this eigenfunction affects neither the MLE nor the Hessian. This explains the behaviour seen in Figure 4.

The following simulation study shows that simple inference procedures can be based on the statistics suggested by the global geometry and that these inference procedures offer a close approximation to inference based on the full likelihood function. In contrast, simply using the observed information in place of the expected information can lead to a poor approximation of likelihood-based inference. Throughout the study, we test $H_0 : \theta = \theta_0 := 0$ versus $H_A : \theta > \theta_0$.

Example 3 *Modified circle*. In this model the relationship between the mean and the parametrization θ in the trivariate normal family $MN_3(\mu, I_{3\times 3})$ is given by

$$\mu = (r \sin(\theta/w),\ r\cos(\theta/w),\ d\sin(\theta))',$$

where r, d and w are fixed hyperparameters.

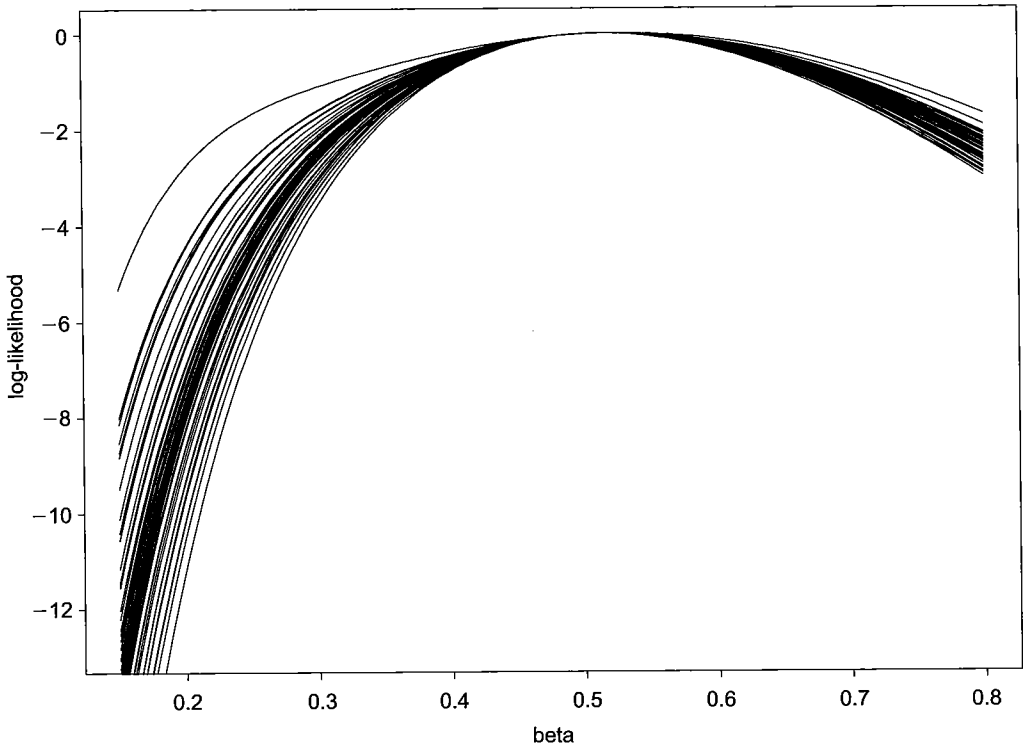

Figure 4. Fifty realizations of the log-likelihood with a constant maximum likelihood estimate for the tanh-link model.

The study compares the signed generalized likelihood ratio test (GLRT) to test statistics of the form

$$(\hat{\theta} - \theta_0)W,$$

where W is either a constant, equal to $\sqrt{I_{\text{obs}}}$, the square root of the observed Fisher information, or equal to the square root of the auxiliary statistic I_{global} suggested by the global geometry. In this example, with $r = 3$, $d = 0.1$, $w = 24$, the global geometry suggests

$$I_{\text{global}} = (r\cos(\hat{\theta}/w)x_1 + r\sin(\hat{\theta}/w)x_2)/w^2,$$

since the significant variation lies in the (x_1, x_2) plane rather than the x_3 direction used by the Hessian.

The critical values for the test statistic defined with $W = \sqrt{I_{\text{global}}}$ can be obtained from the standard normal distribution. Using 1000 samples, the $\alpha = 0.05$ level (i.e., the critical value is $z_{0.95}$) global test rejects the null when the null is true 5.8% of the time for the modified circle model. Using $z_{0.95}$ as critical value for $(\hat{\theta} - \theta_0)\sqrt{I_{\text{obs}}}$ leads to rejection of the null when the null is true 22.8% of the time. This shows that I_{obs} is a poor approximation to the variance of $\hat{\theta}$, but it does not show that I_{obs} is an inferior summarizer

of information. To put these tests on equal ground, the critical value is chosen empirically using 1000 samples under the null hypothesis. This means the power curves for each test have ordinate equal to α at θ_0. Figure 5 shows these power curves. The power curve of $(\hat{\theta} - \theta)\sqrt{I_{\text{global}}}$ is much closer to the power curve of the signed GLRT test than is the test based on $(\hat{\theta} - \theta_0)\sqrt{I_{\text{obs}}}$. Hence the information recovered by the global geometric considerations is what is required by inference and compares well to the 'gold standard' GLRT.

4. Discussion

In this paper we have used the variation of the shape of the log-likelihood function as a way of determining the information content of a statistic. We have not addressed the question of how to extract the information in such a statistic for inference. In the examples we have explicitly avoided conditioning since we have not discussed the ancillary properties of the proposed statistics. We point out, though, that there are important links with the global shape of the log-likelihood function and conditional inference. In particular, the so-called

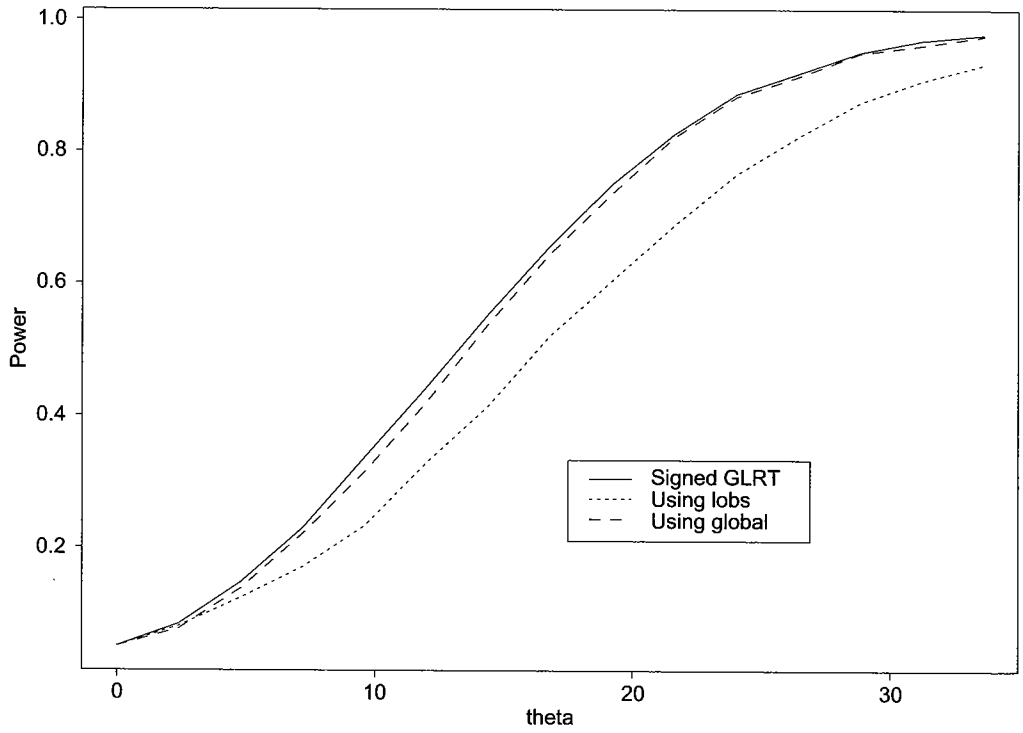

Figure 5. Power curves of the signed generalized likelihood ration test, the test based on I_{obs}, and the test based on I_{global} for the modified circle model.

directed likelihood ancillary (Barndorff-Nielsen and Cox 1994, pp. 227–229; Sweeting 1995; Skovgaard 2001) uses more information about the likelihood than is contained in a local region of its maximum. In general the 'shape' of the log-likelihood function forms a second-order ancillary statistic which can have better properties than the purely locally based ancillaries such as the one proposed by Efron and Hinkley (1978).

The examples in this paper are such that calculation of the required covariance structures is straightforward. In general cases, since these structures are all moments under the data-generation process $p(x, \phi)$, one approach which has worked well in informal studies is to estimate these structures, directly or through bootstrapping. Again this is an area which requires further study.

Acknowledgements

The authors would like to thank the editor and referees for their extremely helpful reading of an earlier version of this paper. Part of this work was undertaken while Marriott was visiting the Institute of Statistics and Decision Sciences, Duke University.

References

Amari, S. (1990) *Differential-Geometric Methods in Statistics, 2nd edn.* Berlin: Springer-Verlag.

Barndorff-Nielsen, O.E. and Cox, D.R. (1994) *Inference and Asymptotics*. London: Chapman & Hall.

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

Efron, B. and Hinkley, D.V. (1978) Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information (with discussion). *Biometrika*, **65**, 457–487.

Kass, R.E. and Vos, P.W. (1997) *Geometrical Foundations of Asymptotic Inference*. New York: Wiley.

Papoulis, A. (1984) *Probability, Random Variables, and Stochastic Processes*. New York: McGraw-Hill.

Rudin, W. (1973) *Functional Analysis*. New York: McGraw-Hill.

Skovgaard, I.M. (2001) Likelihood asymptotics. *Scand. J. Statist.*, **28**, 3–32.

Sweeting, T.J. (1995) A Bayesian approach to approximate conditional inference. *Biometrika*, **82**, 25–36.

Vos, P.W. (1991) Geometry of f-divergence. *Ann. Inst. Statist. Math.*, **43**, 515–537.

Received May 2003 and revised November 2003