difficulty. Indeed the reference prior for the ordered sequence $(\theta_1, \cdots, \theta_m)$ is (see Berger and Bernardo, 1989 for details)

$$\pi_R(\theta_1, \cdots, \theta_m) = \pi(\theta_1)\pi(\theta_2 \mid \theta_1) \cdots \pi(\theta_m \mid \theta_1, \theta_2, \cdots, \theta_{m-1}) = (\pi^{-m}) \prod_{i=1}^{m} [\theta_i^{-1/2}(1 - \textstyle\sum_{j=1}^{i} \theta_j)^{-1/2}],$$

and the corresponding marginal reference distribution for θ_1 is

$$\pi_R(\theta_1 \mid y_1, \cdots, y_m, n) = \pi_R(\theta_1 \mid y_1, n) = Be(\theta_1 \mid y_1 + \tfrac{1}{2}, n - y_1 + \tfrac{1}{2}),$$

no matter how many cells are considered.

ADDITIONAL REFERENCES

BERGER, J. O. and BERNARDO, J. M. (1989). Estimating a product of means: Bayesian analysis with reference priors. *J. Amer. Statist. Assoc.* **84** 200–207.

BERGER, J. O. and BERNARDO, J. M. (1989). Ordered group reference priors with applications to multinomial and variance components problems. Technical Report 01/89, Dept. Estadística, Presidencia de la Generalidad Valenciana, Spain.

BERNARDO, J. M. (1979). Reference posterior distributions for Bayesian inference (with discussion). *J. Roy. Statist. Soc. Ser. B* **41** 113–147.

DAWID, A. P., STONE, M. and ZIDEK, J. V. (1973). Marginalization paradox in Bayesian and structural inference (with discussion). *J. Roy. Statist. Soc. Ser. B* **35** 189–223.

STEIN, C. M. (1959). An example of wide discrepancy between fiducial and confidence intervals. *Ann. Math. Statist.* **30** 877–880.

Comment

C. R. Rao

Geometric ideas do help in suggesting intuitive solutions to some complex problems and also in obtaining explicit solutions to specific problems through geometric methods. In his paper, "The geometry of asymptotic inference," Dr. Kass has demonstrated these two aspects by providing us with an excellent review of the past work and presenting some new ideas on the use of differential geometry in interpreting and developing statistical methodology. As Dr. Kass observed, differential geometry is a branch of mathematics "which is largely unfamiliar to most statisticians and may seem rather technical." I hope his paper will create some interest and encourage research in the differential geometric approach to statistical problems. However, I am tempted to share the caution expressed by Dr. D. J. Finney, in a similar situation, referring to some recent papers in multivariate analysis: "Amongst the many papers on statistical science published today, some appear to find outlets to mathematical theory without materially assisting scientific research." One may not fully subscribe to Dr. Finney's view, but the message is clear that enrichment of statistical methodology can take place only if its development is motivated by practical problems that are formulated in statistical terms. In this process, sophisticated mathematics could be used. I hope and believe as Dr. Kass does, that although "no claim can be made as yet that differential geometric research has made inroads into a large class of problems that is otherwise unreachable, the methods are so powerful, and the connections with statistics so plausible, that some further developments, of great methodological importance, might well occur."

In introducing differential geometric methods in statistics, I was motivated by the problem of discrimination between "populations" or "probability distributions " (p.d.'s), which naturally led to the need to introduce a metric in the space of p.d.'s. With a distance defined between two p.d.'s, it is possible to study the configuration of a given set of p.d.'s in terms of clusters and their hierarchical relationships.

In the case of a parametric family of p.d.'s characterized by a set of densities $\{f(x, \boldsymbol{\theta}): \boldsymbol{\theta} \in \Theta\}$, the metric was introduced by furnishing the parameter space Θ with a Riemannian quadratic differential metric (QDM)

$$(1) \qquad \sum g_{ij} d\theta_i d\theta_j$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots)'$, and (g_{ij}) is the Fisher information matrix (see Rao, 1945).

Using the QDM, one can compute the geodesic distance between any two p.d.'s represented by any two parameters θ and ϕ, which we denote by $D_g(\theta, \phi)$.

C. R. Rao is Eberly Professor of Statistics at Pennsylvania State University and Adjunct Professor at the University of Pittsburgh. His mailing address is: Department of Statistics, Pond Laboratory, Pennsylvania State University, University Park, Pennsylvania 16802.

Three immediate uses of such a distance were explored.

(1) Given a sample from a population (p.d.), how to test the hypothesis that the associated parameter θ has a given value θ_0? A possible test criterion for this purpose is $D_g(\hat{\theta}, \theta_0)$ where $\hat{\theta}$ is an estimate of θ based on the sample. If we have two samples from two populations, then the hypothesis that the associated parameters θ and ϕ are equal can be tested by using the test criterion $D_g(\hat{\theta}, \hat{\phi})$ where $\hat{\theta}$ and $\hat{\phi}$ are estimated from the samples. The large sample distribution of these criteria were briefly discussed in Rao (1945), although the power functions associated with such tests were not investigated.

(2) Given a set of populations with parameter values $\theta_1, \theta_2, \cdots$, the set of estimated distances based on samples from each population

$$\{D_g(\hat{\theta}_i, \hat{\theta}_j),\ i, j = 1, 2, \cdots\}$$

could be used in cluster analysis to study the relationships between populations.

(3) Gradual evolutionary changes (in the parameter θ) of a population over time can be represented by a path in the space of p.d.'s. A possible path between two given states (θ and ϕ) of a population at two different points of time is the geodesic between θ and ϕ. Such paths may be of interest in evolutionary studies of biological populations. (See Rao, 1987, for further comments on the choice of a metric).

(4) Suppose that we have a sample X from a population. The p.d.'s based on X provide a geometry with the QDM

$$\Sigma\Sigma g_{ij}^X d\theta_i d\theta_j \tag{2}$$

where (g_{ij}^X) is the Fisher information matrix based on the whole sample X. The corresponding geometry based on a statistic $T = T(X)$ has the QDM

$$\Sigma\Sigma g_{ij}^T d\theta_i d\theta_j \tag{3}$$

where (g_{ij}^T) is the Fisher information matrix based on the statistic T. How do we compare the informative geometries derived from (2) and (3)?

The problem of comparing the geometries based on X and $T = T(X)$ led me to study Fisher's work on comparing the information matrices

$$(g_{ij}^X) \quad \text{and} \quad (g_{ij}^T) \tag{4}$$

and develop the concepts of first- and second-order efficiencies of an estimator.

Fisher thought of an estimator T as providing a summary of data. To what extent can we replace the sample X by T? I have shown elsewhere (Rao, 1948), that in the context of statistical inference in large samples, the score function $\dot{l}(X, \theta)$, where $l(X, \theta)$ is the log-likelihood based on X, plays a fundamental role as a pivotal quantity. A comparison between X and T could then be made by studying the difference

$$D_n = \dot{l}(X, \theta) - \dot{l}(T, \theta) \tag{5}$$

where $l(T, \theta)$ is the loglikelihood based on T. The first-order efficiency refers to the property

$$n^{-1/2}D_n \to 0 \quad \text{in probability} \tag{6}$$

and the second-order efficiency to the quantity

$$\lim_{n\to\infty} \operatorname{cov}(D_n) \tag{7}$$

or the asymptotic covariance matric of D_n. It is seen that (see Rao, 1973, page 330)

$$\operatorname{cov}(D_n) = i^X(\theta) - i^T(\theta) \tag{8}$$

where $i^X(\theta)$ and $i^T(\theta)$ are Fisher information matrices based on X and T, respectively. Since the expression D_n in (5) together with the properties (6) and (7) are difficult to study, I have given alternative formulations of the first-order efficiency as the property

$$|\, n^{-1/2}\dot{l}(X, \theta) - \alpha - n^{1/2}(T_* - \theta)\,| \to 0 \quad \text{in probability} \tag{9}$$

with a suitable T_* as a function T, which is consistent for θ, and the second-order efficiency as the quantity defined as the minimum asymptotic covariance matrix of

$$\dot{l}(X, \theta) - n^{1/2}\alpha - n[\beta(T_* - \theta) + \lambda(T_* - \theta)^2] \tag{10}$$

when minimized with respect to λ. In the problems I have examined, both the definitions {(6), (7)} and {(9), (10)} appeared equivalent, although it is not true in general. It would therefore be of interest to work out the conditions under which these definitions give the same results.

Assuming the possible uses in statistical inference of geodesic paths, geodesic distances and other induced characteristics of the geometry in the space of p.d.'s such as the curvature and α-connections introduced by Amari (1985), several questions arise.

(1) What is an appropriate choice of the QDM in a given statistical problem? Amari (1985) and Kass (in the present paper) stressed the property of invariance of the QDM based on the Fisher information matrix under transformations of the parameters as well as variables. There are other choices of the QDM such as those based on a quadratic entropy (see Section 3 in Rao, 1987), which have the same property. Can we lay down some criteria for the choice of a QDM?

(2) Mathematically speaking, the parameters defining the affine connections can be arbitrary, and it is not clear how they could be chosen in a given statistical problem. Dr. Kass has not given an adequate discussion of this aspect of the geometry. Some further comments will be useful.

(3) It would be of interest to obtain an expansion of the form

$$i^X(\theta) - i^T(\theta) = \alpha + \frac{\beta}{n} + \cdots .$$

We know the expression for α and its geometric interpretation. What about β?

(4) I believe that the choice of a prior distribution is governed by the nature of the parameter and previous knowledge (though vague) about it and should not depend on what experiment is conducted to have further information on it. Jeffreys' invariant prior may have nice properties but it seems to depend on how observations are generated, which may not be acceptable to Bayesians.

ADDITIONAL REFERENCES

RAO, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proc. Cambridge Philos. Soc.* **44** 50–57.

RAO, C. R. (1973). *Linear Statistical Inference and Its Applications*, 2nd ed. Wiley, New York.

Comment

N. Reid and D. A. S. Fraser

We congratulate Professor Kass on a very clear and interesting account of the role of differential geometry in asymptotic inference. In particular, his discussion of information loss and recovery through conditioning, and the geometric interpretation of this, adds substantially to the long-standing discussion initiated in Fisher's early work.

The use and implications of conditional analysis are central to the topics in the paper. In this discussion, we expand a little on arguments for and justifications of conditioning, and the use of geometric methods to motivate this.

In the setting discussed in Section 3.1, we can write

$$p_Y(y \mid \theta) = p_{T|A}(t \mid a, \theta)p_A(a) \tag{1}$$

where $Y = (T, A)$ is sufficient, A is ancillary, and the Jacobian has been absorbed into the support differentials. This factorization suggests, as the paper indicates, that inference about θ may be based on the conditional distribution of T given A, without loss of information about θ. Section 3.1.1 gives formal clarity to Fisher's general analysis of information loss and is valuable in giving a precise interpretation of the phrase "without loss of information about θ."

Other arguments can also provide some interpretation of the phrase above. For example the likelihood function obtained from the conditional distribution is the same as the likelihood function from the distribution of the full data Y. Another motivation for conditioning on A when the factorization in (1) holds is that the variability in the outcome that is described by the marginal distribution of A is irrelevant for inference about θ; this is an underlying theme in Fisher's early work expanded in Fisher (1961) and is very clearly presented in the weighing machine example of Cox (1958). Fisher frequently used the term "relevant subset" to refer to the set of sample points having the observed value for the ancillary statistic. However, it seems clear that he attached additional meaning to the term, derived from the physical context from which the statistical problem arose. Indeed, this additional interpretation may well have been primary in Fisher's interpretation of conditioning and the definition of the correct probabilities to use in applications. There does seem to be no fully satisfactory formalization of such "relevant subsets" based on the statistical model alone. The derivation of the Likelihood Principle from the Conditionality Principle discussed in Evans, Fraser and Monette (1986) bears on this.

Most discussions of conditioning are motivated by a few very compelling examples. Subsequent attempts to formalize the operating principle to enable extension to more realistic settings are widely divergent. One development, primarily initiated by Birnbaum (1962, 1972) and Basu (1959, 1964) (see also Buehler, 1982), isolates ancillarity as the essential feature; the discussion of this approach and its relation to Bayesian inference and the likelihood principle is well summarized in Berger and Wolpert (1985).

Another development of conditioning in Fraser (1968, 1979) extends and formalizes one aspect of

N. Reid is Professor in the Department of Statistics at the University of Toronto. Her mailing address is: Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S 1A1. D. A. S. Fraser is Professor in the Department of Mathematics at York University. His mailing address is: Department of Mathematics, York University, Downsview, Ontario, Canada M3J 1P3.