KJELL A. DOKSUM[1] AND ALBERT Y. LO[1]

University of California, Berkeley, and Rutgers University and SUNY at Buffalo

1. Introduction. Diaconis and Freedman (D & F) have made important and interesting contributions to the problem of determining in which situations nonparametric Bayes estimates do and do not work. Investigating when statistical principles and techniques break down is an important enterprise which is not well enough appreciated. Thus, even though we often experience in life that nature serves up distributions consistent with Murphy's Law,[2] in applied research, there is a tendency to believe that nature provides nice simplistic distributions and models. Economists are starting to realize that this belief can lead to large errors in prediction. In other fields, it may take longer to discover similar problems.

We find it both surprising and interesting that inconsistency can occur when the prior on the location parameter, the Dirichlet parameter in the law of the error distribution, and the distribution sampled, are all "nice" and symmetric about zero. D & F (1986a (hereafter I), Section 3) suggest using the "device of imaginary results" or the "what if" method to deal with the inconsistency. This procedure involves modifying the prior after computing the posterior for "imaginary data sequences." In Section 2 below, we discuss the properties of a different and much simpler (subjectively speaking) approach which amounts to computing a posterior distribution based on partial information or to presmoothing before computing the posterior. In Section 3 we show that this "partial posterior" idea can be linked to partial likelihood.

But first we will focus on the following intriguing D & F statement:

> Any of the classical estimators, such as the mean or the median will be consistent in this situation, so the Bayes estimates do worse than available frequentist procedures. (D & F I, Remark 4, Section 1. See also D & F (1986b; hereafter II), Section 1.)

This statement refers to models where the "Bayes" procedure is given the job of coping with the infinitely dimensional nuisance parameter F as well as location while the "frequentist" procedure essentially only has to deal with location since any nuisance parameter difficulties have been removed by assuming symmetry. Thus we think that a fairer comparison would be the nonparametric Bayes procedure versus the semiparametric frequentists procedure where the pair (θ, F) is estimated using semiparametric maximum likelihood techniques.

Rather than pursuing this last remark, we claim that the D & F results lead to the conclusion that what is needed in the nonparametric framework are Bayes procedures for location that are not distracted by the problem of dealing with an infinitely dimensional nuisance parameter. Thus we propose using the posterior

[1] Research partially supported by National Science Foundation Grants DMS83-01716 and MCS81-02523-01.

[2] When something can go wrong, it will.

distribution of θ given the trimmed sample mean as the basis for Bayes procedures for location. More generally, we can condition on any estimate T of location. In a sense, this corresponds to focusing on location and smoothing to get rid of distributional unpleasantness that distracts from the location problem before computing the posterior. For instance, for the counterexample density h (D & F I, Figure 1), $\bar{X}$ is nearly normal already for $n = 12$.

In the next section we discuss the consistency, asymptotic normality, and efficiency of such procedures and find that they have high posterior efficiency. Moreover, these procedures can be justified on pragmatic grounds. To paraphrase C. Eisenhart (see Tukey, 1954) "the practical efficiency is the product of the statistical efficiency of the technique and the probability that the technique will ever be used." Our proposed procedure is very simple: $\mathscr{L}(T|\theta)$ is approximately normal; thus with a normal prior, $\mathscr{L}(\theta|T)$ will approximately be the usual normal theory posterior. Moreover, $\mathscr{L}(\theta|T)$ will inherit the robustness properties of T.

2. Robust and consistent Bayes procedures. Let $X_i = \theta + \varepsilon_i$, $i = 1, \dots, n$, where $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. with fixed distribution function F. Thus given θ, X_i has distribution $F_\theta(x) = F(x - \theta)$. Let θ have bounded and continuous prior density $\pi(\theta)$. We suppose that T is a translation invariant estimate of θ, i.e., $T(x + c) = T(x) + c$, $c \in (-\infty, \infty)$, and that $T \to 0$ a.s. $[F]$. Moreover, we assume that $\mathscr{L}(\sqrt{n}\,T|0) \to \mathscr{N}(0, \sigma^2(F))$.

THEOREM 1. *Under the above conditions, the posterior distribution of θ given the partial information provided by T converges to the $\mathscr{N}(T, n^{-1}\sigma^2(F))$ distribution. More precisely, if $\pi(\theta_0) > 0$, then $\mathscr{L}(\sqrt{n}(\theta - T)|T) \to \mathscr{N}(0, \sigma^2(F))$ a.s. $[F_{\theta_0}]$.*

The proof will appear in a forthcoming paper by the authors.

Note that F is not assumed to be symmetric. If T is the sample median, the conditions will be satisfied if F has median zero.

For two posterior distributions $\mathscr{L}(\theta|T_1)$ and $\mathscr{L}(\theta|T_2)$, it is natural to define the Bayes asymptotic relative efficiency (BARE) as the ratio $e_F(T_1, T_2) = \sigma_2^2(F)/\sigma_1^2(F)$ of the posterior variances in Theorem 1. For instance, if T_1 is the normal scores estimate (Hodges and Lehmann, 1963) and $T_2 = \bar{X}$, then $e_F(T_1, T_2) \geq 1$ for all F symmetric about zero. Thus, in terms of BARE, the normal scores Bayes procedure is uniformly more efficient than the familiar Bayes procedure based on $\mathscr{L}(\theta|\bar{X})$. Similarly, let T_3 be an adaptive estimate of θ when F is symmetric (Stein, 1956; Stone, 1975). Then modulo proving strong consistency of T_3, $\mathscr{L}(\theta|T_3)$ is an adaptive Bayes posterior with asymptotic posterior variance $\sigma_3^2(F) = 1/nI(F)$, where $I(F)$ is the Fisher information. Thus, from the point of view of asymptotic Bayes theory for the location of a symmetric distribution, conditioning on an adaptive estimate may make more sense than putting a prior on F.

REMARK 2.1. For the above model, Lo (1984) has shown that $E(\theta|X) \to \theta_0$ a.s. $[F_{\theta_0}]$ and $\mathscr{L}(\theta|X) \to \delta_{\theta_0}$ a.s. $[F_{\theta_0}]$ in the undominated case, where δ_{θ_0} is point mass at θ_0. The results of D & F show that these results fail to hold if F_{θ_0} is replaced by their counterexample distribution H (D & F I, Figure 1).

Next we turn to the model where F is also random with a Dirichlet distribution $\mathscr{D}(\alpha)$. We want to illustrate that by using a posterior based on partial information, we can get consistency. To this end, let T denote the sample median. Since $\alpha/\alpha(R)$ and the counterexample distribution H (D & F I, Figure 1) both have median zero, then $T \to 0$ a.s. $[\alpha/\alpha(R)]$ and $T \to 0$ a.s. $[H]$. Let $\alpha_\theta(t) = \alpha((-\infty, t - \theta])/\alpha(R)$ and $H_\theta(t) = H(t - \theta)$, then $T \to \theta_0$ a.s. $[\alpha_{\theta_0}]$ and $T \to \theta_0$ a.s. $[H_{\theta_0}]$. These results can be used to show:

THEOREM 2. *With the above conventions, if* $\pi(\theta_0) > 0$, *then* $\mathscr{L}(\theta|T) \to \delta_{\theta_0}$ *a.s.* $[H_{\theta_0}]$. *If* $\theta\pi(\theta)$ *is bounded, then* $E(\theta|T) \to \theta_0$ *a.s.* $[H_{\theta_0}]$.

The proof will appear in a forthcoming paper by the authors.

We think that the results of this section in conjunction with those of D & F give compelling reasons for flexible Bayesians who desire robust procedures to consider posterior distributions based on partial information. It is a matter of giving a little (give up Bayes efficiency for some idealized model) in order to gain a lot (high Bayes efficiency and consistency over a wide class of models including the D & F counterexample distribution). How much is gained (and lost) should be the focus of research of pragmatic Bayesians everywhere.

REMARK 2.2. In the D & F spirit of "true confessions" (D & F I, Section 3), what are we anyway? The answer is: Statisticians! Beyond that, one of us is a Bayesian, but he will no doubt be excommunicated as a heretic after the above remarks.

3. The history of partial information. Cox (1972, 1975) introduced the idea of partial likelihood. Kalbfleisch and Prentice (1973) showed that in many interesting cases, the partial likelihood coincides with the marginal likelihood and the rank likelihood. The latter is the probability distribution of the rank vector considered as a function of the parameters and had been used earlier by Hoeffding (1950) to generate optimal rank procedures. Savage and Saxena (Savage, 1969) proposed using the posterior distribution given the ranks in a nonparametric Bayesian context. This is also a form of smoothing before computing the posterior: The ranks map the data set $-150, 6, -2, 3$ into $-4, 3, -1, 2$. In fact, for a sample from the D & F counterexample distribution H, the signed ranks have a uniform distribution over the space of possible signed ranks (e.g., Bickel and Doksum, 1977, page 360). Recently, Pettitt (1983) has proposed useful approximations to the posterior given the ranks.

REFERENCES

BICKEL, P. J. and DOKSUM K. A. (1977). *Mathematical Statistics: Basic Ideas and Selected Topics.* Holden-Day, San Francisco.

COX, D. R. (1972). Regression models and life tables. *J. Roy. Statist. Soc. B* **34** 187–202.

COX, D. R. (1975). Partial likelihood. *Biometrika* **62** 269–276.
DIACONIS, P. and FREEDMAN, D. (1986a). On the consistency of Bayes estimates. *Ann. Statist.* **14** 1–26.
DIACONIS, P. and FREEDMAN, D. (1986b). On inconsistent Bayes estimates of location. *Ann. Statist.* **14** 68–87.
HODGES, J. L., JR. and LEHMANN, E. L. (1963). Estimates of location based on ranks. *Ann. Math. Statist.* **34** 598–611.
HOEFFDING, W. (1950). "Optimum" nonparametric tests. *Proc. Second Berkeley Symp.* 83–92. Univ. California Press.
KALBFLEISCH, J. D. and PRENTICE, R. L. (1973). Marginal likelihood based on Cox's regression and life model. *Biometrika* **60** 267–278.
LO, A. Y. (1984). Consistency in the location model: The undominated case. *Ann. Statist.* **12** 1584–1587.
PETTITT, A. N. (1983). Likelihood based inference using signed ranks for matched pairs. *J. Roy. Statist. Soc. B* **45** 287–296.
SAVAGE, I. R. (1969). Nonparametric statistics: A personal review. *Sankhyā* **31** 107–144.
STEIN, C. (1956). Efficient nonparametric testing and estimation. *Proc. Third Berkeley Symp.* **1** 187–196.
STONE, C. J. (1975). Adaptive maximum likelihood estimators of a location parameter. *Ann. Statist.* **3** 267–284.
TUKEY, J. W. (1954). Unsolved problems of experimental statistics. *J. Amer. Statist. Assoc.* **49** 706–731.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720

DEPARTMENT OF STATISTICS
SUNY AT BUFFALO
4230 RIDGE LEA ROAD
BUFFALO, NEW YORK 14226

HANI DOSS

Florida State University

It is generally acknowledged that it is hard to think about priors in high (but finite) dimensional spaces. Subjective Bayesians know that it is hard to elicit a prior from an individual when the dimension is 3 or 4. Diaconis and Freedman's interesting results on an inconsistent Bayes rule involving a reasonably natural prior show how far off our intuition can be when we pass to an infinite dimensional setting. In this discussion, we present other peculiarities, in addition to the inconsistent behavior, that arise when one uses the symmetrized Dirichlet prior. The discussion concludes with a few remarks on an alternative way of constructing priors on c.d.f.'s.

1. The symmetrized Dirichlet priors. The setup considered by Diaconis and Freedman is the following:

$$X_i = \theta + \varepsilon_i, \qquad i = 1, 2, \dots, n, \qquad \varepsilon_i \text{ are i.i.d.} \sim F.$$

The parameters θ and F are independent, θ having a density f, and F being distributed according to $\bar{\mathscr{D}}_\alpha$, with α absolutely continuous.

Let $\theta_{ij} = \frac{1}{2}(X_i + X_j)$, and let $\#(\theta_{ij})$ denote the number of distinct pairs (X_k, X_l) such that $\frac{1}{2}(X_k + X_l) = \theta_{ij}$. (The pairs (X_k, X_l) and (X_s, X_r) are called distinct if the sets $\{X_k, X_l\}$ and $\{X_s, X_r\}$ are distinct.) The number $\#(\theta_{ij})$ will be called the multiplicity of θ_{ij}. The posterior distribution of θ given $X_1, \dots, X_n$ is denoted $\bar{\pi}_n$.