DISCUSSION

S.E. FIENBERG (*University of Minnesota*):

The three papers presented at this Session have been linked together under a common title. In fact they are only loosely related to one another. The key links would appear to be (1) between the Makov paper on sequential learning, and Section 6 of Smith's paper on change-point problems, (2) the use in all three papers of the idea of recursive updating (this appears explicitly in the Makov and Smith papers, and only implicitly in the Harrison and Smith paper via the use of smoothed data for the tension and alienation variables in their prison riot example of Section 7). Having noted these links between the three papers, I now turn to a separate discussion of each.

I found Makov's review of Bayesian-like approaches to unsupervised sequential learning problems most interesting. This review is especially welcome since most of the work on this topic has appeared outside the mainstream statistical journals. Clearly the problem is a difficult one, and Makov and the others who have worked on various aspects of it are to be congratulated for the progress they have made.

All three cases considered by Makov assume that the p.d.f.'s for an observation x, given that it comes from class H_i, are of the functional form:

$$f_i(x|\theta_i) = f(x|\theta_i, H_i)$$

i.e., there is a common function form for the p.d.f.'s. Moreover, the number of classes, k, is given. The more general problem of mixture with f_i's of possibly different form, and unknown k, has been discussed quite recently by Good and Gaskins (1980 and the ensuing discussion). The computational complexities of the various approaches to more general "bump-hunting" problems make Makov's restrictions quite reasonable for statistical purposes. I should also mention the graphical methods of Fowlkes (1979) for studying mixtures of normals where k is unkown.

I have three questions related to the procedures for Case A discussed in this paper, which may have relatively brief answers:

(1) Has there been any investigation of the adequacy of approximating a mixture of Dirichlets by a single Dirichlet? Good (1967) has noted an example of the inadequacy of a single Dirichlet when the true prior is a mixture.

(2) Isn't part of the problem with the DD, (MDD,) and *PT* methods in your Figure 1 and in other studies due to choice of a "weak" prior?

(3) It would appear that the *QB* computations at step n are not invariant with respect to the ordering of $x_1, \ldots, x_{n-1}$. Is this the case, and if so is it something that a good Bayesian striving for coherence should worry about?

Finally, I note that all of the methods in this paper assumne that observations arrive sequentially, one at a time. Has there been work on related problems when observations arrive in batches?

Smith's paper provides us with a quick, guided tour of the Bayesian approach to change-point problems. It begins in the land of exchangeable subsequences and a consideration of problems representable in such form, and then proceeds with a series of brief stops to explore the more general problems of changes in regression-like

structures where the exchangeable subsequences approach is not directly applicable. While the tour has been quick, offering little opportunity for dalliance with any one problem, it has covered much territory in a spirited fashion, and may well whet our appetite for return visits to selected locations.

My comments and queries focus primarily on the simplest of the problems Smith describes in connection with binomial data in Section 2, but I suspect related questions can be raised about the other problems discussed in the later sections. Although the method of analysis described in Section 2 for the Lindisfarne Scribes problem seems quite general, I believe further attention needs to be given to various consistency questions such as the following: (1) If a change-point at $r_1 = 5$ has high posterior probability when $K = 1$ is assumed, does it necessarily follow that $r_i = 5$ will be included in the pair of change-points with the highest posterior probability when $K = 2$? (2) Is it possible that when we place positive prior probabilities on $K = 0,1,2,3$ we can get Bayes factors favoring $K = 2$ at say $(r_1, r_2) = (4,5)$, but when we place positive prior probabilities on $K = 0,1,2,3,4$, we get Bayes factors favoring $K = 4$ at $(r_1, r_2, r_3, r_4) = (4,5,6,7)$? Such consistency properties would seem highly desirable, but would seem to depend on the specification of the priors, $p(\theta_1,\theta_2, \ldots, \theta_{K+1})$. Perhaps Professor Smith has already explored some of these matters in detail.

I also have some concerns regarding the beta structure used for the Lindisfarne Scribes problem. Smith notes that the assignment of independent beta priors may well be unreasonable, but then he goes on to use them nonetheless due to the computational simplicity they provide. Although I have no compelling reasons to suggest in their support, two alternatives that may bear further examination are: (a) a Dirichlet for the joint density of $\theta_j/\Sigma_{i=1}^{k}\ \theta_i$, $j = 1,2, \ldots, K$; or (b) variants of generalized Dirichlets. The major advantage to these densities (aside from the dependencies they introduce) is that it can be represented as a product of independent betas for the random variables $\theta_j/\sum_{i\leq j}\ \theta_i$. This property may be helpful in achieving the consistency properties I referred to above.

Having revisited the land of exchangeable subsequences with you, I would encourage you to take Smith's complete guided tour for yourselves and choose your own location for an extended visit and prolonged statistical investigation.

While Professor Harrison's oral presentation of this paper can be viewed as nothing short of a tour de force, after several readings of the written version of the paper I am at a loss in my assessment of its contributions to Bayesian decision making. Some of the mathematical aspects of the paper are very interesting, and the discussion of expected loss functions with multiple minima seems quite novel. But there appears to be a fundamental discontinuity in my appreciation and understanding of the paper, as I go from the mathematical formulations to their application. Let me elaborate.

In the initial results they describe, Harrison and Smith investigate decision problems involving bounded utility functions, and they are interested in the behavior of the expected loss, $E[\delta,u]$, with respect to the belief distribution of future outcomes, ϕ, and where δ is the decision, and u represents environmental variables expressible in terms of the parameters of the distribution of ϕ and the loss function, $L(\delta,\phi,u)$. In Section 3, they illustrate that, if the bounded loss function is a monotonically increasing function of δ-ϕ and the distribution of future outcomes is unimodal, then

$E[\delta,u]$ may have two minima. The keys to this result are two: (1) a parameter, α, that appears only in L, (2) monotonic behavior of the scores function over an interval linked to kinks in L. When α takes on one set of values we get the first Bayes decision, and when it takes a different set of values we get a different optimal decision. Despite the fact that slight variations in α may lead to different decision, we must recognize that different values of α do correspond to *different* loss functions, even though they have the same general shape. If one of the two Bayes decisions has smaller expected loss, the decision maker, who after all determines his own loss function, might do well to alter his value of α to achieve the minimum. If, as in the case of the exponential distribution there is a value of α leading to an interval of Bayes decisions with the same expected loss, it does not matter which δ the decision maker chooses since the expected loss does not change. Still, this basic result of Harrison and Smith is somewhat disquieting and bears further examination.

Once we move to multimodal densities for beliefs and for multimodal loss functions, it is not quite so surprising that the possibility of multiple Bayes decisions exists. The main examples Harrison and Smith use to illustrate such situations involve belief distributions which are mixtures of normals and loss functions which are mixtures of Lindley's conjugate normal loss functions. The simplest example here is what the authors refer to as the $E_1^*[\delta]$ model, and involves just the normal distribution with its conjugate loss function, where one of the parameters of the loss function, k, increases with δ. That this situation leads to bifurcating Bayes behavior simply heightens the latent suspicion I have long harboured regarding the appropriateness of Lindley's conjugate loss functions. Yet this result, like the earlier one, is somewhat disquieting.

Up to this point my comments have been technical ones, and have focussed on the mathematical developments described in the paper. The catastrophic discontinuity comes when we turn to the "applications" of this theory.

The first major application of the theory is via an E_2^* model for prison riots. I have been involved in a study of prison-related rehabilitation activities in the United States, have visited with various corrections officials, and actually have spent a little time in a major corrections facility. Thus, I was especially keen to see how a catastrophy-theory like Bayes decision model could be use in "illuminating the dynamics of prison riots".

Now prisons are complex institutions, and to think that an accurate portrayal of behavior in prisons can be made by looking at three crudely defined and artificially interpreted variables seems naive at best. The model *assumed* by Harrison and Smith involves a single normal belief distribution and a bimodal mixture of two conjugate loss functions. Why did they pick such a model? We are told in such loose, heuristic, and ambiguous terms that even the statistically uneducated reader might scream: Stop! What I find even more distressing in the material presented is that we are shown no attempts at model criticism or parameter estimation, the key features of statistical inference when models are used as part of the scientific method (see the related discussion of the role of models in the paper by George Box, given at this conference). You may think from a reading of Section 7.2 that aspects of the requisite data analysis (Bayesian or otherwise) are contained in the referenced papers of Zeeman, Hall, Harrison, Marriage, and Shapland (1976, 1977) but this simply is not so. Although

these papers do contain a more detailed description of the data plotted in Figure 7.2, the motivation and justification for the model and the assessment of the adequacy of the model's fit to the data are attended to in a manner just as facile as in the present paper. Even with such a loose approach as the authors choose to present, the model shifts in midstream as a result of "a higher tolerance of tension in the institution after the first mass protest"!

Next we come to the interpretation of the E_2^* model in prison disturbance context. I would claim that all anyone can get out of the model "applied to the data" is what the authors have put into it to begin with. The catastrophes in the prison behavior they "account for" are really only a restatement of the fact that the model contains discontinuities (see the related critique of applied catastrophe theory in the behavioral sciences by Sussman and Zahler, 1978a, 1978b). The Bayesian Bloodhound referred to in Section 7.3, after reading such a description of statistical modelling, would clearly accept this torment no longer, and would visciously attack the authors until they completed a more satisfactory job of analysis and modelling. All of this is not to say that the E_2^* model is inappropriate for the prison riot example (although I have my suspicions). Rather I believe that the authors have not presented very effective evidence in support of their claims.

Finally, to illustrate my concerns with the other major application described in Section 8, I will give my own "application", henceforth to be known as: "Bayesians, Luggage, and the Butterfly Catastrophe". The decision maker involved is a recently-married Bayesian statistician who upon arrival at the Valencia Airport with his wife en route to this Meeting discovers that their luggage has not arrived with them. The loss function involved is much like the one in Section 8.2. There is (1) "pressure from above" by his wife to stay at the airport until the luggage arrives, (2) "pressure from below" by all the other participants who are waiting for the statistician and his wife aboard the bus that is to take them to the Meeting, (3) the statistician's own criterion and beliefs which fall somewhere in the middle. Since this structure is essentially the same as in Section 8.2.1 it should be clear to all readers that we are faced with an example of a butterfly catastrophe. The two conflicting extreme Bayesian decisions can be translated into (1) staying overnight at a hotel near the airport waiting for the luggage to arrive, and (2) immediate departure on the bus. The "middle course" decision is a little too complex to describe here (it involves a non-exponential waiting time distribution with a heavy tail), but we hope to publish a detailed description at a later time. Needless to say, the butterfly cross-section was very helpful in resolving the conflict in this particular problem, as I will indicate quite shortly.

Does the mathematical phenomenon of a butterfly catastrophe follow from my assumptions in a nontrivial way? Indeed, does it really follow at all? Or is this implementation of the E_3^* model simply the consequence of a vague specification, and some hand-waving (perhaps I should say "wing-flapping")? For this example, I readily admit to a contrived "application" of the models and descriptions in the Harrison and Smith paper. I don't believe the resulting butterfly catastrophe tells us anything of practial value at all. Yet I find my own description not all that much different from that of Section 8 of the Harrison and Smith paper. I believe the value of their models in real

applications can only be judged by a more careful statistical treatment than the one we are offered in this paper.

All in all, I found the Harrison and Smith paper both stimulating and highly provocative. I look forward to seeing elaborations of their ideas in the future. Lest it appear that I am finishing my comments on a note of dispair, let me note the "luggage example" described above was factual, and that my use of it in this discussion did have one practical consequence. The beleaguered statistician in question (who will remain anonymous) did in fact decide to board the waiting bus and travel without his luggage to the site of the Meeting in Las Fuentes. My description of his plight in the oral presentation of this discussion inspired me to loan him, along with other apparel, my *t*-shirt with the brilliantly-colored image of the famous Dunk Island (Australia) blue butterfly on its chest. Never let it be said that Bayesian decision theory does not have its useful applications!

J.M. BERNARDO (*Universidad de Valencia*):

The need for approximations in the problem discussed by Mr. Makov is fairly clear to me. However, I would like to know more about the quality of the approximation he proposes. For instance, one could try to estimate the expected distance, in some well specified sense, between the exact Bayes posterior predictive distribution

$$\{\, p(x_n \epsilon H_i | x_1,\ldots,x_n),\ i = 1,\ldots,k\}$$

and its quasi-Bayes approximation. Moreover, since the true source of the x_i's is never known it might happen that wrong allocations are piled up thus making convergence to the correct allocation as new observations occur difficult, or perhaps impossible.

P.J. BROWN (*Imperial College, London*):

I should like to amplify a point raised by Professor Fienberg concerning the adequacy of the Dirichlet distribution. The problem with the Dirichlet is that it has a very straightjacketed variance-covariance structure. Indeed it involves virtual independence apart from correlations resulting from normalisation to unity. Elsewhere Brown (1976), I have documented some unfortunate features of the Dirichlet prior. To see that an unsupervised learning situation may have a rather different variance-covariance structure consider the following example. There are $K = 3$ populations which are $N(\theta_i,1)$, $i = 1,2,3$. Imagine the case where θ_2 and θ_3 are close together and quite distant from θ_1. Then unsupervised learning will quickly and accurately determine π_1 and $\pi_2 + \pi_3$ but π_2 and π_3 will be highly correlated together and will have a low correlation with π_1. In this situation the Dirichlet representation will not be able to reflect these second order properties. Thus although Professor Makov's scheme will result in eventual convergence to π_1, π_2 and π_3 it may be difficult to discern the reliability of one's estimates at any stage. Use of an approximating multivariate normal distribution would get around this problem but would of course involve heavier computation.

A.P. DAWID (*The City University, London*):

I want to stress the need for care in setting up models of the kind that Makov has been working with. I am sure these are appropriate for the engineering applications with which he is concerned, but I am none too happy when I see them used in other fields. In particular, I am extremely doubtful about their general suitability in the setting of medical diagnosis, as in the work of Titterington and others.

In this context, the classes $\{H_i\}$ represent diseases, and the observation x corresponds to medical symptoms: thus π may be thought of as describing the prevalence of disease, and θ the "clinical pictures" of the diseases. My unease stems from the seeming possibility, when using a mixture model as described, of obtaining consistent estimates of the parameters. This mean that, if we collect a large enough data-base of patients and record *only* their symptoms (never discovering what diseases they are suffering from), we can nevertheless gain accurate knowledge of both disease prevalences and clinical pictures. My possibly naive reaction to this remarkable state of affairs is one of distrust: how can one learn about anything other than the marginal distribution of symptoms from data such as these? If our model says that we can, it may be a signal that we should discard the model.

It is easy to set up alternative models which behave more reasonably. Instead of splitting up the joint distribution of disease D and symptoms S as $f(d,s|\psi) = f(d|\pi)f(s|d,\theta)$, as is normally done, decompose it instead as $f(d,s|\psi) = f(s|\alpha)f(d|s,\beta)$. (There are good practical reasons for regarding this as more meaningful in a diagnostic context: see Dawid, 1976). We are just as much at liberty to make assumptions about the new parameters (α,β) as about (π,θ). In particular, it does not seem unreasonable to me to assume that α and β are, *a priori*, independent. If so, it is easy to show that observation of patients' symptoms alone will modify the distribution of α, but leave that of β unchanged. Now for purposes of diagnosis (prediction of D from S) only β is relevant: consequently (and, I consider, quite reasonably) such a data-bank is entirely valueless.

A simple example may make my point. Suppose D takes only two values, d_1 and d_2, and likewise S takes values s_1 or s_2. Let $\pi_i = P(D = d_i)$, $\theta_{ij} = P(S=s_j|D=d_i)$, and suppose that π and θ are *a priori* independent, with $\pi_1 \sim \beta(a_1.,a_2.)$, $\theta_{11} \sim \beta(a_{11},a_{12})$, $\theta_{21} \sim \beta(a_{21},a_{22})$ all independently (where $a_{i.} = a_{i1} + a_{i2}$). This is an example of Case C of Makov's classification, although it does not satisfy the condition that mixtures should be identifiable. In fact, for this model, a data-bank of unconfirmed cases is quite useless for diagnosis. For, letting $\alpha_j = P(S=s_j)(= \pi_1\theta_{1j} + \pi_2\theta_{2j})$, $\beta_{ij} = P(D=d_i|S=s_j)$ $(=\theta_{ij}\pi_i/\alpha_j)$, we have $P(d,s|\psi) = P(s|\alpha)\,P(d|s,\beta)$, where it is easily found that α and β are *a priori* independent, with, in fact, $\alpha_1 \sim \beta(a_{.1},a_{.2})$, $\beta_{11} \sim \beta(a_{11},a_{21})$ $\beta_{12} \sim \beta(a_{12},a_{22})$, all independently. So the considerations of the last paragraph apply.

The moral of all this is that, when we choose a particular mathematical model to represent a real-world process, and make seemingly harmless assumptions (such as "identifiability of mixtures"), we must be careful that any deductions we make are "*qualitatively stable*" in the sense that similar conclusions would be derived from other reasonable ways of modelling the process. (The term 'reasonable" here depends, of course, on the particular application).

Much as I appreciate Makov's contributions to signal detection, I fear that his

models may come to be used all too uncritically in other applications, for which they fail to be qualitatively stable.

J.M. DICKEY (*University College of Wales*):

A comment may be of some interest here relative to Professor Smith's paper and other papers in this conference in which an inference is made between nested sampling models. The Bayes factor, or ratio of posterior odds to prior odds in favour of one sampling model versus another, depends on each prior distribution conditional on a model. But what prior distributions should one use and how should they be related? The obvious answer is, of course, that pair of tractable distributions which most closely models actual prior uncertainty conditional on each of the models. Savage's *condition continuity* offers a reasonable guide to such a choice (Dickey and Lientz 1970, Gunel and Dickey 1974). This requires that the prior distribution within the smaller nested model be identical to the conditional distribution induced in the usual way from the joint prior distribution in the larger model. (In this case, the Bayes factor will equal Savage's density ratio, the ratio of posterior to prior densities of the conditioning constraint parameter at the null value).

To see that Professor Smith's choices do not satisfy condition continuity, consider the joint density of the regression coefficient θ and the variance σ^2. The pair (θ,σ^2) are *dependent* under the larger model; the smaller model is obtained by a constraint on θ; hence one will not satisfy condition continuity by having the prior distributions of σ^2 identical under the two models. Professor Smith takes them identical. On the other hand, my own papers on Bayes factors for the normal linear model use condition continuity (Dickey 1971, eq. (5.40), 1974 Prop. 4.2).

Note that by the Borel-Kolmogorov nonuniqueness mentioned in my discussion to Professor Hill's paper in these Proceedings, the answer in each case to the question of whether condition continuity is satisfied will depend on the choice of conditioning constraint variable in terms of which the question is framed. If it is not satisfied for a given pair of distributions for one choice of conditioning variable, perhaps it will be for another choice of conditioning variable. (The same smaller model can often be defined by various essentially different constraints in the larger model). In fact, we shall see this happen for the generalizations of Jeffreys' Bayes factors to be presented later in these Proceedings by Professor Zellner. If the new parameter is used, $\eta = \sigma^{-1}\theta$, and if η is independent of σ, the condition on η will not have an effect on the distribution of σ.

A theorem can even be stated showing that an *arbitrary* given distribution in the smaller model can be obtained by condition continuity from the larger model by suitable choice of conditioning variable. This then would seem to make condition continuity a mathematically vacuous requirement. However, I should like to point out that in practice there are often *natural* conditioning variables η, that is, variables for which one would like to define the smaller model as the consequence of additional information of the form, "η lies in some small hyperinterval centred at the point η_o". This is often the case when the overall mixed type distribution, having positive prior probability attached to the smaller model, is intended as an approximation to a continuous density on the full parameter space with a high mound or ridge over a neighborhood surrounding the constraint set.

In practice, one must be careful to model real uncertainty, and not let the mathematics do one's thinking for one. Professor Smith very wisely took his prior parameters in the binomial Lindisfarne scribe problem so that the uncertainty concerning a single scribe alone was the same as the uncertainty concerning the first scribe among many scribes, instead of the same as if he had been told the many were one. If he had used condition continuity based on any linear conditioning variable, such as $\eta = (\theta_2 - \theta_1, \ldots, \theta_{K+1} - \theta_1)$, then the conditional distribution of θ_1 given $\eta = 0$ would have been beta with parameters $\alpha_1 + \ldots + \alpha_{K+1} - K$ and $\beta_1 + \ldots + \beta_{K+1} - K$. For $\alpha_i < 1$ and $\beta_i < 1$, i = 1, ..., $K+1$, and K large, this distribution of θ_1 would have had a small variance, instead of the same variance as θ_1 under the larger model. That is, if one were told that only one scribe was involved, one's opinion would have been less vague than one's opinion concerning the first scribe of many. Of course, if $\alpha_i < 1$ and $\beta_i < 1$ for all i, then the conditional opinion would have been more, rather than less vague than the unconditional opinion. In fact, for small enough (positive) α_i and β_i, the conditional distribution would have been degenerate, even though the joint distribution was proper. (I am greteful to Professor P.R. Freeman and Professor A.P. Dawid for personal discussions on this example).

J.B. KADANE (*Carnegie-Mellon University*):

The assignment of prior distributions under the different models is a very sensitive matter for model selection. There is no particular justification for the independent beta prior of equation (6), for example. Allowing dependent priors can change the answers in the direction of more scribes, and in fact, can in principle suggest up to 13 scribes. Thus Smith's conclusion that there were probably 3 scribes is heavily dependent on the "perhaps unsatisfactory" assumption (6). The same type of comment applies to the ARMA regression and examples. I welcome, however, the interesting applied problems in this paper, especially the Kidney transplant data.

A more decision-theoretic, and I believe more satisfying approach is developed by Lindley (1968) and continued by Kadane and Dickey (1980).

T. LEONARD (*University of Warwick*):

I think that Professor Smith's approach to change-point inference is very useful and interesting, but I wonder whether he might be over specialising his model simply to facilitate a particular type of conclusion? It would seem to be more natural to assume a Kalman-type model permittting different process levels at each time-stage (e.g. the Harrison-Stevens steady model). A whole range of posterior conclusions could then be reached to suit the practical situation at hand. In particular, we could find the restricted Bayes estimates, for the process levels, amongst a suitable class of step functions, thus providing a very simple way of detecting change-points. Further restricted Bayes procedures would cope with the more complicated situations discussed by Professor Smith. This seems to me to provide a conceptually and technically simple way of coping with change-points, and avoids choosing an unduly complex model simply to cope with a single very special type of posterior conclusion. These aspects have been discussed by Leonard (1978).

REPLY TO THE DISCUSSION

U.E. MAKOV (*Chelsea College, London*):

I wish to thank our discussants for their interesting questions and comments and make the following points:

1.- The choice of a Dirichlet prior in *case A* was made simply to exploit the conjugacy property. As described in the paper, neither the 'unfortunate properties of the Dirichlet prior' nor the possible inadequacy of approximating a mixture of Dirichlets by a single Dirichlet affect the desirable assymptotic properties of the QB procedure. However, these inadequacies are bound to affect the small sample properties of the procedure and I suspect that in acute situations, like the one suggested by Dr. Brown, the QB might be painfully slow.

One possible remedy, which was tried numerically,. is to approximate a mixture of size n by a mixture of size k, $k < n$. In Makov (1978) the mixture of Dirichlets is allowed to grow so long as is computationally possible (rather than collapsing the mixture after *each* observation). Thereafter, the mixture build-up can be restarted. In Makov (1978), the mixture, once collapsed, was replaced by a single Dirichlet and no further growth was allowed. In Smith and Makov (1980), in the context of *case B*, the approximating mixture consists of Gaussian p.d.f's. Here, in the case of detection and estimation of jumps in linear systems, the combined quality of the detector/estimator is considerably improved when the number of terms in the approximating density is increased.

2.- The QB procedure is not invariant to the order of the observations, an undesirable property in small samples. One possible solution is to take the observations in batches, where each batch is processed coherently (Bayes) and then approximated by a QB procedure. (In Smith and Makov (1980) batches of 2 and 4 observations were used). Another possibility is to choose two (or more) possible sequences (the most likely in some sense) and then to average the QB estimators for these sequences.

3.- The consistency of the QB recursion (when so proved), as opposed to the possible asymptotic bias of the DD and PT, is inherited in the mathematical structure of the recursion and is invariant to the choice of prior. According to Stochastic-Approximation theory, certain properties of the regression $E\{\pi\text{-}w_1\}$ (see (18) above) are required for consistency. The analysis of such regressions shows that consistency is not affected by the choice of priors but by the degree of overlap between the densities of the corresponding classes. While the QB remains consistent for any degree of overlap (though its rate of convergence is affected), the other methods are consistent only for overlaps below a certain threshhold (or signal-to-noise ratio larger than some value). This also explains why the QB is asymptotically immune to initial errors (or wrong allocations).

4.- The only QB qualities investigated were asymptotic unbiasness and relative efficiency. We have no 'distance measure' to compare the QB with the coherent Bayes procedure.

5.- Prof. Dawid expresses doubts about the possibility of obtaining consistent estimator for both the 'clinical picture' and 'prevalences of disease'. His reaction cannot be contradicted as no proof of such consistency exists for *case C*, to which he refers. As for *case A*, and several models in *case B*, such consistency is proved on the

basis of identifiability. When this assumption cannot be made, the QB should not be adopted, nor should any other procedure which is based on an inappropriate model!

There are, however, cases in the medical context where the assumption of identifiability is acceptable. For instance (see Hermans and Habbema (1975)), in the diagnosis of Haemophilia carriership the identifiable mixture consists of two bivariate normal densities whose means and covariances are estimated from the data, while the mixing parameters are established through genetical considerations. Though the QB may prove to be consistent for this problem, I have my own reservations about the adequacy of its use (and indeed of the exploitation of uncorfirmed cases as a whole) in the case of small samples. (See Makov (1980) for further details).

I am not at all sure how qualitatively stable is the choice of linear discriminant function in medical diagnosis. However, in recent papers*, (O'Neill, 1978; Ganesalingam and McLachlan, 1978), it is shown that the ratio of the relevant (*asymptotic*) information contained in unclassified observation to that of classified observation is quite considerable for a statistically interesting range of separation of the populations. In Ganesalingam and Mclachlan (1979), simulation studies of *small* misclassified samples produced satisfactory results.

A.F.M. SMITH (*University of Nottingham*):

The points which have concerned the discussants of my paper also concern me.

(i) I agree with Professor Kadane that model selection criteria should really be derived using an appropriate loss or utility framework. My current work on these problems is now proceeding along decision-theoretic lines.

(ii) While the consistency properties mentioned by Professor Feinberg might seem appealing at first sight, it is not clear to me that they would be implied by all specifications of priors: I have not succeeded (yet) in sorting out precisely when they would hold.

(iii) The general points raised by Professor Dickey concerning "condition continuity" are very interesting and have been well-aired at this conference. As he himself admits, however, there is often considerable arbitrariness in the way in which a smaller model is derived from a larger by conditioning and this leaves us with the problem of providing a rationale for any particular choice. Dickey is, of course, correct in noting that my regression specification violates condition continuity, but I am also using the improper form $p(\sigma) \propto \sigma^{-1}$ and suspect that when *this* pragmatic approximation doesn't make me feel too uncomfortable neither will I feel too bad about the other.

(iv) Dr. Leonard may well be correct in suggesting that other formulations of the change-point problem could lead to simpler ways of detecting change; I look forward to seeing further details.

---

* I am indebted to Dr. D.M. Titterington for these references.

P.J. HARRISON (*Warwick University*):

We wish to thank Professor Fienberg for his amusing comments and hope to clarify some of the points which seem to have been misunderstood. We shall begin by dealing with the theoretical points.

(i) It is *not* the kinks in our loss function L which generate the pertinent discontinuities. Smooth loss functions with no jumps can exhibit the same kind of discontinuous trajectory of the corresponding Bayes decision. It is simply *easiest* to illustrate this behaviour by using a double step loss function which just happens to be discontinuous.

(ii) Although different values of α give different loss functions it should be noted that any loss function combines a utility with a function representing quantifiable loss (see DeGroot, 1970). It would be a brave man who would suggest that he knew this utility function *precisely* or, indeed, that it did not change with the decision-maker's environment. However, we have shown that slightly different utilities can give rise to extremely different Bayes decisions even when the posterior density is from a smooth and well-known family.

(iii) We hold that bounded loss functions should always be used in a Bayesian analysis. The discontinuities we discuss here cannot be considere a *fault* of using the normal conjugate loss function. Indeed, if we use quadratic loss, for example, then the corresponding expected loss function does not represent the decision-maker's dilemma when he is faced with a very bimodal posterior density. This indicates to us that this form of analysis must be lacking in some fundamental way. Under practical considerations bounded loss is a necessity due to the boundedness of the resources of the decision-maker. Theoretically the use of unbounded loss $L(\delta\text{-}\theta)$ can give rise to some very awkward paradoxes. For example if $L(\delta\text{-}\theta)$ is convex then it is easily shown that the corresponding Bayes decisions always depend solely on the comparative steepness at $\pm\infty$ of the two tails of a posterior density on the real line (see Kadane and Chuang, 1978). We must therefore reconcile ourselves to the fact that under any sensible analysis it is possible to get these sudden changes in decision.

We shall now reply to the discussion about some of the examples that we presented. Obviously in the time and space available we were only able to give the bare bones of the structure of the analysis of what are vast macro-models. A little more detail will be presented in Smith, Harrison and Zeeman (1979) and Smith (1980) has developed the ideas given in Zeeman et al (1974). Because of this lack of space we considered it most important to indicate how discontinuous phenomena can be analysed in a Bayesian way so that the underlying discontinuities of the system are not obliterated by our model. It is obvious, but not always realised, that before we can *criticize* a statistical model we must choose one. Either we acknowledge the absence of developed theory, and construct our own model, or we undertake the often very difficult task of translating a theoretical model (e.g. chemical or psychological) into an appropriate statistical one. For the prison riot case study we had to translate the theories of Konrad Lorentz first into a mathematical and then a Bayesian model. This translation was informative in itself, necessitating, for example, the use of *bounded* loss structures in our description. Since Lorentz's model is qualitative it generates a *class* of statistical models containing posterior densities and loss functions

geometrically "similar" to the normal and double conjugate respectively. Any sensibly parametrised model in this class will give the same kind of geometry to the data. We worked with the normal and its conjugate solely for computational ease. Professor Fienberg's contention that the reader can get nothing out of the model other than the data is quite untrue since the model is now in a statistical form and therefore the parameters can be estimated. These estimates together with other information can be combined by any self-respecting Bayesian to give predictive distributions for the outbreaks of violence as functions of 'Alienation" and "Tension". These are in fact being used in some British institutions. However this methodology, being long and technically dull, would have been out of place amongst the contributions presented at Valencia.

It seems very common, in forecasting and other fields, for a statistician to construct a model with no regard to the dynamics of the underlying process and just "fitting" it. Consequently the practitioner has little information communicated to him other than a short term forecast which he probably could have achieved by eye anyway. We sincerely hope that Professor Fienberg is not proposing this as the ideal (and only) function of a statistician. Although we enjoyed the paper presented by Professor Box at this conference, we felt that he might have emphasized the importance of picking a sensible class of models to begin with. Unfortunately it seems likely that his paper may be used by some statisticians as an excuse to avoid essential thought.

REFERENCES IN THE DISCUSSION

BROWN, P.J. (1976). Remarks on some statistical methods for medical diagnosis. *J. Roy. Stat. Soc.* **A 139**, 104-107.

DAWID, A.P. (1976). Properties of diagnostic data distributions. *Biometrics* **32**, 647-658.

DICKEY, J.M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *Ann. Math. Statist.* **42**, 204-223.

— (1974). Bayesian alternatives to the F test and the least squares estimate in the normal linear model. In *Studies in Bayesian Econometrics and Statistics* (S.E. Fienberg and A. Zellner, eds.) 515-554. Amsterdam: North Holland.

DICKEY, J.M. and LIENTZ, B.P. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *Ann. Math. Statist.* **41**, 214-26.

FOWLKES, E. (1979). Some methods for studying the mixture of two normal (log-normal) distributions. *J. Amer. Statist. Assoc.* **74**, 561-575.

GANESALINGAM, S. and MCLACHLAN, G.J. (1978). The efficiency of a linear discriminant function based on unclassified initial samples. *Biometrika* **65**, 658-662.

— (1979). Small sample results for a linear discriminant function estimated from a mixture of normal populations. *J. Statist. Comput. Simul.* **9**, 151-158.

GOOD, I.J. (1967). A Bayesian significance test for multinomial distributions (with discussion). *J. Roy. Statist. Soc. B* **29**, 399-431.

GOOD, I.J. and GASKINS, R.A. (1980). Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite Data. *J. Amer. Statist. Assoc.* **75**, 42-73.

GUNEL, E. and DICKEY, J.M. (1974). Bayes factors for independence in contingency tables. *Biometrika* **61**, 545-57.

HERMANS, J. and HABBEMA, J.D.F. (1975). Comparison of five methods to estimate posterior probabilities. *EDV in Medizin and Biologie,* 14-19.

KADANE, J.B. and DICKEY, J.M. (1980). Bayesian Decision Theory and the Simplification of Models. To appear in *Criteria for Evaluation of Econometric Models,* (J. Kmenta and J. Ramsey, eds.)

KADANE, J.B. and CHUANG, O.T. (1978). S table decision problems. *Ann. Statist.* **6**, 1095-1110.

LEONARD, T. (1978). Density Estimation, Stochastic Processes, and Prior Information (with discussion). *J. Roy. Statist. Soc. B* **40**, 113-146.

LINDLEY, D.V. (1968). The Choice of Variables in Multiple Regression (with discussion). *J. Roy Statist. Soc. B* **30**, 31-66.

MAKOV, U.E. (1978). An algorithm for sequential unsupervised classification. *Proceedings in Computational Statistics, "Comsptat 1978"* (Corstan, L.C.A. and Hermans, J., eds.).

— (1980). The statistical problem of unconfirmed cases in medicine. To appear in *Teoria delle Decisioni in Medicina.* (E. Girelli-Bruni ed.). Verona: Bertani.

O'NEILL, T.J. (1978). Normal discrimination with unclassified observations. *J. Amer. Statist. Assoc.* **73**, 821-826.

SMITH, A.F.M. and MAKOV, U.E. (1980). Bayesian detection and estimation of jumps in linear systems. *Proceedings of the IMA Conference on "The Analysis and Optimization of Stochastic Systems".* (O.R.L. Jacobs *et al.* eds.), 333-346. New York: Academic Press.

SMITH, J.Q. (1980). The Prediction of Prison Riot. *J. Math. Statist. Psychol.* (To appear).

SUSSMAN, H.J. and ZAHLER, R.S. (1978a). A critique of applied catastrophe theory in the behavioral sciences. *Behavioral Science,* **23**, 383-389.

— (1978b), Catastrophe theory as applied to the social and biological sciences: A critique. *Synthese* **37**, 117-216.