DISCUSSION

I.J. GOOD (*Virginia Polytechnic Institute and State University*):

Dr. Dalal discussed the application of the Dirichlet-process priors to continuous problems. In my work on categorical data, I found it necessary to use *mixtures* of Dirichlet priors (Good, 1965, 1967, 1976; Good & Crook, 1974; Crook & Good, 1980). In Good (1978) I asked whether it would be useful to use mixtures of Dirichlet processes for continuous data, such as for testing independence in continuous bivariate distributions. Also, can we apply "Ockham's hyperrazor" by somehow selecting the Dirichlet processes so that only one hyperparameter is required? If so, this could be given a hyperprior as in the categorical work.

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

One of the interesting things in non-parametric statistics is the interpretation of various interesting quantities as U-statistics. For example, Wilcoxon's statistic is an estimate of $P\,[X < Y]$. Have the modifications of Dirichlet processes been studied to see whether Wilcoxon's statistic can be justified as an estimator from this point of view?

T. LEONARD (*University of Warwick*):

Professor Dalal's convolution of the Dirichlet process seems to me to involve some really brilliant ideas. It will be regarded as one at the important contributions in the area of non-parametric density estimation. His generalisation of the Dirichlet process avoids the pitfalls faced by Ferguson, for example the problems of spiky posterior estimates and specific prior covariance structures. His prior distribution is very general and simply formulated and leads to appealingly smooth posterior estimates. He is to be congratulated on achieving an original idea of such beautiful and wide-ranging simplicity.

When specifying his prior distribution, I think that it would be helpful if Professor Dalal worked in terms of his prior covariance kernel, as well as his prior mean value function, since this would demonstrate precisely how he intends to smooth his estimates. This would also highlight the similarity between his approach and that of the early work of Whittle, who just specified the first two-moments of his prior. By completely specifying his prior Dalal achieves the same generality as Whittle, but he does not run into problems of negative posterior estimates, and he is also able to make posterior probability statements about the unknown density, as well as providing point estimates.

Professor Dalal's posterior estimates are constrained to the class of kernel estimates and I wonder whether this is a property of the type of prior distribution assumed? My own approach constructs a prior in logit space where it seems very natural to think in terms of linear relationships and covariance kernels, and my estimates assume a general non-linear form rather different from kernel estimates. The following rather undesirable properties of kernel estimates are avoided under my approach:

1) The overspread-out nature of kernel estimates (the estimated variance is always greater than the sample variance)
2) The dependence of bandwidth upon sample size in order to achieve asymptotic consistency, or under Whittle's approach the contraction to delta functions as the sample size increases.
3) The problem that when there are only a moderate number of observations kernel estimates will either oversmooth or possess bumps in the tails.

I think that the great strenght of a Bayesian approach to nonparametric density estimation lies in the fact that it permits us to model the density via its prior estimate whilst avoiding any constraint on the posterior estimate to belong to a parameterized family. It for example provides a particularly viable alternative to classical tests for fit, since we simply need to investigate differences between the posterior estimate and the prior hypothesised estimate.

A. O'HAGAN (*University of Warwick*):

Professor Dalal has shown us a very interesting formulation of nonparametric inference. The so-called nonparametric problems are characterised in his approach, and in the earlier work of Ferguson, by a vast number of parameters. I believe this feature is inevitable: even when inference centres on some subparameter like the median, Professor Dawid has shown in his paper at this meeting that nuisance parameters cannot be dismissed without careful consideration. Given that there really are infinitely many parameters, *only* a Bayesian approach is feasible. The problem is underidentified (or overparametrised) and no amount of data will give sufficient information to render the prior irrelevant. In particular, the way in which the prior relates parameters to each other influences strongly the shape of posterior inferences. Ferguson's Dirichlet process, for example, yields discontinuous posterior means. It is not enough that the prior should look sensible; it must also give sensible posterior inferences, and it is quite proper for Professor Dalal to seek for priors which give posterior inferences having sensible shapes.

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

I hope that all who have enjoyed Professor Dalal's elegant presentation and admired his undoubted mathematical ingenuity will forgive me if I express the philistine sentiment that exercises involving contemplation of completions of spaces of mixtures of Dirichlet processes have very little to do with interpreting data in the light of personal judgment, and, whatever else they are, are *not* Bayesian Statistics.

Instead of seeking a tractable way of representing the uncontemplatable (i.e. measures having large support over gigantic spaces of distributions), we should first of all decide what aspects of the problem we *are* able to contemplate and *then* seek a tractable representation.

As an example of what I have in mind, suppose we want to make inferences about location, given up to 50 observations from an (unknown) member of the (assumed) location-scale family. I *can* contemplate qualitative features that may be relevant -such as heaviness of tails, skewness, etc.- and I *can* realize that with samples of this size there is little point in seeking a prior measure with large support in the location-scale family. (We simply cannot distinguish other than quite crude qualitative differences between distributions.) Instead, a sufficiently rich mixture should result from a prior with a sensibly chosen representative *finite* support. One such crude choice which incorporates heavy, and light-tailed departures from Normality, together with skewness in both directions, is a finite mixture model consisting of the Normal, Uniform, Laplace, Right-Exponential and Left-Exponential distributions.

This has the added advantage that all the necessary Bayesian manipulations can be carried out analytically. (See Spiegelhalter, 1978.)

I have always understood "Nonparametric" to mean "Enormous Parameter (Model) Space", where "enormous" signifies "too big to have to think meaningfully about". I suggest, therefore, that we should be very circumspect about any theory which couples "Nonparametric" with the word "Bayesian".

S.R. DALAL (*Rutgers University*):

Professor Good during his discussion at the conference inquired about the suitability of symmetric Dirichlet distributions and associated processes as priors for nonparametric problems. Use of these priors in contingency tables leads to manageable numbers of hyperparameters and some ease in numerical computations (Good, 1976). Unfortunately, in many interesting nonparametric problems, the interesting sets are of various sizes, and thus, the kind of symmetry inherent in contingency tables is absent. This rules out the use of symmetric Dirichlet distributions. However, as indicated in the paper we can use Dirichlet symmetric processes whenever some appropriate invariance structures can be assumed. Professor Good's comment on the use of "Ockham's hyperrazor" needs further investigation.

Professor Kadane has raised an interesting and an important issue related to justification of classical nonparametric procedures based on U statistics through the nonparametric Bayes theory. This line of inquiry has already been followed in Professor Ferguson's fundamental paper. He showed that in the problem of estimation of $\int FdG$ with a squared error loss, the Bayes estimate is a convex linear combination involving the Mann-Whitney statistic. Similar justification can be provided for several other nonparametric procedures. For example, my work with Professor Phadia has shown that Kendall's π can also be similarly interpreted from Bayesian point of view.

Dr. Leonard has been very kind in praising my work on density estimation. The applicability and usefulness of my approach can be judged only after examining the complexity of the estimates, the large sample properties (e.g. consistency rates of convergence), etc. In this regard, the references furnished by Dr. Leonard to his work (1973), Whittle's work (1958) and Good and Gaskin's work (1971) will be very useful.

Dr. Leonard is also quite correct in pointing out that the posterior estimates are constrained to the class of kernel estimates because of the nature of the prior. However, in the important problem of unimodal density estimation this is not a constraint. Dr. Leonard has also been able to convince me that it would be helpful to work in terms of covariance kernels. I think this deserves detailed investigation.

I do concur with Dr. O'Hagan's comment on the inevitability of the parametrization by large number of parameters in Bayes formulation of nonparametric problems. This is not to say that in such a formulation no amount of data will give sufficient information to render the prior irrelevant. In fact, I think that some sort of generalized version of the theory of precise measurement would hold and accordingly the precise nature of the large number of parameters involved would be unimportant.

Professor Smith comments that we would be circumspect about any theory which couples 'Nonparametric' with the word 'Bayesian'. I disagree with his logic. Much recent works shows that suchs an alliance is not an unholy one. This is also best illustrated in the usual one sample problem where observations are obtained as differences of pairs of measurements. Here the assumption of symmetry is easily justified and beliefs about the point of symmetry may also be easily parametrized. Savage's theory of precise measurement tells us that the precise formulation of beliefs about the point of symmetry is immaterial. However, an incorrect specification of the model does have serious consequences for the Bayesian (e.g. Berk, 1966). In this instance, whithout any additional information, the Bayesian nonparametric theory is certainly a viable contender to any other form of Bayes analysis. Also, if Dirichlet symmetric processes are used as priors, then a generalization of Savage's theory of precise measurement suggests that the parameter α of the process need not be precisely specified.

Professor Smith also contends that the results related to completion of spaces of mixtures of Dirichlet processes are not part of Bayesian statistics. This may be true in a narrow sense. However, disregarding its Bayesian implications will be a mistake. The result which Professor Smith refers to says that a Bayesian, in quest for a suitable prior for a nonparametric problem, need not go beyond the class of mixtures of Dirichlet processes. A parametric counterpart would say that the Bayesian need not go beyond the class of mixtures of natural conjugate priors. (Dalal and Hall, 1977).

REFERENCES IN THE DISCUSSION

BERK, R.H. (1966). Limiting behavior of posterior distributions when the model is incorrect. *Ann. Math. Statist.* **37**, 51-58.

CROOK, J.F. and GOOD, I.J. (1980). On the application of symmetric Dirichlet distributions and their mixtures to contingency tables, Part II. *Ann. Statist.* (in press).

GOOD, I.J. (1965). *The estimation of Probabilities: An Essay on Modern Bayesian Methods.* Cambridge, Massachusetts: The M.I.T. Press.

— (1967). A Bayesian significance test for multinomial distributions. *J. Roy. Statist. Soc. B* **29**, 399-431 (with discussion). *Corrigendum* **36** (1974), 109.

— (1976). On the application of symmetric Dirichlet distributions and their mixtures to contingency tables. *Ann. Statist.* **4**, 1159-1189.

— (1978). Review of Ferguson, Thomas S., "Prior distributions on spaces of probability measures". *Ann. Statist.* **2**, (1974) 615-629; *Math. Rev.* **55**, 1546-1547.

GOOD , I.J. and CROOK, J.F. (1974). The Bayes/non-Bayes compromise and the multinomial distribution. *J. Amer. Statist. Assoc.* **69**, 711-720.

GOOD, I.J. and GASKINS, R.A. (1971). Non-parametric roughness penalties for probability densities. *Biometrika* **58**, 255-277.

LEONARD, T. (1973). A Bayesian method for histograms. *Biometrika* **60**, 297-308.

SPIEGELHALTER, D.J. (1978). *Adaptive inference using a finite mixture model.* Ph.D. Thesis. London: University College.

WHITTLE, P. (1958). On the smoothing of probability density functions. *J. Roy. Statist. Soc. B* **20**, 334-343.