Nonparametric Bayes Decision Theory

S.R. DALAL
Rutgers University[*]

SUMMARY

A summary of the seminar with the same title is presented. Ferguson's fundamental work on the theory of Dirichlet processes is elucidated and their shortcomings are discussed. Some modifications are also proposed and illustrated. Some of the intricate mathematical issues related to the definitions and the proofs are not discussed for the sake of clarity and brevity. The development related to unimodal processes, briefly mentioned in the last section, will appear as a joint work with Professor W.J. Hall elsewhere.

Keywords: NONPARAMETRIC BAYES; DIRICHLET PROCESSES; DIRICHLET INVARIANT PROCESSES; ESTIMATION OF A SYMMETRIC C.D.F.; ESTIMATION OF A UNIMODAL DENSITY; ESTIMATION OF A POINT OF SYMMETRY.

1. INTRODUCTION

Nonparametric theory deals with the problems of inference when the underlying distribution is not specified in terms of a parametric family. This theory can be gainfully employed in many situations as models are seldom more than approximations to reality, and the procedures which are optimal for a given parametric family (i.e. the 'Idealized Model') may perform poorly even for models which are near to the idealized model (e.g. Tukey (1960), Huber (1964)).

However, 'classical' nonparametric theory disregards much of the existing knowledge about the idealized model. Further, evaluations and comparisons are usually carried out asymptotically at specific parametric models.

To avoid these shortcomings, it is useful to think that there is an idealized model and that the observed distribution is a (possibly random) perturbation of the idealized model. This approach has been used by Huber (1972) and

[*] Presently at Bell Labs, Murray Hill, N.J. 07974, U.S.A.

others to create an elegant theory of robustness. Here, we explore an alternative approach of Ferguson (1973), who derived, and suggested the use of Dirichlet processes as priors for nonparametric problems. Specifically, we shall review Dirichlet processes (Section 2), note their anomalies and inadequacies (Section 3), and suggest some modifications.

2. DIRICHLET PROCESSES

Let $(\chi, B(\chi))$ be the sample space and P^* be the space of all probability measures on $(\chi, B(\chi))$. P^*, the parameter space for many nonparametric problems is quite large and consequently many procedures turn out to be minimax. Hence, the Bayes criterion becomes more relevant.

For the Bayes framework, it is necessary to consider a class of priors over P^*, i.e., a class of *random probability measures,* which is a) *mathematically tractable,* b) *rich*, and c) *easy to parameterize.* Several procedures have been suggested toward this end, notably by Dubins and Freedman (1966), Kraft and Van Eeden (1964), Rolph (1968), Ferguson (1973), Doksum (1974) and Sethuraman (1979). In statistical inference Ferguson's priors, Dirichlet processes, have been more often used that the other procedures, because of their intuitive properties and tractability, e.g., Ferguson (1974), Susarla and Van Ryzin (1976), Phadia and Susarla (1979), Berry and Christensen (1979).

Mixtures of Dirichlet processes, proposed by Antoniak (1974), have also been used in Bio-assay and regression-type problems. Relatively few applications not related to Dirichlet processes are available. For example, Ferguson and Phadia (1979) have dealt with censored data problems using Doksum's *neutral to right processes.* Also, some new non-Dirichlet-process priors developed by Sethuraman (1979) may prove to be useful. We shall, however, follow Ferguson's approach with some modification. Before delineating the modifications, we define and briefly state some elementary properties of Dirichlet processes below.

Definition. A random probability P is a Dirichlet process if there exists a finite, finitely, additive measure, α, such that for every measurable partition $B_1,\ldots,B_k$ of χ, $(P(B_1),\ldots,P(B_k))$ has a Dirichlet distribution with parameters $(\alpha(B_1),\ldots,\alpha(B_k))$. We then write $P \epsilon DP(\alpha)$ and denote the corresponding random probability measure by P.

Elementary properties. Let $\alpha = M \cdot Q$, where M is a positive number, and Q is a probability measure on $(\chi, B(\chi))$. Then $P \in D(\alpha)$ implies that $\mathcal{E}P = Q$. Q, thus, can be thought of as ideal distribution. Further, the number M can be viewed as the prior example size (e.g. Novick and Hall, 1965). Using these properties DP priors can be *easily parameterized.*

The second desirable property, Richness, mentioned earlier, has two aspects. First, richness of support is essential to deal with a broad class of nonparametric problems. Secondly, if one is to restrict attention to a specified class of priors, it is essential for this class to have members capable of approximating any prior belief. We call this latter aspect *adequacy*. Both of these issues can be examined by imposing a 'natural' topology on P^*, the space of all probability measures, and P^{**}, the space of all random probability measures.

For a lack of the "natural" topology, various topologies can be considered (e.g. Ferguson (1973), Dalal (1978), Dalal and Hall (1980)). By considering the weak* topology on P^* obtain by imbedding P^* on the product space $[0,1]_{\beta}{}^{(\chi)}$, Ferguson (1973) showed that all α-absolutely continuous distributions are in the support of $DP\,(\alpha)$. Dalal (1978) showed that this kind of imbedding leads to random probability measures which select finitely additive probability measures on $(\chi, B\,(\chi))$ with probability one. Further, although the class of Dirichlet processes is not *adequate* in terms of approximating a given belief, a convex hull of this class of mixtures of Dirichlet processes (MDP) is adequate in this regard (see Dalal (1978), Dalal and Hall (1980)).

The mathematical tractability of any class of priors can only be ascertained by examining the ease with which the posterior and various simple expectations are obtained. With respect to Dirichlet processes, Ferguson showed that, given a sample $X_1,\ldots,X_n$, the posterior is $DP\,(\alpha + \Sigma\delta_{xi})$, where δ_x is the unit mass degenerate at x. This conjugate prior property has been extensively used in applications.

3. SHORTCOMINGS AND MODIFICATIONS WITH APPLICATIONS

First, we discuss an anomaly (discreteness), and an inadequacy (to deal with invariant problems) of Dirichlet processes. This is followed by some modifications to overcome these defects. A few illustrative example are also given.

3.1. *Shortcomings*

i) **Discreteness.** It is known that *DP*'s are discrete with probability one (e.g. Blackwell (1973), Berk and Savage (1977)). This discreteness is more than a technical aberration. In some applications this has led to non-intuitive answers. Further, the posterior changes the masses only at the observed sample points. Intuitively, however, it would be appealing to have a prior which increases the probability of a neighborhood instead.

ii) **Invariance.** In nonparametric problems, one is permitted to have nonparametric beliefs, e.g. symmetry of the underlying distribution (i.e. in the one-sample problem), exchangeability, spherical symmetry, etc. However *DP* and the other priors defined so far live on the class of all probability measures. It would be useful to also have priors giving probability one to invariant (under a group g) probability measures.

3.2. *Modifications*

Our approach, simply stated, is to modify the sample paths of a *DP* (i.e. the distributions selected by a *DP*) to eliminate these shortcomings. The modified process, although closely related to the *DP*, is usually more complex. However, one can use the updated version of the *DP* to manipulate the posterior of the modified process.

3.2.1. *Modifications related to Invariance*

Let $g = \{g_1,\ldots,g_k\}$ be a finite group of measurable transformations on $(\chi,B\,(\chi))$ and P be a random probability measure. Define a new random probability measure Q by the mapping $Q\,(A) = \frac{1}{k}\,\Sigma P\,(g_iA)$. Clearly Q selects g-invariant distributions with probability one. Such a scheme can also be used with a compact topological group to obtain invariant distributions with probability one. When P is a Dirichlet process with g-invariant α, Q is called the Dirichlet Invariant process ($DIP(\alpha)$). These kind of processes have been studied in Dalal (1979a). The behavior of *DIP*s is similar to *DP*s, e.g. if $x_1,\ldots,x_n$ is a sample from $Q\epsilon DIP$, then $Q|X_1,\ldots,X_n$, the posterior of Q, is $DIP(\alpha + k^{-1}\sum_i\sum_j g_iX_j)$.

Using *DIP*'s Bayes estimates of various functionals can be obtained. Some illustrative applications are considered below.

i) **Estimation of a symmetric c.d.f.** Consider the problem of estimating a c.d.f., F_μ symmetric about a known point μ. Let the loss function be $L(F_\mu,\hat{F}) = \int(F_\mu(t) - \hat{F}(t))^2dW(t)$, where W is a finite prespecified weight function. For Bayes estimation, consider the prior $DIP(\alpha)$, where $\alpha = M{\cdot}Q$ and the group g is $\{g_1,\, g_2\}$; $g_1(x) = 2\mu\text{-}x$, $g_2(x) = x$. Let G be the c.d.f. corresponding to Q. The Bayes estimate then can be shown to be a convex linear combination of the initial guess G and the μ-symmetrized empirical c.d.f. $\hat{F}_n$ (Dalal, 1979a), i.e.

$$\hat{F}_\mu = \frac{M}{M+n}\,G + \frac{n}{M+n}\,\hat{F}_n.$$

As n becomes larger, the dependence on the initial guess G becomes weaker. Also the expression for $\hat{F}_\mu$ suggests that M can be thought of as the prior sample size, as discussed earlier.

The above Bayes formulation can also be exploited to get a minimax estimate,

$$\hat{F}_\mu = \frac{1}{4(1+\sqrt{n})} + \frac{1}{2(\sqrt{n}+n)}\Sigma(\delta_{x_i} + \delta_{2\mu-x_i}) + \frac{1}{2(1+\sqrt{n})}\delta_\mu .$$

Bayes estimates of F_μ for μ unknown have also been obtained in Dalal (1979a).

ii) **Estimation of the unknown center of symmetry.** Consider the usual one sample problem of estimating the center of symmetry of an arbitrary distribution F. Specifically, assume the following model $Y_i = \mu + \Delta_i$ where the Δ_i are i.i.d. with an arbitrary distribution, F, symmetric about 0. For the Bayes formulation, Let $F \epsilon DIP(\alpha)$, $\alpha = M \bullet G$, and μ have the non-informative uniform distribution over the reals. Then the Bayes estimate using squared error loss $(\mu-\hat{\mu})^2$ can be found. In the case of the idealized model, G, being standard normal (density φ), and assuming all distinct observations the Bayes estimate $\hat{\mu}$ is (Dalal, 1979b)

$$\hat{\mu} = \frac{M}{M+n}\bar{y} + \frac{n}{M+n}\mu^* ,$$

where

$$\mu^* = (\sum_{i<j} w_{ij}\frac{y_i+y_j}{2} / \sum_{i<j} w_{ij})$$

and

$$w_{ij} = (\prod_{k \neq i,j} \varphi(y_k - \frac{y_i+y_j}{2}))(\varphi\frac{y_i-y_j}{2}).$$

$\mu^{*'}$ is a weighted mean of pairwise averages. The weight given to the pair $\frac{y_i+y_j}{2}$ is inversely proportional to the distance between y_i and y_j, and the distance of $\frac{y_i+y_j}{2}$ from the rest of observations. Thus, this estimate is robust. Numerical and other investigations on this estimate are considered in Dalal (1979b).

3.2.2. *Modifications related to continuity unimodality*

In density estimation problems, the usual estimates are obtained by smoothing out the functionals of the empirical c.d.f. This is usually accomplished by convoluting with different kernels.

This kind of idea can be used to smooth out the sample paths of the *DP*'s followed by sampling from the smoothed process. Some of these ideas have been considered in a joint work with Professor W.J. Hall of the University of Rochester. Using this, Bayes estimates of general densities, unimodal densities, modes, etc., can be obtained. Some details have been worked out in Dalal and Hall (1977).

ACKNOWLEDGEMENTS

I wish to thank Professor W.J. Hall for various helpful discussions. This research was supported by NSF Grant MCS 78-02160 and Rutgers University Research Council.

REFERENCES

ANTONIAK, C. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2**. 1152-1174.

BERK, R.H. and SAVAGE, I.R. (1977). Dirichlet processes produce discrete distributions: An elementary proof. *Tech. Rep.* Rutger University.

BERRY, D.A. and CHRISTENSEN R. (1979). Empirical Bayes estimation of a binomial parameter via mixtures of Dirichlet processes. *Ann. Statist.* **7**, 558-69.

BLACKWELL, D. (1973). Discreteness of Ferguson selections. *Ann. Statist.* **1** 356-358.

DALAL, S.R. (1978). A note on the adequacy of mixtures of Dirichlet processes. *Sankhya, A*, **40**, 185-91.

— (1979a). Dirichlet Invariant processes and applications to nonparametric estimation of symmetric distribution functions. *Stoch. Proc. and Appl.* **9**, 99-107.

— (1979b). Nonparametric and robust Bayes estimation of location. *Proc. Optimizing Methods in Statistics.* 141-166. New York: Academic Press.

DALAL, S.R. and HALL, G.J., Jr. (1980). On approximating parametric Bayes models by nonparametric Bayes models. *Ann. Statist.* **8**, 664-672.

DALAL, S.R. and HALL, W.J. (1977). Unimodal density estimation. *Tech. Rep.* Rutgers University.

DOKSUM, K. (1974). Tailfree and neutral random probabilities and their posterior distributions. *Ann. Prob.* **2**, 183-201.

DUBINS, L.E. and FREEDMAN, D.A. (1966). Random distribution function. *Proc. 5th Berkeley Symp.* **2**. 183-214. Berkeley: University of California.

FERGUSON, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209-230.

— (1974). Prior distributions on spaces of probability measures. *Ann. Statist.* **2**, 615-629.

FERGUSON, T.S. and PHADIA, E.G. (1979). Bayesian nonparametric estimation based on censored data. *Ann. Statist.* **7**, 163-86.

HUBER, P.J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35**, 73-101.

— (1972). Robust statistics: a review. *Ann. Math. Statist.* **43**, 1041-1067.

KRAFT, C.H. and VAN EEDEN, C. (1964). Bayesian bio-assay. *Ann. Math. Statist.* **35**, 886-890.

NOVICK, M.R. and HALL, W.J. (1965). A Bayesian indifference procedure. *J. Amer. Statist. Assoc.* **60**, 1104-17.

PHADIA, E.G. and SUSARLA, V. (1979). An empirical Bayes approach to two sample problem with censored data. *Comm. in Statist.* **8**, 1327-1351.

ROLPH, J.E. (1968). Bayesian estimation and mixing distributions. *Ann. Math. Statist.* **39**, 1289-1302.

SETHURAMAN, J. (1979). Personal Communication.

SUSARLA, V. and PHADIA, E.G. (1976). Empirical Bayes Testing of a distribution function with Dirichlet process priors. *Comm. in Statist. A,* **5**, 4505-69.

SUSARLA, V. and VAN RYZIN, J. (1976). Nonparametric Bayesian estimation of survival curves from incomplete observations. *J. Amer. Statist. Assoc.* **71**, 897-902.

TUKEY, J.W. (1960). A Survey of Sampling from Contamined Distributions. In *Contributions to Probability and Statistics.* (Olkin ed.) Stanford: University Press.