Commentary on *A Note on the Dirichlet Process*

Jayaram Sethuraman

Consider the standard inference problem where the data $\mathbf{X} = (X_1, \dots, X_n)$, taking values in $\mathcal{X}^n$, consist of independent observations from a common distribution F or probability measure P on $(\mathcal{X}, \mathcal{B})$. The essentials of Bayesian analysis are well developed for this problem, when the common distribution F is completely specified in terms of a finite number k of parameters $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$. A prior distribution for this problem is just a suitable distribution on a subset in R^k described by any restrictions the parameters $\boldsymbol{\theta}$ must satisfy. The calculation of posterior distribution become routine from this point onwards. Special classes of prior distributions called conjugate families have the property that the posterior distribution is also in the same class. Computations become simpler when the conjugate family is itself parametrized by a finite number of parameters. If the prior distribution in a conjugate family can be described by the parameters $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_m)$ then the posterior distribution based on data $\mathbf{X}$ would also be in the same conjugate family with parameters $\boldsymbol{\lambda}^{\mathbf{X}} = (\lambda_1^{\mathbf{X}}, \dots, \lambda_m^{\mathbf{X}})$; the term “updated” has been used to describe the parameters of the posterior distribution.

When $\mathcal{X}$ consists of k points, the common distribution P is a discrete distribution and it can be completely described by $\mathbf{p} = (p_1, \dots, p_k)$ satisfying $p_i \geq 0, i = 1, \dots, k, \sum_1^k p_i = 1$. In view of the linear dependence among these parameters it is enough to specify a prior distribution just for $(p_1, \dots, p_{k-1})$; in other words, P is parametrized by $k - 1$ parameters satisfying the above conditions. A natural conjugate family is the finite dimensional Dirichlet distribution $\mathcal{D}(\lambda_1, \dots, \lambda_k)$, which is the distribution of $(\frac{Z_1}{Z}, \dots, \frac{Z_{k-1}}{Z})$ where $Z = Z_1 + \dots + Z_k$ and $Z_1, \dots, Z_k$ are independent Gamma random variables with parameters $\lambda_1, \dots, \lambda_k$ respectively. The parameters $(\lambda_1, \dots, \lambda_k)$ will have to satisfy the conditions $\lambda_1 \geq 0, \dots, \lambda_k \geq 0, \sum_1^k \lambda_i > 0$. When all the λ_i are positive, the distribution $\mathcal{D}(\lambda_1, \dots, \lambda_k)$ can be defined in a more familiar way by a pdf proportional to $\prod_1^{k-1} p_i^{\lambda_i-1} (1 - \sum_1^{k-1} p_j)^{\lambda_k-1}$. The data $\mathbf{X} = (X_1, \dots, X_n)$ which are i.i.d. P can also be summarized by its empirical distribution function $F_n(j) = \frac{1}{n} \sum_1^n I(X_i = j), j = 1, \dots, k$. In this case the posterior distribution can easily be shown to also be the finite dimensional Dirichlet distribution $\mathcal{D}(\lambda_1 + nF_n(1), \dots, \lambda_k + nF_n(k))$.

To perform nonparametric inference when the common distribution P is not restricted to such parametric classes one should study classes of probability distributions for P which varies in the space $\mathcal{P}$ of all probability measures on $(\mathcal{X}, \mathcal{B})$. A natural σ -field in $\mathcal{P}$ is $\sigma(\mathcal{P})$, the smallest σ -field such that sets of the form $\{P : P(B) < r\}$ where $B \in \mathcal{B}$ and $r \in [0, 1]$. One can consider P as $(P(B), B \in \mathcal{B})$ and thus take P as an element of $[0, 1]^\infty$ satisfying the familiar countable additivity assumptions. One can also consider P to be the vectors $(P(B_1), \dots, P(B_k))$ satisfying

J. Sethuraman (✉)

Distinguished Professor of Statistics, Florida State University, Tallahassee, FL 32306

e-mail: sethu@ani.stat.fsu.edu

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_8, © Springer Science+Business Media, LLC 2011

$P(B_i) \geq 0, i = 1, \dots, k, \sum_1^k P(B_i) = 1$ over all finite measurable partitions $(B_1, \dots, B_k)$ of $\mathcal{X}$ and satisfying some other conditions. Ferguson (1973) used the above description to define a probability measure on $\mathcal{P}$ as follows. Let α be a non-zero finite measure on $(\mathcal{X}, \mathcal{B})$. Ferguson (1973) defined $\mathcal{D}_\alpha$, the Dirichlet process with parameter α , to be the probability measure on $(\mathcal{P}, \sigma(\mathcal{P}))$ under which $(P(B_1), \dots, P(B_k))$ has the finite dimensional Dirichlet distribution $\mathcal{D}(\alpha(B_1), \dots, \alpha(B_k))$, for each measurable partition $(B_1, \dots, B_k)$ of $\mathcal{X}$.

The paper of Ferguson (1973) revolutionized the subject of nonparametric Bayes methods by showing that when $\mathcal{D}_\alpha$ is used as a prior distribution for P , the posterior distribution becomes the Dirichlet process $\mathcal{D}_{\alpha+nF_n}$ where F_n is the empirical measure of the data $\mathbf{X}$. This means that Dirichlet processes form a conjugate family of priors in the standard nonparametric problem. This paper further showed how to obtain Bayes estimates of several functions of P . It also established that Dirichlet process $\mathcal{D}_\alpha$ is concentrated on the subset of all discrete probability measures in $\mathcal{P}$.

The papers Blackwell (1973) and Blackwell and MacQueen (1973) appeared in the same journal issue where Ferguson's article appeared. These papers gave other ways to define a Dirichlet prior and to establish its properties.

Berk and Savage (1979) also gave an elementary proof of the result that Dirichlet processes concentrate on the collection of discrete measures.

The treatment of measure theoretical issues involved in all this was not satisfactory. For instance it was not clear that the Dirichlet process was well defined as a probability measure on $(\mathcal{P}, \sigma(\mathcal{P}))$. Do we know that the set $\mathcal{P}_0 \stackrel{\text{def}}{=} \{P : P \text{ is a discrete probability measure}\}$ is in $\sigma(\mathcal{P})$ before asserting that it had probability 1 under a Dirichlet process? How general can the space $\mathcal{X}$ be? These and other such questions remained.

The paper of Basu and Tiwari (1982) is a delightful paper that clears up all these questions. The paper starts out by describing the nature of general Bayes inference. It describes carefully and in detail the properties of finite-dimensional Dirichlet distributions which form a conjugate family in the standard nonparametric problem concerning random variables taking values in $\mathcal{X}$ consisting of a finite number of points. A quick extension when $\mathcal{X}$ is countable is presented next.

The case where $\mathcal{X}$ is a Borel space is the main focus of this paper. (Separable complete metric spaces are examples of Borel spaces). Borrowing from the ideas in Blackwell (1973), Basu and Tiwari establish the existence and properties of a Dirichlet process in this general case, with care and in an elementary way.

The highlight of this paper consists of Sections 6, 7 and 8. A very clear exposition of several measurability issues and the existence of the Dirichlet process are presented in these sections. It is shown that the function $P \rightarrow P_d(\mathcal{X})$ which gives the total sum of the probability masses of the discrete part of P is a measurable function from $(\mathcal{P}, \sigma(\mathcal{P}))$ to $[0, 1]$. Again it is established that the collection, $\mathcal{P}_0$, of all discrete probability measures on $(\mathcal{X}, \mathcal{B})$ is a measurable set in $(\mathcal{P}, \sigma(\mathcal{P}))$. It is only after this has been established, it makes sense to say that Dirichlet processes gives probability 1 to this set and this fact is established next. It is further shown that the collection $\mathcal{P}'$ of all probability measures whose support is $\mathcal{X}$ is also a measurable set in $(\mathcal{P}, \sigma(\mathcal{P}))$. It is only after this has been established, it makes sense to say that $\mathcal{D}_\alpha$ has support $\mathcal{P}'$ if α has support $\mathcal{X}$ and this result is also part of the paper.

David Blackwell's visit to Florida State University in 1978–79 gave Basu, Tiwari and myself inspiration to work on Dirichlet processes. In the course of my lectures at that time I discovered a constructive definition of a Dirichlet process which does not place restrictions of the space $\mathcal{X}$. This result appears in Sethuraman (1994).

References

- [1] Basu, D. and Tiwari, R. C. (1982) A note on Dirichlet processes *Statistics and Probability Essays in Honor of C. R. Rao Ed. Kallianpur, G, Krishniah, P. R. and Ghosh, J. K.* North Holland Publ. Co. 82–103.

- [2] Berk, R. and Savage, I. R. (1979) Dirichlet processes produce discrete measures: An elementary proof *Contributions to Statistics. Jaroslav Hájek Memorial Vol.* Academia, North Holland, Prague 25–31.
- [3] Blackwell, D. (1973) Discreteness of Ferguson selections *Ann. Statist.* **1** 356–358.
- [4] Blackwell, D. and MacQueen, J. B. (1973) Ferguson distributions via Pólya schemes *Ann. Statist.* **1** 353–355.
- [5] Ferguson, T. S. (1973) A Bayesian analysis of some nonparametric problems *Ann. Statist.* **1** 209–230.
- [6] Sethuraman, J. (1994) A constructive definition of Dirichlet priors (1994) *Statistica Sinica* **4** 639–650.