PUBLICATIONS MATHÉMATIQUES ET INFORMATIQUES DE RENNES

CHARLES H. KRAFT

**Posterior Distributions for Non-Parametric Priors**

*Publications des séminaires de mathématiques et informatique de Rennes*, 1969-1970, fascicule 2
« Séminaire de probabilités et statistiques », , exp. nº 6, p. 1-7
<http://www.numdam.org/item?id=PSMIR_1969-1970___2_A6_0>

# POSTERIOR DISTRIBUTIONS FOR NON - PARAMETRIC PRIORS

by

Charles H. KRAFT

(University of Montreal)

This work was done while the author was visiting the Department of Mathematics of the University of Rennes. It was partially supported by the National Research Council of Canada.

## 1. INTRODUCTION.

The purpose of this note is to give a formula for the calculation of the conditional, given a sample $U_1, \ldots, U_n$ from F, distribution of a randomly selected distribution function F. The sole restriction on the method of selection is that F is chosen, with independent interpolation, by the method of Kraft and Van Eeden [5].

Ferguson [3] gives a method of selecting a prior which also admits a formula for the calculation of the posterior. His selection has the advantage that it can be used to describe a prior for a distribution on a completely arbitrary sample space. If the sample space is the unit interval the method here includes Ferguson selection (see Antoniak [1]) as well as selections which concentrate on absolutely continuous distributions.

The method, described in [4] (see also [6]), of concentrating the prior on absolutely continuous distribution functions F on $[0,1]$ requires that $\mathcal{E}F(x) = x$. This method can be adapted (see [5]) to concentrate on absolutely continuous distributions G on the real line by letting $G(x) = F(H_0(x))$ for a fixed absolutely continuous $H_0$. On this case, of course, $\mathcal{E}G = H_0$.

. PRIORS, SAMPLES, AND POSTERIORS.

Let $\{X(\frac{k}{2^m})\}$, $m = 1,2,3, \ldots,$

$k = 1,3,5, \ldots, 2^m - 1$, be sequence of completely independent random variables each taking in $[0,1]$. It can be supposed that the $X(\frac{k}{2^m})$ have densities $P_{X(\frac{k}{2^m})}$ with respect to a fixed measure $\mu$ on $[0,1]$.

Let $F_m$ be the distribution function that gives mass to the intervals $[0, \frac{1}{2^m}]$, $(\frac{1}{2^m}, \frac{2}{2^m}]$, ..., $(\frac{2^m-1}{2^m}, 1]$ as determined by the density

$$P_m = \prod_{i=1}^{m} q_i \text{ where}$$

$$1/2\, q_1 = X(1/2).I[0,1/2] + [1-X(1/2)] . I(1/2,1]$$

$$1/2\, q_2 = X(1/4)\, I[0,1/4] + (1-X(1/4) \quad I(1/4,1/2] +$$

$$+ X(3/4)\, I(1/2, 3/4] + (1-X(3/4))\, I(3/4)1]$$

$$\vdots$$

$$1/2\, q_m = \sum_{\substack{k=1 \\ k\,odd}}^{2^{m-1}} X(\frac{k}{2^m})\, I(\frac{k-1}{2^m}, \frac{k}{2^m}] + (1-X(\frac{k}{2^m})\, I(\frac{k}{2^m}, \frac{k+1}{2^m}]$$

Let F then be the right continuous distribution function determined by $\lim_m F_m(\frac{i}{2^j})$,

$j = 1,2,\ldots$

$i = 1,3,5,\ldots, 2^j-1.$

The following alternate definition of F ;

$$F(1/2) = X(1/2) \qquad \text{or} \qquad \begin{array}{l} F(1/2) - F(0) = X(1/2) \\ F(1) - F(1/2) = 1-X(1/2) \end{array}$$

$$F(1/4) = X(1/4)\, X(1/2)$$
$$F(3/4) = X(1/2) + X(3/4)\left[1-X(1/2)\right]$$

or

$$F(1/4) - F(0) = X(1/2)\, X(1/4)$$
$$F(1/2) - F(1/4) = X(1/2)\left[1-X(1/4)\right]$$
$$F(3/4) - F(1/3) = \left[1-X(1/2)\right] X(3/4)$$
$$F(1) - F(3/4) = \left[1-X(1/2)\right]\left[1-X(3/4)\right]$$

etc...

makes it clear that F is determined by successive interpolations with the variables $X\left(\frac{k}{2^m}\right)$. The distribution obtained for F will be described by saying F is determined by interpolation with independent $X\left(\frac{k}{2^m}\right)$.

After F is determined, let $U_1, \ldots, U_n$ be a sample of n independent observations with $P(U_i \leq t) = F(t)$. Define random variables $\{n_{m,j}\}$, m=1,2,3,... $j = 1,2,\ldots, 2^m$ by $n_{mj}$ = (the number of $U_i$ in $\left(\frac{j-1}{2^m}, \frac{j}{2^m}\right)$) where as above the interval for j=1 includes 0 while those for j > 1 are open on the left and closed on the right. It is clear that the $\{n_{m,j}\}$ determine, uniquely, the sample cumulative $G_n(t) = \frac{\text{Of } U_i \leq t}{n}$.

With these definitions the following theorem and immediate corollary can be given.

Theorem.

The conditional, given $U_1,\ldots,U_n$, distribution of F is that of $F^{G_n}$ where $F^{G_n}$ is determined by independant interpolation with variables $\{\mathcal{Z}\left(\frac{k}{2^m}\right)\}$ and $\mathcal{Z}\left(\frac{k}{2^m}\right)$ has the density with respect to μ.

$$P_{\mathcal{Z}\left(\frac{k}{2^m}\right)}(x) = \frac{x^{n_{m,k}}(1-x)^{n_{m,k+1}}}{\mathcal{E}\left[X\left(\frac{k}{2^m}\right)\right]^{n_{m,k}}\left[1-X\left(\frac{k}{2^m}\right)\right]^{n_{m,k+1}}} \cdot P_{X\left(\frac{k}{2^m}\right)}(x)$$

Corollary.

$\mathcal{E}(F|U_1,\ldots,U_n)$ is the distribution function determined by interpolation with the numbers

$$a\frac{k}{2^m} = \frac{\mathcal{E}\left[X(\frac{k}{2^m})\right]^{n_{m,k+1}} \left[1-X\ (\frac{k}{2^m})\right]^{n_{m,k+1}}}{\mathcal{E}\left[X(\frac{k}{2^m})\right]^{n_{m,k}} \left[1-X\ (\frac{k}{2^m})\right]^{n_{m,k+1}}}$$

Proof.

Let $P^{G_n(t)}(F(t)\ A)$ denote the probability that F(t) is in A when F is determined by the independent $\{Z(\frac{k}{2^m})\}$ and let $P(F(t)\in A)$ denote the probability that F(t) is in A when F is determined by interpolation with the independent $\{X(\frac{k}{2^m})\}$. If

$$1)\qquad \int_B P^{G_n(t)}(F(t)\in A)\ d\ P = P(F(t)\in A,\ G_n(t)\in B)$$

for all sets B in $\sigma(G_n(t))$, then $P^{G_n(t)}$ will be the stated conditional probability.

It is sufficient to show that 1) holds for $A = \bigcap_{i=1}^{1} (F(t_i) \in J_i$ and $B = \bigcap_{i=1}^{1'} (G_n(t_i') \in J_i')$ where $J_i$ and $J_i'$ are subsets of the unit interval. Because the processes F(t) and $G_n(t)$ are, with probability one, determined by their values on the dyadic rationals, it will be sufficient to allow the $t_i$ and $t_i'$ bo be of the form $\frac{k}{2^m}$ where $2^m$ is their least common denominator. Hence, it is sufficient to prove that 1) holds if A is measurable with respect to $\sigma(F(\frac{1}{2^m}),\ F(\frac{2}{2^m}) - F(\frac{1}{2^m}),\ldots,$ $1 - F(\frac{2^{m-I}}{2^m}))$ and B is measurable with respect to $\sigma(n_{m,1},\ldots,n_{m,2^m})$. In this case, $P(F(t)\in A,\ G(t)\in B)$ is the integral over A x B of

$$\left\{ \prod_{\substack{i=1,\dots,m \\ j=1,3,\dots,2^m-1}} P_{X(\frac{j}{2^i})} \right\} K(n_{m,1}, \dots, n_{m,2^m}) \prod_{\substack{i=1,\dots,m \\ j=1,3,\dots,2^m-1}} \left[X(\frac{j}{2^i})\right]^{n_{ij}} \left[1-X(\frac{j}{2^i})\right]^{n_{i,j+1}} ..$$

Because the $X(\frac{j}{2^i})$ are independent and $n_{j,i} + n_{j,i+1} = n_{j-1,i+1}$, i odd, the marginal probability of $(n_{m,1}, \dots, n_{m,2^m})$ is $K(n_{m,1}, \dots, n_{m,2^m})$ times the products of the expectations in the denominator of

$$\prod_{\substack{i=1,\dots,m \\ j=1,3,\dots,2^m-1}} P_{Z(\frac{j}{2^i})} \qquad \text{Q. E. D.}$$

A somewhat different way to describe priors for distribution functions was given by Dubins and Freedman [2]. Their way involves interpolation with random variables $\left[X(\frac{k}{2^m}), F(X(\frac{k}{2^m}))\right]$ The above formula has an interpretation for this interpolation if the $n_{mj}$ are the numbers of observations betwenn $X(\frac{j}{2^m})$ and $X(\frac{j+1}{2^m})$. However, the conditional distribution so obtained is not that of F given the observation since the $n_{mj}$ are now functions of nature ' s strategy.

## 3. THE SUPPORT OF THE PRIOR.

Suppose that the support of $P_{X(\frac{k}{2^m})}$ is all of $[0,1]$ and P (F is continuous) = 1. Then the support of the distribution of F is the space of all distribution functions with respect to the topology of weak convergence and containe the continuous distribution functions with respect to the topology point-wise convergence. These facts are immediate upon noting that the map of the product of the coordinate spaces of the variables $X(\frac{k}{2^m})$ into the space of distribution functions, which is obtained by regarding the points of the coordinate spaces as degenerate random variables, is continuous with respect to the point-wise convergence in both spaces when the map is restricted to the continuous distribution functions.

Métivier [6], has shown another result, namely that, if the support of each $X(\frac{k}{2^m})$ is the closed unit interval, then the support, with respect to weak convergence, of the prior defined by interpolation with the $\{X(\frac{k}{2^m})\}$ is the space of all distribution functions.

## 4. ACKNOWLEDGEMENT.

The author wishes to thank Anatole Joffe, Michel Métivier, and Constance Van Eeden for helpful discussions about the suject of this paper.

## REFERENCES

[1] Charles Edward ANTONIAK :

"Mixtures of Dirichlet Processes with applications to Bayesian non parametric problems".

Abstract of dissertation, University of California. LOS ANGELES 1969 (manuscript of 56 pp.)

[2] LESTER E., DUBINS and David A. FREEDMAN :

"Random Distribution Functions"

Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. III, pp. 183-214.

[3] Thomas S. FERGUSON :

"A Bayesian analysis of some non parametric problems"

(manuscrit of 43 pp.)

[4] Charles H. KRAFT :

"A class of distribution function processes which have derivatives"

J. Appl. Probability, Vol. 1 (1964) pp. 385-388.

[5] Charles H. KRAFT, and Constance VAN EEDEN :

"Bayesian Bio - Assay"

Ann. Math. Stat. Vol. 35 (1964), pp. 886-890.

[6] Michel METIVIER :

(manuscript of 8 pp.). To be published.