PROPERTIES OF SAMPLES FROM DISTRIBUTIONS CHOSEN FROM A DIRICHLET PROCESS

Yamato, Hajime
Department of Mathematics, Faculty of Science, Kagoshima University

https://doi.org/10.5109/13358
PROPERTIES OF SAMPLES FROM DISTRIBUTIONS
CHosen FROM A DIRICHLET PROCESS*

By

Hajime YAMATO**

Abstract

The joint distributions of samples from distributions chosen from a Dirichlet process with nonatomic parameter are given and the conditional distributions of the samples are derived, by the method different from Yamato [4]. By making use of the above result, the expectations of functions of the samples are evaluated.

1. Introduction

The Dirichlet process was introduced by Ferguson [2] for Bayesian nonparametric inference. It is well-known that a distribution chosen from a Dirichlet process is discrete with probability one. The purpose of this paper is to show properties of samples from distributions chosen from a Dirichlet process with nonatomic parameter by the method different from Yamato [4] and to give its application. The author assumes that readers are familiar with the Dirichlet process. For the definition of the Dirichlet process, see Ferguson [2].

Let $R$ be the real line and let $B$ be the $\sigma$-field of Borel sets. Let $\alpha$ be a nonnull finite measure on $(R, B)$. $Q(\cdot)$ denotes $\alpha(\cdot)/\alpha(R)$ and $M$ denotes $\alpha(R)$. We list some properties of the Dirichlet process for the later use.

LEMMA 1 (Ferguson [2]). Let $P$ be a Dirichlet process on $(R, B)$ with parameter $\alpha$ and let $X$ be a sample of size 1 from $P$. Then for $A \in B$

$$P(X \in A) = Q(A).$$

Let $X_1, \ldots, X_n$ be a sample of size $n$ from a distribution $P$ chosen from a Dirichlet process on $(R, B)$ with parameter $\alpha$. Then, as stated in Korwar and Hollander [3], we can view the observations $X_1, \ldots, X_n$ as being obtained sequentially as follows: Let $X_1$ be a sample of size 1 from $P$; having obtained $X_1$, let $X_2$ be a sample of size 1 from the conditional distribution $P$ given $X_1$; and so on until $X_1, \ldots, X_n$ are obtained. Thus by Lemma 1 we have the following lemma, which is essentially similar to the statement of Zehnwirth [5, p. 16].

LEMMA 2. Let $P$ be a Dirichlet process on $(R, B)$ with parameter $\alpha$ and let $X_1, \ldots, X_n$ be a sample of size $n$ from $P$. Then we can view $X_1$ has the distribution $Q$

---

* This research was partly supported by the Grant-in-Aid for Scientific Research Project No. 58540117 from the Ministry of Education.

** Department of Mathematics, Faculty of Science, Kagoshima University, Kagoshima 890, Japan.
and for \( k = 1, \ldots, n - 1 \) the conditional distribution \( X_{k+1} \) given \( X_1, \ldots, X_k \) is given by
\[
\left( MQ(\cdot) + \sum_{j=1}^k \delta_{X_j}(\cdot) \right) / (M+k),
\]
where for \( x \in X \), \( \delta_x \) denotes the measure on \((R, B)\) giving the mass one to the point \( x \).

In Section 2, we shall give the joint distribution of samples from distributions chosen from a Dirichlet process with nonatomic parameter, by the method different from Yamato [4]. Furthermore, we shall derive the conditional distribution of the samples, which is essentially similar to Theorem 3.1 of Yamato [4].

We shall use the above result to evaluate expectation of functions of the samples for nonatomic parameter in Section 3.

2. Properties of Samples

Let \( R \) be the real line and let \( B \) be the \( \sigma \)-field of Borel sets. Let \( \alpha \) be a nonnull finite measure on \((R, B)\) and nonatomic. \( Q(\cdot) \) denotes \( \alpha(\cdot)/\alpha(R) \) and \( M \) denotes \( \alpha(R) \).

Let \( X_1, \ldots, X_n \) be a sample of size \( n \) from a distribution \( P \) chosen from a Dirichlet process on \((R, B)\) with parameter \( \alpha \). We can consider that the sample \( X_1, \ldots, X_n \) is obtained sequentially, as stated in Section 1. For nonnegative integers \( m(1), \ldots, m(n) \) with \( \sum_{i=1}^n m(i) = n \), let \( (X_1, X_2, \ldots, X_n) \in C(m(1), \ldots, m(n)) \) be the event that there are \( m(1) \) distinct values of \( X \) that occur only once, \( m(2) \) that occur exactly twice, \( \ldots, m(n) \) that occur exactly \( n \) times. We denote the sample \((X_1, \ldots, X_n)\) with \((X_1, \ldots, X_n) \in C(m(1), \ldots, m(n))\) by \((X_{11}, \ldots, X_{1m(1)}, X_{21}, X_{22}, \ldots, X_{2m(2)}, X_{31}, \ldots, X_{n1}, \ldots, X_{n1})\). Note that if \( m(n) \geq 1 \) then \( m(1) = \ldots = m(n-1) = 0 \) and \( m(n) = 1 \). If \( m(1) = 2 \) and \( X_s \neq X_t \) with \( s < t \) are different from the remainders, then \( X_1 = X_s, X_2 = X_t \). Suppose that \( m(j) = m(1 < j < m) \).

For any \( A_{ij} \in B(i=1, \ldots, n, j=1, \ldots, m(i)) \),
\[
P(X_{ij} \in A_{ij}, \ldots, X_n \in C(m(1), \ldots, m(n)))
= n! M^{r(n)} \prod_{i=1}^n \prod_{j=1}^{m(i)} Q(A_{ij}) / M^{(n)} \prod_{i=1}^n (m(i) !)^{m(i)},
(2.1)
\]
where \( M^{(n)} = M(M+1) \ldots (M+n-1) \).

Before proving Lemma 3 we shall prepare Lemma 4. For nonnegative integers \( m(1), \ldots, m(n) \) with \( \sum_{i=1}^n m(i) = n \), let \((X_n, X_{n-1}, \ldots, X_1) \in C_\circ(m(1), \ldots, m(n))\) be the event that \( X_n, X_{n-1}, \ldots, X_{n-(m(1)-1)} \), in that order, are unique in the sample and occur only once; that \( X_{n-(m(1)-1)}, \ldots, X_{n-(m(1)+2m(2)-1)} \) occur twice each in the order \( X_{n-(m(1)+2m(2)-1)}, \ldots, X_{n-(m(1)+2m(2)-1)} \) and etc. We use the similar notations to Antoniak [1] with respect to \( C_\circ \) and \( C_\circ \). We denote the sample \((X_n, X_{n-1}, \ldots, X_1) \in C_\circ(m(1), \ldots, m(n))\) by \( Y_{11}, \ldots, Y_{1m(1)}, Y_{21}, Y_{22}, \ldots, Y_{2m(2)}, Y_{31}, \ldots, Y_{m(1)+2m(2)-1m(2)}, \ldots \). Similarly we denote the realization of the above sample, \( x_n, x_{n-1}, \ldots, x_1 \), by \( y_{11}, \ldots, y_{1m(1)}, y_{21}, y_{22}, \ldots, y_{2m(2)}, y_{31}, \ldots, y_{m(1)+2m(2)-1m(2)}, \ldots \). Then we have the following
Lemma 4. For any $A_{ij} \in B(i=1, \ldots, n, j=1, \ldots, m(i))$

$$P(Y_{ij} \in A_{ij}(i=1, \ldots, n, j=1, \ldots, m(i)), (X_n, \ldots, X_1) \in S_0(m(1), \ldots, m(n)))$$

$$= \prod_{i=1}^{n}((i-1)! M)^{m(i)} \prod_{i=1}^{n} \prod_{j=1}^{m(i)} Q(A_{ij}) / M^{i(n)} . \quad (2.2)$$

Proof. At first we shall prove the lemma for $n=2$. Two non-negative integers $(m(1), m(2))$ with $m(1)+2m(2)=2$ are $(2, 0)$ and $(0, 1)$. Let $X_1, X_2$ be a sample of size 2.

For $(X_2, X_1) \in S_0(m(1), m(2))$ with $m(1)=2$ and $m(2)=0$, we have $Y_{11}=X_1, Y_{12}=X_2$. For any $A_1, A_2 \in B$, from Lemma 2 we have

$$P(Y_{11} \in A_2, Y_{12} \in A_1, (X_2, X_1) \in S_0(2, 0))$$

$$= P(X_2 \in A_2, X_1 \in A_1, X_2 \neq X_1)$$

$$= \int_{A_1} \left( \int_{A_2} P(X_2 \in A_2, X_2 \neq x_1| x_1) dQ(x_1) \right).$$

Since from Lemma 2, given $X_1=x_1$, $X_2$ has the distribution $(\alpha(\cdot)+\delta_{x_1}(\cdot))/(M+1)$ and $\alpha$ is nonatomic, we have

$$P(Y_{11} \in A_2, Y_{12} \in A_1, (X_2, X_1) \in S_0(2, 0))$$

$$= \int_{A_1} \alpha(A_2)/(M+1) dQ(x_1) = Q(A_1)\alpha(A_2)/(M+1)$$

$$= M^{m(1)} Q(A_1) Q(A_2) / M^{(2)}$$ with $m(1)=2, m(2)=0$.

For $(X_2, X_1) \in S_0(m(1), m(2))$ with $m(1)=0$ and $m(2)=1$, we have $Y_{21}=X_2=X_1$. For any $A \in B$, from Lemma 2 we have

$$P(Y_{21} \in A, (X_2, X_1) \in S_0(0, 1))$$

$$= P(X_2=x_1 \in A) = \int_{A} P(X_2=x_1| x_1) dQ(x_1) = \int_{A} 1/(M+1) dQ(x_1)$$

$$= M^{m(1)} Q(A) / M^{(2)}$$ with $m(1)=0, m(2)=1$.

Thus the lemma holds for $n=2$. Next we assume that the lemma holds for $n \geq 2$ and show that it holds for $n+1$. We denote the sample $X_{n+1}, X_n, \ldots, X_1$ with $(X_{n+1}, X_n, \ldots, X_1) \in S_0(m(1), \ldots, m(n)+1)$ and $\sum_{i=1}^{n+1} m'(i)=n+1$ by $Y_{11}, \ldots, Y_{1m'(1)}, Y_{21}, Y_{22}, \ldots, Y_{2m'(2)}, Y_{n+1}$. For a sample of size $n+1$ we have two cases: The one is that $X_{n+1}$ occurs only once and the other is that $X_{n+1}$ equals to the previous observation.

For the case that $X_{n+1}$ occurs only once, we have $m'(1) \geq 1, m'(n+1)=0$ and for $A_{ij} \in B(i=1, \ldots, n, j=1, \ldots, m(i))$

$$p_1 = P(Y_{ij} \in A_{ij}(i=1, \ldots, n, j=1, \ldots, m(i)), (X_{n+1}, \ldots, X_1) \in S_0(m(1), \ldots, m(n)+1))$$

$$= \int_{D_1} P(X_{n+1} \in A_{11}, X_{n+1} \neq x_1, \ldots, x_n| x_1, \ldots, x_n) dH(x_1, \ldots, x_n),$$

where $H(x_1, \ldots, x_n)$ is the joint distribution of $X_1, \ldots, X_n$ and

$$D_1 = \{(x_1, \ldots, x_n) | (x_n, \ldots, x_1) \in S_0(m(1), \ldots, m(n)), m(1)=m'(1)-1,$$ $m'(i)(i=2, \ldots, n), y_{1, j-1} \in A_{ij}(j=2, \ldots, m'(1)),$
\[
y_{ij} \in A_{ij}(i=2, \ldots, n, j=1, \ldots, m'(i)).
\]

Since from Lemma 2, given \( X_1, \ldots, X_n, X_{n+1} \) has the distribution \( \left( \alpha(\cdot) + \sum_{i=1}^{\infty} \delta_{x_i}(\cdot) \right) / (M+n) \) and \( \alpha \) is nonatomic,
\[
p_1 = \int_{D_1} \alpha(A_{11}) / (M+n) dH(x_1, \ldots, x_n)
= [\alpha(A_{11}) / (M+n)] P(X_1, \ldots, x_n) \in D_1
= [\alpha(A_{11}) / (M+n)] P(Y_{i,j-1} \in A_{ij}(j=2, \ldots, m'(1)),
Y_{i,j} \in A_{ij}(i=2, \ldots, n, j=1, \ldots, m'(i)),
(X_n, \ldots, X_1) \in \mathcal{C}_0(m'(1)-1, m'(2), \ldots, m'(n))).
\]

Since we assume that the lemma holds for \( n \) and \( m'(n+1)=0 \),
\[
p_1 = [\alpha(A_{11}) / (M+n)] M^{m'(1)} / \prod_{i=2}^{n} ((i-1)! M)^{m'(i)}
\times \prod_{j=2}^{m'(1)} Q(A_{ij}) / M^{m'(1)}
\times \prod_{i=2}^{m'(i)} Q(A_{ij}) / M^{m'(1)}
= \prod_{i=1}^{n+1} ((i-1)! M)^{m'(i)} / \prod_{i=1}^{n+1} \prod_{j=1}^{m'(i)} Q(A_{ij}) / M^{m'(i)}.
\]

(2.3)

For the case that \( X_{n+1} \) equals to the previous observation, at first we consider the case of \( m'(n+1)=1 \) and next the case of \( m'(n+1)=0 \). In case of \( m'(n+1)=1 \) where \( X_1, \ldots, X_{n+1} \) are all equal, for \( A_{n+1,1} \in \mathcal{B} \) we have
\[
p_2 = P(Y_{n+1,1} \in A_{n+1,1}, (X_{n+1}, \ldots, X_1) \in \mathcal{C}_0(m'(1), \ldots, m'(n+1)), m'(n+1)=1)
= \int_{D_2} P(X_{n+1}=x_n | x_1, \ldots, x_n) dH(x_1, \ldots, x_n),
\]
where \( D_2 = \{(x_1, \ldots, x_n) | x_1 = \cdots = x_n \in A_{n+1,1}\} \). Since from Lemma 2 given \( X_1, \ldots, X_n, X_{n+1} \) has the distribution \( \left( \alpha(\cdot) + \sum_{i=1}^{\infty} \delta_{x_i}(\cdot) \right) / (M+n) \) and \( \alpha \) is nonatomic,
\[
p_2 = \int_{D_2} n / (M+n) dH(x_1, \ldots, x_n)
= [n / (M+n)] P(X_n = \cdots = X_1 \in A_{n+1,1})
= [n / (M+n)] P(Y_{n+1} \in A_{n+1,1}, (X_{n+1}, \ldots, X_1) \in \mathcal{C}_0(m(1), \ldots, m(n)), m(n)=1).
\]

We assume that the lemma holds for \( n \) and therefore
\[
p_2 = [n / (M+n)] (n-1)! M Q(A_{n+1,1}) / M^{m'(n)}
= (n! M)^{m'(n+1)} Q(A_{n+1,1}) / M^{m'(n+1)} \quad \text{with } m'(n+1)=1.
\]

(2.4)

Finally we consider the case that \( X_{n+1} \) equals to the previous observation and \( m'(n+1)=0 \). Since \( m'(1)=0 \), we suppose that there exists an integer \( k \) such that \( 2 \leq k \leq n, m'(1)=\cdots=m'(k-1)=0, m'(k) \geq 1 \) and \( m'(n+1)=0 \).

For \( A_{ij} \in \mathcal{B}(i=k, \ldots, n, j=1, \ldots, m'(i)) \), we have...
Properties of samples from distributions chosen from a Dirichlet process

\[ p_3 = P(Y_{ij} \in A_{ij}(i = k, \ldots, n, j = 1, \ldots, m'(i)), (X_{n+1}, \ldots, X_i) \in C_0(m'(1), \ldots, m'(n+1))) \]

\[ = \int_{D_3} P(X_{n+1} = x_n = \ldots = x_{n-k+1} | x_1, \ldots, x_n) dH(x_1, \ldots, x_n), \]

where

\[ D_3 = \{(x_1, \ldots, x_n) | (x_n, \ldots, x_1) \in C_0(m(1), \ldots, m(n)), m(i) = 0 \}

\[ (i = 1, \ldots, k - 2), m(k - 1) = 1, m(k) = m'(k) - 1 \]

\[ m(i) = m'(i)(i = k + 1, \ldots, n), y_{k-1}, i \in A_{k1} \]

\[ y_{k, j-1} \in A_{kj}(j = 2, \ldots, m'(k)), y_{ij} \in A_{ij}(i = k + 1, \ldots, n, j = 1, \ldots, m'(i)) \].

By the similar argument to \ref{132}, we have

\[ p_3 = \frac{1}{D_3(n)} \left( \frac{(k - 1)}{(M + n)} \right) (X_1, \ldots, X_n) \in D_3 \]

\[ = \left( \frac{(k - 1)}{(M + n)} \right) (X_1, \ldots, X_n) \in D_3 \]

\[ = \left( \frac{(k - 1)}{(M + n)} \right) P(Y_{k-1, 1} \in A_{k1}, Y_{k, j-1} \in A_{kj}(j = 2, \ldots, m'(k)), \]

\[ Y_{ij} \in A_{ij}(i = k + 1, \ldots, n, j = 1, \ldots, m'(i)), \]

\[ (X_n, \ldots, X_1) \in C_0(0, \ldots, 0, 1, m'(k) - 1, m'(k + 1), \ldots, m'(n)) \]

\[ = \left( \frac{(k - 1)}{(M + n)} \right) \prod_{i=k+1}^{n} (i - 1) ! M^{m'(i) - 1} \]

\[ \prod_{i=k+1}^{n} \prod_{j=1}^{m'(i)} Q(A_{ij}) / M^{m'(i)} \]

\[ = \frac{n}{D_3(n)} \left( \frac{(k - 2)}{(M + n)} \right) \prod_{i=k+1}^{n} \prod_{j=1}^{m'(i)} Q(A_{ij}) / M^{m'(i)} \]

\[ = \frac{n}{D_3(n)} \left( \frac{(k - 2)}{(M + n)} \right) \prod_{i=k+1}^{n} \prod_{j=1}^{m'(i)} Q(A_{ij}) / M^{m'(i)} \]

\[ = \frac{n}{D_3(n)} \left( \frac{(k - 2)}{(M + n)} \right) \prod_{i=k+1}^{n} \prod_{j=1}^{m'(i)} Q(A_{ij}) / M^{m'(i)} \]

\[ = \frac{n}{D_3(n)} \left( \frac{(k - 2)}{(M + n)} \right) \prod_{i=k+1}^{n} \prod_{j=1}^{m'(i)} Q(A_{ij}) / M^{m'(i)} \]

\[ = \frac{n}{D_3(n)} \left( \frac{(k - 2)}{(M + n)} \right) \prod_{i=k+1}^{n} \prod_{j=1}^{m'(i)} Q(A_{ij}) / M^{m'(i)} \]

\[ = \frac{n}{D_3(n)} \left( \frac{(k - 1)}{(M + n)} \right) \prod_{i=k+1}^{n} \prod_{j=1}^{m'(i)} Q(A_{ij}) / M^{m'(i)} \]

\[ = \frac{n}{D_3(n)} \left( \frac{(k - 1)}{(M + n)} \right) \prod_{i=k+1}^{n} \prod_{j=1}^{m'(i)} Q(A_{ij}) / M^{m'(i)} \]

\[ = \frac{n}{D_3(n)} \left( \frac{(k - 1)}{(M + n)} \right) \prod_{i=k+1}^{n} \prod_{j=1}^{m'(i)} Q(A_{ij}) / M^{m'(i)} \]

From the evaluations of \( p_1, p_2, p_3 \), we know that the lemma holds for \( n + 1 \) and thus proved it by induction.

**Proof of Lemma 3.** Lemma 4 also holds for \( (X_1, \ldots, X_n) \in C_0(m(1), \ldots, m(n)) \).

The number of ways that \( n \) observations \( X_1, \ldots, X_n \) are permuted differently with \( (X_1, \ldots, X_n) \in C(m(1), \ldots, m(n)) \) and \( \sum_{i=1}^{n} m(i) = n \) is \( \frac{n!}{\prod_{i=1}^{n} [m(i)! (i!)^{m(i)}]} \). To multiply the right-hand side of (2.2) with \( (X_1, \ldots, X_n) \in C_0(m(1), \ldots, m(n)) \) by this number yields (2.1).

If we take \( A_{ij} = B \) for \( i = 1, \ldots, n, j = 1, \ldots, m(i) \) in Lemma 3, then we have the following lemma which is found in Antoniak [1].

**Lemma 5.** (Antoniak [1]).

\[ P((X_1, \ldots, X_n) \in C(m(1), \ldots, m(n))) = n! M^{\sum_{i=1}^{n} m(i) / M(m(i))} \]

The following theorem is essentially similar to Theorem 3.1 of Yamato [4].

**Theorem 1.** Given \( (X_1, \ldots, X_n) \in C(m(1), \ldots, m(n)), X_{11}, X_{12}, \ldots, X_{1m(1)}, X_{21}, X_{22}, \ldots, X_{2m(2)}, \ldots, X_{nm(n)} \) are independent and identically distributed with the distribution \( Q \).

**Proof.** For any \( A_{ij} \in B(i = 1, \ldots, n, j = 1, \ldots, m(i)) \), by Lemma 3 and 5 we have
3. Expectation of Random Functionals

By the use of Theorem 1 we shall prove the following theorem (Yamato [4]) for nonatomic parameter $\alpha$. Our method of proof is different from Yamato [4]. $R^n$ is the $n$-dimensional Euclidean space and $B^n$ is the $\sigma$-field of Borel subsets of $R^n$ for $n=2, 3, \ldots$.

**Theorem 2 (Yamato [4]):** Let $h(x_1, \ldots, x_n)$ be a real-valued measurable function defined on $(R^n, B^n)$ and symmetric in $x_1, \ldots, x_n$. Let $P$ be a Dirichlet process on $(R, B)$ with parameter $\alpha$. Let $X_1, \ldots, X_n$ be a sample from $P$. Then

\[
E[h(X_1, \ldots, X_n)] = \sum_{m(1), \ldots, m(n)} \left( \frac{n!}{M(m(1)) \cdots M(m(n))} \frac{m(i)!}{m(i)} \right) \prod_{i=1}^{n} \frac{dQ(x_{ij})}{m(i)}
\]

provided all integrals of the right-hand side exist. Where $\sum^*$ denotes the summation over all $n$ nonnegative integers $m(1), \ldots, m(n)$ satisfying $\sum_{i=1}^{n} m(i) = n$ and in the arguments of the integrand of the right-hand side $x_{is}$ appears at exactly $i$ times for $i=1, 2, \ldots, n$ and $s=1, \ldots, m(i)$.

**Proof.** We give the proof for nonatomic parameter $\alpha$. From Theorem 1, for nonnegative intergers $m(1), \ldots, m(n)$ with $\sum_{i=1}^{n} m(i) = n$, given $(X_1, \ldots, X_n) \in C(m(1), \ldots, m(n))$, $X_1, \ldots, X_{1m(1)}, X_{21}, \ldots, X_{2m(2)}, \ldots, X_{nm(n)}$ are independent and identically distributed with the distribution $Q$. $h$ is symmetric in $x_1, \ldots, x_n$. Therefore we have

\[
E[h(X_1, \ldots, X_n) | (X_1, \ldots, X_n) \in C(m(1), \ldots, m(n))] = \sum_{m(1), \ldots, m(n)} \left( \frac{n!}{M(m(1)) \cdots M(m(n))} \frac{m(i)!}{m(i)} \right) \prod_{i=1}^{n} \frac{dQ(x_{ij})}{m(i)}
\]

which exists for each $n$ nonnegative integers $m(1), \ldots, m(n)$ with $\sum_{i=1}^{n} m(i) = n$ by the assumption. Since by Lemma 5 for each $n$ nonnegative integers $m(1), \ldots, m(n)$ with
\[ \sum_{i=1}^{n} \ln(i) = n, \]

\[ P( (X_1, \ldots, X_n) \in \mathcal{C}(m(1), \ldots, m(n)) ) = n! \frac{M^{\sum_i m(i)}}{M^{\mathcal{C}(n)}} \prod_{i=1}^{n} (m(i)! i^{m(i)}), \]

taking expectation of (3.2) we have (3.1).

References


Communicated by S. Kano
Received September 21, 1983