A note on a proper Bayesian bootstrap

Pietro Muliere* Piercesare Secchi

Dipartimento di Economia Politica e Metodi Quantitativi
Università di Pavia

Abstract

We introduce a random probability distribution which approximates, in the sense of weak convergence, the Dirichlet process and supports a Bayesian resampling plan named proper Bayesian bootstrap

Keywords: Dirichlet process, Bayesian bootstrap.
AMS 1991 Subject Classification: 62G09, 60B10.

*Dip. Econ. Pol. Met. Quant., Via San Felice 5, I-27100 Pavia, Italy. e-mail: pmuliere@eco.unipv.it

from P_0. Set $P_m^* \in \mathcal{P}$ to be the empirical distribution of $X_1^*, \ldots, X_m^*$ defined by

$$P_m^* = \frac{1}{m} \sum_{i=1}^{m} \delta_{X_i^*}$$

where δ_x indicates the point mass at x. Write $\mathcal{H}_m^*$ for the distribution of P_m^* on $(\mathcal{P}, \sigma(\mathcal{P}))$.

Roughly, the following definition introduces a process P such that, conditionally on P_m^*, $P \in \mathcal{D}(wP_m^*)$.

Definition 1 *A random element $P \in \mathcal{P}$ is called a Dirichlet-Multinomial process with parameters (m, w, P_0) ($P \in \mathcal{DM}(m, w, P_0)$) if it is a mixture of Dirichlet processes on $(\Re, \mathcal{B})$ with mixing distribution $\mathcal{H}_m^*$ and transition measure α_w.*

It follows from the definition that, if $P \in \mathcal{DM}(m, w, P_0)$, then, for every finite measurable partition $B_1, \ldots, B_k$ of $\Re$ and $(y_1, \ldots, y_k) \in \Re^k$,

$$\Pr\left(P(B_1) \le y_1, \ldots, P(B_k) \le y_k\right) = \int_{\mathcal{P}} D(y_1, \ldots, y_k | \alpha_w(u, B_1), \ldots, \alpha_w(u, B_k))\, d\mathcal{H}_m^*(u)$$

where $D(y_1, \ldots, y_k | \alpha_1, \ldots, \alpha_k)$ denotes the distribution function of the Dirichlet distribution with parameters $(\alpha_1, \ldots, \alpha_k)$. With different notation, we may say that the vector $(P(B_1), \ldots, P(B_k))$ has distribution

$$\text{Dirichlet}(w\frac{M_1}{m}, \ldots, w\frac{M_k}{m}) \bigwedge_{(M_1, \ldots, M_k)} \text{Multinomial}\,(m, (P_0(B_1), \ldots, P_0(B_k)))\,.$$

For our purposes, the introduction of the Dirichlet-Multinomial process is justified by the following theorem.

Theorem 1 *For every $m > 0$, let $P_m \in \mathcal{P}$ be a Dirichlet-Multinomial process with parameters (m, w, P_0). Then, when $m \to \infty$, P_m weakly converges to a Dirichlet process with parameter wP_0.*

Proof. Given any finite measurable partition $B_1, \ldots, B_k$ of $\Re$, the distribution of the vector $(P_m(B_1), \ldots, P_m(B_k))$ weakly converges to a Dirichlet distribution with parameters $(wP_0(B_1), \ldots, wP_0(B_k))$ when $m \to \infty$. We prove this claim by showing that the moments of any order converge to the corresponding moments of the right Dirichlet distribution. In fact, if $r_1 \ge 0, \ldots, r_k \ge 0$ are k integers,

$$E\left[P_m^{r_1}(B_1) \cdots P_m^{r_k}(B_k)\right] = E\left[\frac{\Gamma(w)}{\Gamma(w\frac{M_1}{m}) \cdots \Gamma(w\frac{M_k}{m})} \frac{\Gamma(w\frac{M_1}{m} + r_1) \cdots \Gamma(w\frac{M_k}{m} + r_k)}{\Gamma(w + \sum_{i=1}^k r_i)}\right]$$

where $(M_1, \ldots, M_k)$ has distribution Multinomial$(m, (wP_0(B_1), \ldots, wP_0(B_k)))$. Therefore

$$\lim_{m\to\infty} E\left[P_m^{r_1}(B_1) \cdots P_m^{r_k}(B_k)\right] =$$
$$= \frac{\Gamma(w)}{\Gamma(wP_0(B_1)) \cdots \Gamma(wP_0(B_k))} \frac{\Gamma(wP_0(B_1) + r_1) \cdots \Gamma(wP_0(B_k) + r_k)}{\Gamma(w + \sum_{i=1}^k r_i)}$$

since, for $i = 1, \ldots, k$, $m^{-1}M_i$ converges in probability to $P_0(B_i)$.

In order to prove that P_m weakly converges to a Dirichlet process with parameter wP_0 it is now enough to show that the sequence of measures induced on $(\mathcal{P}, \sigma(\mathcal{P}))$ by the processes P_m, $m = 1, 2, \ldots$, is tight. We will follow an argument inspired by Sethuraman and Tiwari [1982].

Given $\epsilon > 0$, let K_r, $r = 1, 2, \ldots$, be a compact set of $\Re$ such that

$$P_0(K_r^c) \leq \frac{\epsilon}{r^3}$$

and define

$$M_r = \{P \in \mathcal{P} : P(K_r^c) \leq \frac{1}{r}\}.$$

The set

$$M = \bigcap_{r=1}^{\infty} M_r$$

is compact in $\mathcal{P}$.

Fix r and note that, for every m, the random variable $P_m(K_r^c)$ has distribution

$$\text{Beta}(w\frac{\Theta}{m}, w(1 - \frac{\Theta}{m})) \bigwedge_{\Theta} \text{Binomial}(m, P_0(K_r^c)).$$

Therefore $E[P_m(K_r^c)] = P_0(K_r^c)$ and this implies that

$$\Pr(P_m(K_r^c) > \frac{1}{r}) \leq rP_0(K_r^c) \leq \frac{\epsilon}{r^2}.$$

Hence, for every m,

$$\Pr(P_m \in M^c) \leq \sum_{r=1}^{\infty}(P_m(K_r^c) > \frac{1}{r}) \leq \epsilon \sum_{r=1}^{\infty} \frac{1}{r^2}$$

and this proves that the sequence of measures induced on $(\mathcal{P}, \sigma(\mathcal{P}))$ by the processes P_m, $m = 1, 2, \ldots$, is tight. $\diamond$

Remark 1 We called the process P defined above Dirichlet-Multinomial since, given any finite measurable partition $B_1, \ldots, B_k$ of $\Re$, the distribution of $(P(B_1), \ldots, P(B_k))$ is a mixture of Dirichlet distributions with Multinomial weights. This process must not be confused with the Dirichlet-Multinomial point process of Lo [Lo, 1986, Lo, 1988] whose marginal distributions are mixtures of Multinomial with Dirichlet weights.

3 Connections with the proper Bayesian bootstrap

Let $T : \mathcal{P} \to \Re$ be a measurable function and $P \in \mathcal{D}(wP_0)$ with $w > 0, P_0 \in \mathcal{P}$. It is often difficult to work out analitically the distribution of $T(P)$, even when T is a simple statistical functional like the mean [Hannum.et.al., 1981, Cifarelli and Regazzini, 1990]. However, when P_0 is discrete with finite support one may produce a reasonable approximation of the distribution of $T(P)$ by a Monte Carlo procedure which obtains i.i.d. samples from $\mathcal{D}(wP_0)$. If P_0 is not discrete, we propose to approximate the distribution of $T(P)$ with the distribution of $T(P_m)$, where P_m is a Dirichlet-Multinomial process with parameters (m, w, P_0) and m is large enough.

Of course, since the Continuous Mapping Theorem does not apply to every function T, the fact that P_m weakly converges to P does not always imply that the distribution of $T(P_m)$ is close to that of $T(P)$. However, in a previous work [Muliere and Secchi, 1996], we proved that this is in fact the case when T belongs to a large class of linear functionals or when T is a quantile. In the same paper we also proposed a bootstrap algorithm which produces an approximation of the distribution of $T(P)$ by means of the following steps:

(1) Generate an i.i.d sample $X_1^*, \ldots, X_m^*$ from P_0.

(2) Generate an i.i.d. sample $V_1, \ldots, V_m$ from a Gamma$(\frac{w}{m}, 1)$.

(3) Compute $T(P_m)$, where $P_m \in \mathcal{P}$ is defined by

$$P_m = \frac{1}{\sum_{i=1}^m V_i} \sum_{i=1}^m V_i \delta_{X_i^*}.$$

(4) Repeat steps (1)-(3) s times and approximate the distribution of $T(P)$ with the empirical distribution of the values $T_1, \ldots, T_s$ generated at step (3).

The performance of this algorithm was tested with a few numerical illustrations in Muliere and Secchi [1996] where it was compared with the approximations generated by the Pólya urn scheme [Blackwell and MacQueen, 1973] and with the Bayesian bootstrap procedures described by Rubin [1981] and by Meeden [1993].

It is easily seen that the probability distribution P_m produced at step (3) is in fact a trajectory of a Dirichlet-Multinomial process with parameters (m, w, P_0). We may therefore conclude that the previous algorithm aims at approximating the distribution of $T(P)$ with the distribution of $T(P_m)$, where $P_m \in \mathcal{DM}(m, w, P_0)$, and approximates the latter by means of the empirical distribution of the values $T_1, \ldots, T_s$ generated at step (3).

Remark 2 Step (1) is useless when P_0 is discrete with finite support $\{z_1, \ldots, z_m\}$ and $P_0(z_i) = p_i, i = 1, \ldots, m$, with $\sum_{i=1}^m p_i = 1$. In fact, in this case one may generate at step (3) a trajectory of $P \in \mathcal{D}(wP_0)$ by taking

$$P_m = \frac{1}{\sum_{i=1}^m V_i} \sum_{i=1}^m V_i \delta_{z_i}$$

where $V_1, \ldots, V_m$, are independent and V_i has distribution Gamma$(wp_i, 1)$, $i = 1, \ldots, m$.

We call the algorithm (1)-(4) proper Bayesian bootstrap. To understand the reason for this name consider the following situation. A sample $X_1, \ldots, X_n$ from a process $P \in \mathcal{D}(kQ_0)$, with $k > 0$ and $Q_0 \in \mathcal{P}$, has been observed and the problem is to compute the posterior distribution of $T(P)$ where T is a given statistical functional. Ferguson [1973] proved that the posterior distribution of P is again a Dirichlet process with parameter $kQ_0 + \sum_{i=1}^n \delta_{X_i}$.

In order to approximate the posterior distribution of $T(P)$ our algorithm generates an i.i.d. sample $X_1^*, \ldots, X_m^*$ from

$$\frac{k}{k+n}Q_0 + \frac{n}{k+n}\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}\right)$$

and then, in step (3), produces a trajectory of a process which, given $X_1^*, \ldots, X_m^*$, is Dirichlet with parameter $(k+n)m^{-1}\sum_{i=1}^{m}\delta_{X_i^*}$ and evaluates T with respect to this trajectory. The algorithm is therefore a bootstrap procedure since it samples from a mixture of the empirical distribution function generated by $X_1, \ldots, X_n$ and Q_0 which, toghether with the weight k, elicits the prior opinions relative to P. Because it takes into account prior opinions by means of a proper distribution function, the procedure was termed proper.

The name proper Bayesian bootstrap also distinguishes the algorithm from the Bayesian bootstrap of Rubin [1981] which approximates the posterior distribution of $T(P)$ by means of the distribution of $T(Q)$ whith $Q \in \mathcal{D}(\sum_{i=1}^{n}\delta_{X_i})$. We already noticed in a previous work [Muliere and Secchi, 1996] that there are no proper priors for P which support Rubin's approximation and that the proper Bayesian bootstrap essentially becomes the Bayesian bootstrap of Rubin when k is set to 0 or n is very large.

References

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

BLACKWELL, D. and J.B. MACQUEEN (1973), "Ferguson distributions via Pólya urn schemes". *Ann. Statist.*, 1(2), 353-355.

CIFARELLI, D.M. e REGAZZINI, E. (1990), "Distribution functions of means of Dirichlet process", *Ann. Statist.*, 18(1), 429–442.

FERGUSON, T.S. (1973), "A Bayesian analysis of some nonparametric problems", *Ann. Statist.*, 1(2), 209–230.

HANNUM, R.C., HOLLANDER, M. and N.A. LANGBERG (1981), "Distributional results for random functionals of a Dirichlet process", *Ann. Prob.*, 9, 665–670.

LO, A.Y. (1986), "Bayesian statistical inference for sampling a finite population", *Ann. Statist.*, 14(3), 1226–1233.

LO, A.Y. (1988), "A Bayesian bootstrap for a finite population", *Ann. Statist.*, 16(4), 1684–1695.

MEEDEN, G. (1993), "Noninformative nonparametric Bayesian estimation of quantiles", *Statistics and Probability Letters*, 16, 103–109.

MULIERE, P. e P. SECCHI (1996), "Bayesian nonparametric predictive inference and bootstrap techniques", *Ann. Inst. Statist. Math.*, to appear.

PROHOROV, YU. V. (1956), "Convergence of random processes and limit theorems in probability theory", *Theory Prob. Appl.*, 1, 157–214.

RUBIN, D.M. (1981), "The Bayesian bootstrap", *Ann. Statist.*, 9(1), 130–134.

SETHURAMAN, J. and R. C. TIWARI (1982), "Convergence of Dirichlet measures and the interpretation of their parameter", *Statistical Decision Theory and Related Topics III*, Vol.2, 305–315.