

On a Class of Discrete Nonparametric Prior Distributions

Su una Classe di Distribuzioni a Priori Non Parametriche Discrete

Andrea Ongaro
Dipartimento di Statistica
Università di Milano Bicocca
andrea.ongaro@unimib.it

Carla Cattaneo
Dipartimento di Statistica
Università di Milano Bicocca
carla.cattaneo@unimib.it

Riassunto: In questo lavoro viene proposta una classe molto flessibile e generale di distribuzioni a priori discrete non parametriche. Tale classe include, fra gli altri, il processo di Dirichlet, il processo Poisson-Dirichlet, il processo beta a due parametri, i processi di tipo “stick-breaking” e altre generalizzazioni del processo di Dirichlet. In particolare si è ricavata una espressione particolarmente semplice della forma della distribuzione a posteriori e della distribuzione predittiva. Inoltre si è derivata una regola esplicita di aggiornamento della distribuzione a posteriori, che può essere utilizzata a fini di simulazione.

Keywords: Random probability measure, generalized Dirichlet prior, Predictive distribution.

1. Introduction

In Bayesian inference problems, interest generally lies in the derivation of posterior and predictive distributions. In nonparametric settings, these are often neither easily available, nor easy to simulate.

The Dirichlet process plays a central role in Bayesian nonparametric statistics and various generalizations have been proposed and investigated in the literature. In this work we study a very rich and general class of nonparametric discrete prior distributions, denoted by $\mathcal{P}$, on an arbitrary sample space $(\mathcal{X}, \mathcal{A})$, which includes the Dirichlet process and many of its generalizations. The aim is to give insight into the structure of the posterior and predictive distributions and to provide an explicit updating mechanism of the posterior which can be employed to perform simulation. The problem of choosing a member of the proposed class on the basis of available prior information will only be touched upon.

In Section 2 we shall define the class $\mathcal{P}$ and in Section 3 we shall derive the main results concerning the predictive and posterior distribution. Comments on the results are summarized in Section 4.

2. Definition of the class $\mathcal{P}$

The class $\mathcal{P}$ is defined as follows. Let δ_ξ denote a probability measure degenerate at ξ and let P_0 be a diffuse probability measure on $(\mathcal{X}, \mathcal{A})$, i.e. $P_0(\{x\}) = 0, \forall x \in \mathcal{X}$. Then, a random discrete probability measure P belongs to $\mathcal{P}$ if it can be represented as

$$P = \sum_{i=1}^M \beta_i \delta_{\xi_i}.$$

Here, the ξ_i s, β_i s and M are random elements with distribution specified as follows. The random positions ξ_i s are i.i.d. from P_0 and independent of all the other random elements; conditionally on M , the random weights $\beta_1, \dots, \beta_M$ have an arbitrary distribution Q_M on the $(M - 1)$ -dimensional simplex $S^M = \{\mathbf{x} \in \mathbf{R}^M : x_1 + \dots + x_M = 1, x_i \geq 0, i = 1, \dots, M\}$; $M \geq 1$ is an arbitrary extended integer valued random variable, i.e. M is allowed to take the value $+\infty$.

This definition gives a very rich and flexible structure. Two properties essentially set $\mathcal{P}$ apart from the general class of discrete random probability measures: (1) the random positions ξ_i s are independent of the associated random weights; (2) the positions are generated by i.i.d. random elements from a diffuse measure. Notice, on the other hand, that no restrictions are imposed on the distribution of the weights.

Many relevant classes of nonparametric prior distributions proposed in the literature can be viewed as instances of the class $\mathcal{P}$ (see Ishwaran and James (2001) and references therein for an extensive list of such priors). Besides the Dirichlet process, we mention: the Poisson-Dirichlet process, the beta two-parameter process, stick-breaking priors and other generalizations of the Dirichlet process (Hjort and Ongaro (1999), Hjort (2000)).

3. Posterior and predictive distributions

It is not difficult to check that $\mathcal{P}$ does not form a conjugate class of prior distributions. One way to derive a general expression for the posterior distribution of members of $\mathcal{P}$ is to embed $\mathcal{P}$ into a larger conjugate class. In the next Lemma we establish the form of such enlarged class by showing how a member of this class is updated when a single new observation is taken. First, we need to introduce the following notation.

Given a probability measure Q_n on the simplex S^n , we denote by Q_n^l , $l = 1, \dots, n$, the probability measure absolutely continuous w.r.t. Q_n with Radon-Nikodym derivative $\frac{dQ_n^l}{dQ_n}(x_1, \dots, x_n) = \frac{x_l}{b_l}$ where $b_l = \int_{S^n} x_l dQ_n(x_1, \dots, x_n)$. If $b_l = 0$, we set $Q_n^l = Q_n$.

Moreover, throughout the paper we shall always assume that the random positions ξ_i s are i.i.d. from P_0 and independent of all the other random elements defining the random probability measures of interest.

Lemma. Consider the following random probability measure P_h on $(\mathcal{X}, \mathcal{A})$

$$P_h = \sum_{i=1}^h \gamma_i \delta_{y_i} + \sum_{i=1}^M \beta_i \delta_{\xi_i}$$

where $y_i \in \mathcal{X}$, $i = 1, \dots, h$, are fixed distinct constants, $h \geq 0$ is a fixed integer and $M \geq 0$ is an extended integer valued random variable with probability distribution $p(m)$. Moreover, let Q_{h+M} be the joint distribution on S^{h+M} of $\alpha|M$, where $\alpha = (\gamma_1, \dots, \gamma_h, \beta_1, \dots, \beta_M)$ and assume that $M + h \geq 1$ with probability 1.

Then, the posterior distribution of P_h given a new observation X has the following form.

(1) If $x = y_j$ for some $1 \leq j \leq h$, then $P_h|X = x$ is distributed as

$$\sum_{i=1}^h \gamma'_i \delta_{y_i} + \sum_{i=1}^{M'} \beta'_i \delta_{\xi_i}$$

with $\Pr\{M' = m\} = p(m)E(\gamma_j|M = m)/E(\gamma_j)$ and $(\alpha'|M' = m) \sim Q_{h+m}^j$.

(2) If $x \neq y_j$ for all $1 \leq j \leq h$, then $P_h|X = x$ is distributed as

$$\sum_{i=1}^h \gamma_i'' \delta_{y_i} + \gamma_{h+1}'' \delta_x + \sum_{i=1}^{M''} \beta_i'' \delta_{\xi_i},$$

with $\Pr\{M'' = m\} = p(m+1) \sum_{i=1}^{m+1} E(\beta_i|M = m+1)/E(\sum_{i=1}^M \beta_i)$ for $m \geq 0$. Conditionally on $M'' = m$, the random vector $\alpha'' = (\gamma_1'', \dots, \gamma_{h+1}'', \beta_1'', \dots, \beta_{M''}'')$ has a mixture-type distribution given by α_L^* , where L is an integer valued random variable such that $\Pr\{L = l\} = E(\beta_l|M = m+1)/\sum_{i=1}^{m+1} E(\beta_i|M = m+1)$ for $1 \leq l \leq m+1$ and

$$\alpha_L^*|L = l \sim (\lambda_1, \dots, \lambda_h, \lambda_{h+l}, \lambda_{h+1}, \dots, \lambda_{h+l-1}, \lambda_{h+l+1}, \dots, \lambda_{h+m+1}),$$

where $\lambda = (\lambda_1, \dots, \lambda_{h+m+1}) \sim Q_{h+m+1}^{h+l}$, i.e. $\alpha_L^*|L = l$ is obtained by moving the $(h+l)$ th component of the vector λ to position $h+1$.

Notice that $E(\gamma_j) = \Pr(X = y_j)$ can be assumed to be positive in the case $x = y_j$; the same holds for $E(\sum_{i=1}^M \beta_i) = \Pr(X \neq y_j, j = 1, \dots, h)$ in case $x \neq y_j, j = 1, \dots, h$.

By recursively applying the Lemma, it is possible to derive the general form of the posterior distribution of $P \in \mathcal{P}$ given n observations. To this purpose, we need to introduce some preliminary notation and definitions.

Let $\mathbf{x} = (x_1, \dots, x_n)$ be a vector, whose entries are given observations. If $\mathbf{x}$ contains k distinct values, let $\mathbf{N}(\mathbf{x}) = (n_1, \dots, n_k)$ be the vector of multiplicities, i.e. the numbers of repetitions of each distinct value, which we suppose to be ordered in non-decreasing fashion. Moreover, let $\mathbf{V}(\mathbf{x}) = (v_1, \dots, v_k)$ be the vector of corresponding distinct values; if some distinct values share the same multiplicities, they can be ordered arbitrarily.

Definition. We say that the two n -tuples of observations $\mathbf{x} = (x_1, \dots, x_n)$ and $\mathbf{y} = (y_1, \dots, y_n)$ are equivalent if their multiplicity vectors $\mathbf{N}(\mathbf{x}) = (n_1, \dots, n_k)$ and $\mathbf{N}(\mathbf{y}) = (n'_1, \dots, n'_h)$ are equal, i.e. $k = h$ and $n_i = n'_i$ for all i .

Theorem. Let $\mathbf{X} = (X_1, \dots, X_n)$ be n observations from $P \in \mathcal{P}$. The posterior process $P|\mathbf{X} = \mathbf{x}$ can be represented as

$$\sum_{i=1}^k \gamma_i^{\mathbf{x}} \delta_{v_i} + \sum_{i=1}^{M^{\mathbf{x}}} \beta_i^{\mathbf{x}} \delta_{\xi_i},$$

where the v_i s are the elements of $\mathbf{V}(\mathbf{x})$; conditionally on $M^{\mathbf{x}} = m$, the random vector $\alpha^{\mathbf{x}} = (\gamma_1^{\mathbf{x}}, \dots, \gamma_k^{\mathbf{x}}, \beta_1^{\mathbf{x}}, \dots, \beta_{M^{\mathbf{x}}}^{\mathbf{x}})$, takes values on $S^{k+M^{\mathbf{x}}}$ and $M^{\mathbf{x}} \geq 0$ is an extended integer valued random variable. Furthermore, the distribution of $(M^{\mathbf{x}}, \alpha^{\mathbf{x}})$ depends on $\mathbf{x}$ only through the set of multiplicities, i.e. if $\mathbf{x}$ and $\mathbf{y}$ are two equivalent vectors, then $(M^{\mathbf{x}}, \alpha^{\mathbf{x}}) \sim (M^{\mathbf{y}}, \alpha^{\mathbf{y}})$.

The posterior process displays a rather uncomplicated structure. The two main characteristics of the class $\mathcal{P}$, namely the independence of random positions and random weights, and the i.i.d. distributions of the random positions, are preserved by the posterior process. Furthermore, the distribution of the random weights exhibits a simple dependence on the observations.

The general form of the predictive distribution follows then immediately.

Corollary. If $\mathbf{X} = (X_1, \dots, X_n)$ are n observations from $P \in \mathcal{P}$, then $\forall A \in \mathcal{A}$

$$\Pr(X_{n+1} \in A | \mathbf{X} = \mathbf{x}) = \sum_{i=1}^k c_i^{\mathbf{x}} \delta_{v_i}(A) + (1 - c_1^{\mathbf{x}} - \dots - c_k^{\mathbf{x}}) P_0(A),$$

where the weights $c_i^{\mathbf{x}}$ depend on $\mathbf{x}$ only through the set of multiplicities.

In particular, this implies that $c_i^{\mathbf{x}} = c_j^{\mathbf{x}}$ if $n_i = n_j$. So, for example, if the n observations are all distinct, the predictive distribution takes the form:

$$\Pr(X_{n+1} \in A | \mathbf{X} = \mathbf{x}) = d_n \frac{1}{n} \sum_{i=1}^n \delta_{x_i}(A) + (1 - d_n) P_0(A),$$

i.e. it is a linear combination of P_0 and of the empirical distribution.

4. Conclusions

Despite the generality of the considered class, the posterior distribution of the process maintains quite a simple structure, which makes it easy to interpret and to work with. Theorem in Section 3 gives an easily interpretable description of this structure, whereas the Lemma gives an explicit rule of the updating mechanism of the posterior. The Lemma can also be recursively used to derive an explicit expression of the posterior and predictive distributions of any given new subclass of $\mathcal{P}$ of potential interest. Furthermore, the updating mechanism described in the Lemma can be easily implemented to perform simulation from the posterior distribution of the process whenever simulation procedures are available for the prior process.

The proposed class $\mathcal{P}$ reaches therefore a twofold objective: (1) to provide a unifying framework for a collection of widely used nonparametric priors; (2) to greatly broaden the set of discrete nonparametric priors effectively available for applications.

As far as the problem of choosing a member of the class $\mathcal{P}$ is concerned, we limit ourselves to the following remarks. Clearly, we have $E(P(A)) = \Pr(X_i \in A) = P_0(A)$, i.e. P_0 represents the mean of the random probability measure P as well as the marginal distribution of a single observation. The distribution of the weights β_i s and M can be selected by looking at further moments of P and by investigating the expression of the $c_i^{\mathbf{x}}$ s appearing in the predictive distribution. As an example, it is easy to see that $\text{Var}(P(A)) = P_0(A)(1 - P_0(A))E(\sum_{i=1}^M \beta_i^2)$. Therefore, the desired level of variability of P around P_0 can be achieved by appropriately specifying the second moments of the β_i s. Finally, we note that it is reasonable to let M be an unbounded random variable, possibly degenerated at $+\infty$; this prevents from having a fixed maximum number of possible distinct observations.

References

- Hjort N.L. (2000) Bayesian analysis for a generalised Dirichlet process prior, *submitted*.
Hjort N.L. and Ongaro A. (1999) Nonparametric Bayesian inference using a generalization of the Dirichlet process, *Working Papers n.39. Department of Statistics, University of Padova*.
Ishwaran H. and James L. (2001) Gibbs sampling methods for stick-breaking priors, *Journal of the American Statistical Association*, 96, 161–173.