Open Science Framework DOI:10.31219/osf.io/***

Dirichlet-process mixtures from the viewpoint of exchangeability

P.G.L. Porta Mana
Høgskulen på Vestlandet, Bergen <pgl@portamana.org>

Draft of 10 October 2020; updated 8 September 2021

***

Exchangeability comes into play when we want to make probabilistic inferences about an unbounded sequence of observations. It is the assumption or knowledge that the temporal or spatial arrangement of these observations is irrelevant for our inferences. Under exchangeability the long-run frequency distribution $f(x)\,\mathrm{d}x$ of the possible outcomes x becomes a sufficient statistic for our inferences:[1] the probability for the outcome of any observation is $\mathrm{p}(x \mid f, I)\,\mathrm{d}x = f(x)\,\mathrm{d}x$, where I denotes our prior information and assumptions.

If the long-run frequency distribution is unknown, the probability density of the outcomes of several observations is

$$\begin{aligned}\mathrm{p}(x_1, x_2, \dots \mid I) &= \int \Bigl[\prod_i \mathrm{p}(x_i \mid f, I)\Bigr]\, \mathrm{p}(f \mid I)\, \mathrm{d}f \\ &\equiv \int \Bigl[\prod_i f(x_i)\Bigr]\, \mathrm{p}(f \mid I)\, \mathrm{d}f\ .\end{aligned} \tag{1}$$

We can choose between different coordinate systems, or parameterizations, in the infinite-dimensional manifold[2] of frequency densities $\{f\,\mathrm{d}x\}$, over which the integration occurs; and we can choose different prior probability distributions $\mathrm{p}(f \mid I)\,\mathrm{d}f$ over such manifold. Often these two choices are made at once, because the prior probability density $\mathrm{p}(f \mid I)$ assumes a simple form in a specific coordinate system.

Dirichlet-process mixtures are a way of choosing a coordinate system and a prior probability density on the manifold of frequency densities.

[1] De Finetti 1930; 1937; Dawid 2013. [2] Choquet-Bruhat et al. 1996.

A generic density f is expressed as

$$f(x) = \sum_c q_c \; K(x \mid \theta_c) \;, \tag{2}$$

where (q_c) is a countable distribution and $K(x \mid \theta)$ is some density depending on parameters θ; for example it could be a normal density with mean and variance matrix $(\boldsymbol{m}, \boldsymbol{S}) =: \theta$. The sum is in principle infinite. The density f is thus represented by the coordinates $(\boldsymbol{q}, \boldsymbol{\theta}) \coloneqq (q_c, \theta_c)$.

The probability density for f is in turn re-expressed as a density for $(\boldsymbol{q}, \boldsymbol{\theta})$ having this form:

$$\mathrm{p}(f \mid I) \,\mathrm{d}f = \mathrm{Dir}(\boldsymbol{q} \mid \alpha) \prod_c G(\theta_c \mid I) \,\mathrm{d}\boldsymbol{q} \,\mathrm{d}\boldsymbol{\theta} \;, \tag{3}$$

where Dir is a Dirichlet distribution with parameter α and G is some density function. If the sum over c is infinite then the expression above tends to a Dirichlet process.

In terms of these coordinates, de Finetti's formula (1) becomes

$$\begin{aligned} \mathrm{p}(x_1, x_2, \dots \mid I) &= \int \prod_i \big[\mathrm{p}(x_i \mid \boldsymbol{q}, \boldsymbol{\theta}, I)\big] \; \mathrm{p}(\boldsymbol{q}, \boldsymbol{\theta} \mid I) \,\mathrm{d}\boldsymbol{q} \,\mathrm{d}\boldsymbol{\theta} \\ &\equiv \int \prod_i \Bigg[\sum_c q_c \; K(x_i \mid \theta_c)\Bigg] \, \mathrm{Dir}(\boldsymbol{q} \mid \alpha) \prod_c G(\theta_c \mid I) \,\mathrm{d}\boldsymbol{q} \,\mathrm{d}\boldsymbol{\theta} \;. \end{aligned} \tag{4}$$

If we consider $(\boldsymbol{q}, \boldsymbol{\theta})$ as quantities with unknown values, the their joint probability density function with the outcomes (x_i) is

$$\begin{aligned} \mathrm{p}(x_1, x_2, \dots, \; \boldsymbol{q}, \boldsymbol{\theta} \mid I) &= \prod_i \big[\mathrm{p}(x_i \mid \boldsymbol{q}, \boldsymbol{\theta}, I)\big] \; \mathrm{p}(\boldsymbol{q}, \boldsymbol{\theta} \mid I) \\ &\equiv \prod_i \Bigg[\sum_c q_c \; K(x_i \mid \theta_c)\Bigg] \, \mathrm{Dir}(\boldsymbol{q} \mid \alpha) \prod_c G(\theta_c \mid I) \;. \end{aligned} \tag{5}$$

We can imagine the existence of a discrete quantity C associated with x, such that the quantities $(x_1, C_1), (x_2, C_2), \dots$ have an exchangeable joint probability, and such that

$$\mathrm{P}(x \mid C, \boldsymbol{q}, \boldsymbol{\theta}, I) = K(x \mid \theta_C) \qquad \text{and} \qquad \mathrm{P}(C \mid \boldsymbol{q}, \boldsymbol{\theta}, I) = q_C \;. \tag{6}$$

However, C remains unobserved when x is observed. Then, rewriting the product of sums as sums over a product, $\prod_i \sum_c = \sum_{c_1} \sum_{c_2} \dots \prod_i$, we find

$$
\begin{aligned}
&\mathrm{p}(x_1, x_2, \dots, C_1, C_2, \dots\ \boldsymbol{q}, \boldsymbol{\theta} \mid I) = \\
&\qquad \prod_i \bigl[q_{C_i}\ K(x_i \mid \theta_{C_i})\bigr]\ \mathrm{Dir}(\boldsymbol{q} \mid \alpha)\ \prod_c G(\theta_c \mid I) = \\
&\qquad \mathrm{Dir}(\boldsymbol{q} \mid \alpha) \left[\prod_c \prod_i^{C_i=c} q_c\right] \times \prod_c \left[G(\theta_c \mid I) \prod_i^{C_i=c} K(x_i \mid \theta_c)\right] .
\end{aligned}
\tag{7}
$$

The coordinates $(\boldsymbol{q}, \boldsymbol{\theta})$ can in turn be interpreted as an atomic probability density $Q(\theta)$:

$$
Q(\theta)\, \mathrm{d}\theta \coloneqq \sum_c q_c\ \delta(\theta - \theta_c)\, \mathrm{d}\theta\ . \tag{8}
$$

We have effectively established a transformation from atomic densities on a θ-space to continuous densities on an x-space,

$$
Q(\theta)\, \mathrm{d}\theta \mapsto f(x)\, \mathrm{d}x\ . \tag{9}
$$

This transformation can be seen as a change of coordinates in an infinite-dimensional manifold. From this point of view I shall speak of the "coordinates Q" and of the "coordinates f". There are important topological details behind this point of view[3].

In coordinates Q the long-run frequency density for x is given by

$$
\mathrm{p}(x \mid Q, I) = \int K(x \mid \boldsymbol{m}, \boldsymbol{S})\ Q(\theta)\, \mathrm{d}\theta\ , \tag{10}
$$

and de Finetti's formula (1) becomes

$$
\begin{aligned}
\mathrm{p}(x_1, x_2, \dots \mid I) &= \int \left[\prod_i \mathrm{p}(x_i \mid Q, I)\right] \mathrm{p}(Q \mid I)\, \mathrm{d}Q \\
&\equiv \int \left[\prod_i \int K(x_i \mid \theta)\ Q(\theta)\, \mathrm{d}\theta\right] \mathrm{p}(Q \mid I)\, \mathrm{d}Q\ .
\end{aligned}
\tag{11}
$$

Remember that the densities Q are discrete. The "Dirichlet process" is a probability distribution on a manifold of densities, with support over discrete densities. When we use mixtures of Dirichlet processes we are stating that our prior density $\mathrm{p}(Q \mid I)$ is a Dirichlet process.

[3] See e.g. Antoniak 1974.

Bibliography

(“de *X*” is listed under D, “van *X*” under V, and so on, regardless of national conventions.)

Antoniak, C. E. (1974): *Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems*. Ann. Stat. **2**6, 1152–1174.

Choquet-Bruhat, Y., DeWitt-Morette, C., Dillard-Bleick, M. (1996): *Analysis, Manifolds and Physics. Part I: Basics*, rev. ed. (Elsevier, Amsterdam). First publ. 1977.

Damien, P., Dellaportas, P., Polson, N. G., Stephens, D. A., eds. (2013): *Bayesian Theory and Applications*. (Oxford University Press, Oxford).

Dawid, A. P. (2013): *Exchangeability and its ramifications*. In: Damien, Dellaportas, Polson, Stephens (2013): ch. 2:19–29.

de Finetti, B. (1930): *Funzione caratteristica di un fenomeno aleatorio*. Atti Accad. Lincei: Sc. Fis. Mat. Nat. **IV**5, 86–133. http://www.brunodefinetti.it/Opere.htm.

— (1937): *La prévision: ses lois logiques, ses sources subjectives*. Ann. Inst. Henri Poincaré **7**1, 1–68. Transl. in Kyburg, Smokler (1980), pp. 53–118, by Henry E. Kyburg, Jr.

Kyburg Jr., H. E., Smokler, H. E., eds. (1980): *Studies in Subjective Probability*, 2nd ed. (Robert E. Krieger, Huntington, USA). First publ. 1964.