Bayesian Analysis (2011) **6**, Number 3, pp. 359–386

An enriched conjugate prior for Bayesian nonparametric inference

Sara Wade*, Silvia Mongelluzzo† and Sonia Petrone‡

Abstract. The precision parameter α plays an important role in the Dirichlet Process. When assigning a Dirichlet Process prior to the set of probability measures on $\mathbb{R}^k$, $k > 1$, this can be restrictive in the sense that the variability is determined by a single parameter. The aim of this paper is to construct an enrichment of the Dirichlet Process that is more flexible with respect to the precision parameter yet still conjugate, starting from the notion of *enriched conjugate priors*, which have been proposed to address an analogous lack of flexibility of standard conjugate priors in a parametric setting. The resulting enriched conjugate prior allows more flexibility in modelling uncertainty on the marginal and conditionals. We describe an enriched urn scheme which characterizes this process and show that it can also be obtained from the stick-breaking representation of the marginal and conditionals. For non atomic base measures, this allows global clustering of the marginal variables and local clustering of the conditional variables. Finally, we consider an application to mixture models that allows for uncertainty between homoskedasticity and heteroskedasticity.

Keywords: Bayesian nonparametric inference, conjugate priors, generalized Dirichlet, Dirichlet process, mixture models, Pólya urns, multivariate random distribution functions

1 Motivation

Conjugacy is a desirable property because the posterior distribution remains analytically tractable; this is especially true in nonparametric inference where the posterior distribution of non-conjugate priors can be very complex. The most popular prior in Bayesian nonparametric inference is the Dirichlet Process, and it is conjugate; if $Z_i \mid \mathbf{P} = P$ are independent and identically distributed (i.i.d.) according to P, and $\mathbf{P}$ is a Dirichlet process, $DP(\alpha P_0)$, with precision parameter α and base measure P_0 on the sample space $\mathcal{Z}$, then $\mathbf{P} \mid Z_1 = z_1, \ldots, Z_n = z_n \sim DP(\alpha P_0 + \sum_{i=1}^n \delta_{z_i})$. However, when Z is a random vector and $\mathbf{P}$ is a random probability measure on $\mathbb{R}^k$, $k > 1$, as in many applications, the choice of a Dirichlet process prior implies that the variability is determined by a single parameter, α. Indeed, the precision parameter α plays an important role; it not only reflects the strength of belief in the prior guess of P_0, but also controls the ties configuration in a random sample from $\mathbf{P}$. Thus, having only one degree of freedom, α,

*Department of Decision Sciences, Bocconi University, Milan, Italy, mailto:sara.wade@phd.unibocconi.it

†Department of Decision Sciences, Bocconi University, Milan, Italy, mailto:silvia.mongelluzzo@phd.unibocconi.it

‡Department of Decision Sciences, Bocconi University, Milan, Italy, mailto:sonia.petrone@unibocconi.it

 DOI:10.1214/11-BA614

in the prior can be quite restrictive.

In fact, a similar lack of flexibility arises in a parametric setting; standard conjugate priors for the natural exponential family have only one parameter to control variability. To overcome this issue, a general class of *enriched conjugate priors* (Consonni and Veronese (2001)) have been proposed. A Dirichlet Process, $DP(\alpha P_0)$, is characterized by the fact that the finite dimensional distributions of the probability over any measurable partition,$(C_1, \ldots, C_k)$, of $\mathcal{Z}$, are Dirichlet with parameters $(\alpha P_0(C_1), \ldots, \alpha P_0(C_k))$. The Dirichlet Process inherits conjugacy from the property of conjugacy of the standard Dirichlet distribution prior for multinomial sampling, but also inflexibility from the fact that the Dirichlet distribution, as all standard conjugate priors, has only one parameter to control variability. The question addressed in this paper is whether one can extend the notion of enriched conjugate priors to nonparametric inference and construct a prior on a random probability measure over $\mathbb{R}^k$, that is more flexible than the DP in allowing more parameters to control the variability, yet is still conjugate.

Actually, Doksum's Neutral to the Right Process (Doksum (1974)) is an extension of the enriched conjugate Generalized Dirichlet distribution to a process, providing a more flexible, conjugate prior for *univariate* random distribution functions. The Generalized Dirichlet distribution is defined for a specific ordering of the random probabilities; thus, extension to a multivariate random distribution is not obvious, since there is no natural ordering in $\mathbb{R}^k$.

Therefore, we start our analysis by constructing an enriched Dirichlet prior for a multivariate random distribution when the sample space is finite. To convey the main ideas, we will focus on the case when the random vector Z can be partitioned into two groups, $Z = (X, Y)$, and the sample space can be written as the product of two finite spaces (or in the more general case, the product of two complete separable metric spaces, $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$). In the finite case, the enriched Dirichlet distribution is obtained based on the reparametrization of the joint probabilities in terms of the marginal and the conditionals.

Then, we extend this construction to a process by reparametrizing the joint random probability measure in terms of the marginal and conditionals and assigning independent Dirichlet Process priors to each of these terms. The parameters of the resulting *enriched* Dirichlet process again include a base measure controlling the location, but there are now many more parameters to control the variability. We show that the Dirichlet Process is in fact a special case, which consequently, characterizes the distribution of the random conditionals. Although many desirably properties are maintained, some are necessarily weakened, including a clear asymmetry in the two (groups of) variables, that however may be reasonable in several applications.

The paper is organized as follows. In Section 2, we give a brief overview of enriched conjugate priors for the natural exponential family. In Section 3, we discuss the *enriched* Dirichlet distribution in the finite case as a particular enriched conjugate prior for multinomial sampling and provide a Pólya urn characterization. These notions are extended to a process in Section 4. Finally, a simple application to mixture models is illustrated using data on national test scores to compare schools in Section 5. Proofs

are given in the Appendix.

2 Preliminaries: Enriched Conjugate Priors

For a Natural Exponential Family (NEF) $\mathcal{F}$ on $\mathbb{R}^d$, where d represents the dimension of the sufficient statistics, the likelihood for the natural parameter θ is given by:

$$L_\theta(\theta|\underline{s}, n) = exp(\theta^T \underline{s} - nM(\theta)) \quad \text{for } \theta \in \Theta,$$

where $\underline{s}$ is a d-dimensional vector of the sufficient statistics, $M(\theta) = \log \int exp(\theta^T x)\eta(dx)$, and η is a σ-finite measure on the Borel sets of $\mathbb{R}^d$. The parameter space Θ is the interior of the set $\mathcal{N} = \{\theta \in \mathbb{R}^d : M(\theta) < \infty\}$. More generally, we have a Standard Exponential Family (SEF) if $\Theta \subseteq \mathcal{N}$, and it is non-empty and open.

A family of measures on the Borel sets of Θ whose densities with respect to the Lebesgue measure are of the form $\pi_\theta(\theta|\underline{s}', n') \propto L_\theta(\theta|\underline{s}', n')$ is called the *standard conjugate family of priors* of $\mathcal{F}$ relative to the parametrization θ, where the sufficient statistics, $\underline{s}$, are replaced by parameters, $\underline{s}'$, which control the location of the prior, and the sample size, n, is replaced by a single parameter, n', which controls the precision; see Diaconis and Ylvisaker (1979).

Consonni and Veronese (2001) discuss *enriched* conjugate priors for the NEF, moving from the notion of conditional reducibility. A d-dimensional NEF is called k *conditionally reducible* if the density can be decomposed as the product of k standard exponential families, each depending on their own parameters. The notion of *enriched conjugate priors* involves replacing the sufficient statistics and the sample size with different hyperparameters within each SEF. This means giving independent standard conjugate priors to the parameters of the conditional densities and induces a prior on the original parameter of the NEF which enriches the standard conjugate prior by allowing for k precision parameters. For a deeper discussion, see Consonni and Veronese (2001).

One important example is given by the Generalized Dirichlet distribution of Connor and Mosimann (1969), which provides an enriched conjugate prior for the parameters of a multinomial distribution; see Consonni and Veronese (2001), Example 4. Briefly, if $(N_1, \ldots, N_k)$ is multinomial given $(\mathbf{p}_1 = p_1, \ldots, \mathbf{p}_k = p_k)$, one can decompose the multinomial probability function as

$$p(N_1 = n_1, \ldots, N_k = n_k \mid p_1, \ldots, p_k) =$$
$$p(N_1 = n_1 \mid V_1)p(N_2 = n_2 \mid N_1 = n_1, V_2) \cdots p(N_k = n_k \mid N_1 = n_1, \ldots, N_{k-1} = n_{k-1}, V_k),$$

where each factor in the product is a NEF (namely, binomial), depending on its own parameter, $\mathbf{V}_1 = \mathbf{p}_1, \mathbf{V}_i = \mathbf{p}_i/(1 - \sum_{j=1}^{i-1} \mathbf{p}_j), i = 2, \ldots, k-1$, and $\mathbf{V}_k$ is degenerate at 1. The standard, Dirichlet$(\alpha_1, \ldots, \alpha_k)$ conjugate prior corresponds to assuming $\mathbf{V}_i \overset{indep}{\sim}$ beta$(\alpha_i, \sum_{j=i+1}^{k} \alpha_j)$, $i = 1, \ldots, k-1$. The enriched, or Generalized, Dirichlet conjugate prior allows a more flexible choice of the beta hyperparameters: $\mathbf{V}_i \overset{indep}{\sim}$ beta(α_i, β_i), $i = 1, \ldots, k-1$. It is worth underlining that some properties of the Dirichlet distribution are necessarily weakened. In particular, the Dirichlet prior implies that *any permutation*

of $(\mathbf{p}_1, \ldots, \mathbf{p}_k)$ is completely neutral (the vector $(\mathbf{p}_1, \ldots, \mathbf{p}_k)$ is completely neutral iff $(\mathbf{p}_1, \mathbf{p}_2/(1-\mathbf{p}_1), ..., \mathbf{p}_k/(1-\sum_{j=1}^{k-1}\mathbf{p}_j))$ are independent). The Generalized Dirichlet only assumes that *one* ordered vector $(\mathbf{p}_1, \ldots, \mathbf{p}_k)$ is completely neutral. This makes applications to the bivariate case of contingency tables $\mathbf{p}_{i,j}$ not obvious, since there is no natural ordering in two dimensions. The enriched conjugate prior that we propose in the next section is a simple proposal in this direction.

3 Finite case: Enriched Dirichlet distribution

Let $\{(X_n, Y_n)\}_{n\in\mathbb{N}}$ be a sequence of discrete random vectors with values in $\mathcal{X}\times\mathcal{Y} = \{1,\ldots,k\}\times\{1,\ldots,m\}$, such that $(X_i, Y_i) \mid \mathbf{p} = p \overset{i.i.d}{\sim} p$, where $\mathbf{p}$ is a random probability function with mass $\mathbf{p}_{i,j}$ on $(i,j), i = 1,\ldots,k; j = 1,\ldots,m$. Then, given $\mathbf{p} = p$, the vector of counts $(N_{1,1},\ldots,N_{k,m})$, where $N_{i,j}$ is the number of times the pair (i,j) is observed in a sample $((X_1,Y_1),\ldots,(X_n,Y_n))$, has a multinomial probability function

$$p(n_{1,1},...,n_{k,m-1} \mid p_{1,1},...,p_{k,m-1}) = \frac{n!}{n_{1,1}!...n_{k,m-1}!(n-\sum\limits_{(i,j)\neq(k,m)} n_{i,j})!} p_{1,1}^{n_{1,1}} \cdots p_{k,m-1}^{n_{k,m-1}} \left(1-\sum_{(i,j)\neq(k,m)} p_{i,j}\right)^{n-\sum\limits_{(i,j)\neq(k,m)} n_{i,j}}, \tag{1}$$

for $n_{i,j} \geq 0; \sum_{i=1}^{k}\sum_{j=1}^{m} n_{i,j} = n$. The standard conjugate prior for $(\mathbf{p}_{1,1},\ldots,\mathbf{p}_{k,m})$ is the Dirichlet distribution, which involves replacing the $km-1$ sufficient statistics in (1) with hyperparameters, $\underline{s'} = (s'_{1,1},...,s'_{k,m-1})$, that control the location of the prior, and the sample size with a single hyperparameter, n', that controls the precision of the prior. As discussed in Section 2, a generalized Dirichlet prior is problematic in this case, since there is no natural ordering of the probabilities $\mathbf{p}_{i,j}$.

However, a fairly natural and simple enrichment can be obtained by first applying the linear transformation:

$$\begin{aligned} N_{i+} &= \sum_{j=1}^{m} N_{i,j} \quad \text{for } i = 1,...,k-1, \\ N_{i,j} &= N_{i,j} \quad \text{for } i = 1,...,k \quad j = 1,...,m-1, \end{aligned}$$

followed by the reparametrization:

$$\begin{aligned} \mathbf{p}_{i+} &= \sum_{j=1}^{m} \mathbf{p}_{i,j} \quad \text{for } i = 1,...,k-1, \\ \mathbf{p}_{j|i} &= \frac{\mathbf{p}_{i,j}}{\mathbf{p}_{i+}} \quad \text{for } i = 1,...,k-1 \quad j = 1,...,m-1, \\ \mathbf{p}_{j|k} &= \frac{\mathbf{p}_{k,j}}{1-\sum_{i=1}^{k-1}\mathbf{p}_{i+}} \quad \text{for } j = 1,...,m-1. \end{aligned}$$

Define: $\underline{N}_+ = (N_{1+}, ..., N_{k-1+})$; $\underline{N}_i = (N_{i,1}, ..., N_{i,m-1})$; $\underline{\mathbf{p}}_+ = (\mathbf{p}_{1+}, ..., \mathbf{p}_{k-1+})$, and $\underline{\mathbf{p}}_i = (\mathbf{p}_{1|i}, ..., \mathbf{p}_{m-1|i})$, for $i = 1, ..., k$. Under this linear transformation and reparametrization, the multinomial is a $k+1$ conditionally reducible NEF:

$$p(\underline{n}_+, \underline{n}_1, ..., \underline{n}_k \mid \underline{p}_+, \underline{p}_1, ..., \underline{p}_k) = p(\underline{n}_+ \mid \underline{p}_+) \prod_{i=1}^{k} p(\underline{n}_i \mid \underline{p}_i, \underline{n}_+), \tag{2}$$

$$(N_{i,1}, ..., N_{i,m} \mid n_{i+}, p_{1|i}, ..., p_{m|i}) \sim \text{Mult}\left(n_{i+}, p_{1|i}, ..., p_{m|i}\right) \quad \text{for } i = 1, ..., k,$$

$$(N_{1+}, ..., N_{k+} \mid p_{1+}, ..., p_{k+}) \sim \text{Mult}(n, p_{1+}, ..., p_{k+}).$$

By replacing the sufficient statistics and sample size with different parameters within each SEF in the right hand side of (2), one can create a more flexible conjugate prior. In particular, letting $(\underline{s}'_{(+)}, \underline{s}'_{(1)}, ..., \underline{s}'_{(k)})$ denote the $km-1$ location parameters and $(n'_+, n'_1, ..., n'_k)$ denote the precision parameters, in terms of $(\underline{\mathbf{p}}_+, \underline{\mathbf{p}}_1, ..., \underline{\mathbf{p}}_k)$, the Enriched Dirichlet conjugate prior is:

$$\mathbf{p}_{1+}, ..., \mathbf{p}_{k+} \sim \text{Dir}(s'_{1+}, ..., s'_{k-1+}, n'_+ - \sum_{i=1}^{k-1} s'_{i+}), \tag{3}$$

$$\mathbf{p}_{1|i}, ..., \mathbf{p}_{m|i} \sim \text{Dir}(s'_{i,1}, ..., s'_{i,m-1}, n'_i - \sum_{j=1}^{m-1} s'_{i,j}),$$

where $(\mathbf{p}_{1+}, ..., \mathbf{p}_{k+}), (\mathbf{p}_{1|1}, ..., \mathbf{p}_{m|1}), ..., (\mathbf{p}_{1|k}, ..., \mathbf{p}_{m|k})$ are independent. We get back to the Dirichlet distribution if $n'_i = s'_{i+}$ for $i = 1, ...k-1$ and $n'_+ = \sum_{i=1}^{k} n'_i$.

Remark 1. The Dirichlet distribution on the vector $\underline{\mathbf{p}} = (\mathbf{p}_{1,1}, ..., \mathbf{p}_{k,m})$ defining the random marginal, $\mathbf{p}_x$, $\mathbf{p}_y$, and conditional, $\mathbf{p}_{y|x}$, $\mathbf{p}_{x|y}$, probability functions is characterized by the properties
(i) $\mathbf{p}_x(\cdot)$ and $\mathbf{p}_{y|x}(\cdot|i), i = 1, \ldots, k$ are independent, and
(ii) $\mathbf{p}_y(\cdot)$ and $\mathbf{p}_{x|y}(\cdot|j), j = 1, \ldots, m$ are independent;
see Geiger and Heckerman (1997). The Enriched Dirichlet relaxes that the independence properties holds in both directions. We maintain (i) and allow more degrees of freedom in the distributions of $\mathbf{p}_x$ and $\mathbf{p}_{y|x}$.

Remark 2. Under the linear transformation discussed here, the multinomial could also be viewed as a $km-1$ conditionally reducible NEF; it can be written as the product of $km-1$ SEFs (namely, binomial) each depending on its own parameters. The resulting enriched conjugate prior has $km-1$ parameters to control the precision and can be seen as nested version of Generalized Dirichlet distribution of Connor and Mosimann (1969).

In the rest of the paper, we will use the following parametrization of the distributions (3). Let $\alpha(\cdot)$ be a finite measure on $\mathcal{X}$ and $\mu(\cdot, \cdot)$ be a mapping from $2^{\mathcal{Y}} \times \mathcal{X}$ to $\mathbb{R}_+$ such that for every $x \in \mathcal{X}$, $\mu(\cdot, x)$ is a finite measure on $(\mathcal{Y}, 2^{\mathcal{Y}})$. Then we assume that the parameters in (3) are chosen in terms of $\alpha(\cdot)$ and $\mu(\cdot, \cdot)$:

$$\mathbf{p}_{1+}, ..., \mathbf{p}_{k+} \sim \text{Dir}(\alpha(1), ..., \alpha(k)), \tag{4}$$

$$\mathbf{p}_{1|i}, ..., \mathbf{p}_{m|i} \sim \text{Dir}(\mu(1, i), ..., \mu(m, i)) \quad i = 1, ..., k,$$

with the convention that if $\alpha(i) = 0$ then $\mathbf{p}_{i+}$ is degenerate at 0 and if $\mu(j,i) = 0$ then $\mathbf{p}_{j|i}$ is degenerate at 0. If $\alpha(i) > 0$ and $\mu(j,i) > 0$ for all i, j, then the enriched Dirichlet conjugate prior induced on $(\mathbf{p}_{1,1}, \ldots, \mathbf{p}_{k,m})$ is:

$$f(p_{1,1},...,p_{k,m-1}) = \frac{\Gamma(\alpha(\mathcal{X}))}{\prod_{i=1}^{k}\Gamma(\alpha(i))}\prod_{i=1}^{k-1}(\sum_{j=1}^{m}p_{i,j})^{\alpha(i)-\mu(\mathcal{Y},i)}(1-\sum_{i=1}^{k-1}\sum_{j=1}^{m}p_{i,j})^{\alpha(k)-\mu(\mathcal{Y},k)}$$
$$\times\prod_{i=1}^{k}\frac{\Gamma(\mu(\mathcal{Y},i))}{\prod_{j=1}^{m}\Gamma(\mu(j,i))}\prod_{j=1}^{m-1}p_{i,j}^{\mu(j,i)-1}\prod_{i=1}^{k-1}p_{i,m}^{\mu(m,i)-1}(1-\sum_{(i,j)\neq(k,m)}p_{i,j})^{\mu(m,k)-1}.$$

Clearly, the prior of the marginal probabilities $(\mathbf{p}_{+1}, \ldots, \mathbf{p}_{+m})$ on $\mathcal{Y}$ is no longer a Dirichlet distribution, and in fact, the density may not be available in closed form. But, we can give the following representation in terms of G-Meijer variables (Springer and Thompson (1970)). First, remembering the Gamma representation of the Dirichlet distribution and defining $\mathbf{U}_i \overset{indep}{\sim} \text{Gamma}(\alpha(i), 1)$ and $\mathbf{V}_{ij} \overset{indep}{\sim} \text{Gamma}(\mu(j,i), 1)$, we have the following G-Meijer representation of the vector $(\mathbf{p}_{1,1}, ..., \mathbf{p}_{k,m})$:

$$(\mathbf{p}_{1,1}, ..., \mathbf{p}_{k,m}) \overset{d}{=} \left(\frac{\mathbf{U}_1\mathbf{V}_{11}}{\sum_{i=1}^{k}\mathbf{U}_i\sum_{j=1}^{m}\mathbf{V}_{1j}}, ..., \frac{\mathbf{U}_k\mathbf{V}_{km}}{\sum_{i=1}^{k}\mathbf{U}_i\sum_{j=1}^{m}\mathbf{V}_{kj}}\right),$$

which is independent of $\sum_{i=1}^{k}\mathbf{U}_i\sum_{j=1}^{m}\mathbf{V}_{1j}, ..., \sum_{i=1}^{k}\mathbf{U}_i\sum_{j=1}^{m}\mathbf{V}_{kj}$; where the symbol $\overset{d}{=}$ denotes equality in distribution. Therefore, the marginal probabilities over $\mathcal{Y}$ can be represented as the sum of G-Meijer random variables:

$$(\mathbf{p}_{+1}, ..., \mathbf{p}_{+m}) \overset{d}{=} \left(\sum_{i=1}^{k}\frac{\mathbf{U}_i\mathbf{V}_{i1}}{\sum_{h=1}^{k}\mathbf{U}_h\sum_{j=1}^{m}\mathbf{V}_{ij}}, ..., \sum_{i=1}^{k}\frac{\mathbf{U}_i\mathbf{V}_{im}}{\sum_{h=1}^{k}\mathbf{U}_h\sum_{j=1}^{m}\mathbf{V}_{ij}}\right).$$

3.1 Enriched Pólya Urn

An alternative way to define the Enriched Dirichlet distribution is based on a Pólya urn scheme, which will be useful in extending the distribution to a process. In the bivariate setting, the standard Pólya urn scheme describes the predictive distribution of a sequence of random vectors. An urn contains pairs of balls of color $(i, j) \in \mathcal{X} \times \mathcal{Y}$. A pair of balls is drawn from the urn and replaced along with another pair of balls of the same colors. The random vector, (X_n, Y_n), is equal to (i, j) if the n-th pair drawn is of color (i, j).

Alternatively, we can consider one urn containing just X-balls and k urns, say $Y|i$ urns, containing only Y-balls. We first draw an X-ball from the X-urn and replace it along with another ball of the same color, and then, depending on color of the X-ball, draw a Y-ball from urn associated to X-ball drawn, and replace it along with another ball of the same color. In this case, the random vector, (X_n, Y_n), is equal to (i, j) if the

n-th X-ball drawn is of color i and the Y ball associated with it is of color j. If the number of Y-balls in the $Y|i$ urn is equal to the number balls of color i in the X-urn, the two urn schemes are equivalent.

The Enriched Pólya Urn scheme enriches this urn scheme by relaxing the constraints that the number of Y-balls in the $Y|i$ urn has to equal the number of X-balls of color i in the X-urn for $i = 1, ..., k$. More precisely, the number of balls in each urn is specified as follows:

- $\alpha(i)$ is the number of X-balls of color i
- $\mu(j, i)$ is the number of Y-balls of color j in the $Y|i$ urn

where $\alpha(\mathcal{X}) = \sum_{i=1}^{k} \alpha(i)$ is the total number of balls in the X-urn and $\mu(\mathcal{Y}, i) = \sum_{j=1}^{m} \mu(j, i)$ is the total number of balls in the $Y|i$ urn for $i = 1, ..., k$. This urn scheme implies the following predictive distribution:

$$Pr(X_1 = i, Y_1 = j) = \frac{\alpha(i)}{\alpha(\mathcal{X})} \frac{\mu(j,i)}{\mu(\mathcal{Y}, i)},$$

$$Pr(X_{n+1} = i, Y_{n+1} = j | X_1 = i_1, Y_1 = j_1, .., X_n = i_n, Y_n = j_n)$$
$$= \frac{\alpha(i) + \sum_{h=1}^{n} \delta_{i_h}(i)}{\alpha(\mathcal{X}) + n} \frac{\mu(j,i) + \sum_{h=1}^{n} \delta_{j_h, i_h}(j,i)}{\mu(\mathcal{Y}, i) + \sum_{h=1}^{n} \delta_{i_h}(i)}.$$

Theorem 1. *Let $\{(X_n, Y_n)\}_{n\in\mathbb{N}}$ be a sequence of random vectors taking values in $\{1, ..., k\} \times \{1, ..., m\}$ with predictive distributions characterized by an Enriched Pólya urn scheme with parameters $\alpha(\cdot)$ and $\mu(\cdot, \cdot)$. Then,*

1. *the sequence of random vectors $\{(X_n, Y_n)\}_{n\in\mathbb{N}}$ is exchangeable, and its de Finetti measure is an Enriched Dirichlet distribution with parameters $\alpha(\cdot)$ and $\mu(\cdot, \cdot)$.*
2. *as $n \to \infty$, the sequence of the predictive distributions $p_n(i, j) = Pr(X_{n+1} = i, Y_{n+1} = j | X_1 = i_1, Y_1 = j_1, .., X_n = i_n, Y_n = j_n)$ converges a.s with respect to the exchangeable law to a random probability function, $\mathbf{p}$; and $\mathbf{p}$ is distributed according to the Enriched Dirichlet de Finetti measure.*

The proof is an extension of that used for the standard Pólya urn (see Ghosh and Ramamoorthi (2003), pages 94-95). The first step is to show the sequence of random vectors is exchangeable. Next, computing their finite dimensional distributions and using de Finetti's Representation Theorem, the random vectors are shown to be i.i.d given the random variables $(\mathbf{p}_{1+}, ...\mathbf{p}_{k+}, \mathbf{p}_{1|1}, ..., \mathbf{p}_{m|k}) = (p_{1+}, ...p_{k+}, p_{1|1}, ..., p_{m|k})$ which are distributed according to an Enriched Dirichlet distribution with parameters α and μ. A detailed proof is given in the Appendix.

4 Enriched Dirichlet Process

Assume $\mathcal{X}$ and $\mathcal{Y}$ are complete and separable metric spaces with Borel σ-algebras $\mathcal{B}_X$ and $\mathcal{B}_Y$. Let $\mathcal{B}$ be the σ-algebra generated by the product of the σ-algebras of $\mathcal{X}$ and $\mathcal{Y}$ and

$\mathcal{P}(\mathcal{B})$ be the set of probability measures on the measurable product space $(\mathcal{X} \times \mathcal{Y}, \mathcal{B})$ where $\mathcal{P}(\mathcal{B}_X)$, $\mathcal{P}(\mathcal{B}_Y)$ are similarly defined. For any $P \in \mathcal{P}(\mathcal{B})$, let P_X denote the marginal probability measure, $P_{Y|X}(\cdot|x)$ for $x \in \mathcal{X}$ denote a version of the conditional, and $P_{Y|X}$ denote the entire version of the conditional as an element of $\mathcal{P}(\mathcal{B}_Y)^{\mathcal{X}}$. Here, we consider the Borel σ-algebra under weak convergence on $\mathcal{P}(\mathcal{B})$, $\mathcal{P}(\mathcal{B}_X)$, and $\mathcal{P}(\mathcal{B}_Y)$ and the product σ-algebra on $\mathcal{P}(\mathcal{B}_Y)^{\mathcal{X}}$. We will define a probability measure on $\mathcal{P}(\mathcal{B})$ that is more flexible than the Dirichlet Process with respect to the precision parameter and still retains conjugacy by extending the ideas of the Enriched Dirichlet distribution.

Note that trying to enrich the DP by using the Enriched Dirichlet in place of the Dirichlet as the finite dimensional distributions, i.e., defining a random $\mathbf{P}$ such that $(\mathbf{P}(A_1 \times B_1), \ldots, \mathbf{P}(A_k \times B_m)) \sim$ Enriched Dirichlet distribution, would not succeed because finite additivity holds only with a specification of the parameters that is equivalent to the Dirichlet distribution.

Instead, we use directly the idea of the Enriched Dirichlet distribution, which defines a prior for the joint by first, decomposing it in terms of the marginal and conditionals and then, assigning independent conjugate priors to them. If $\mathcal{X}, \mathcal{Y}$ are general spaces, it is a delicate issue to establish that such an approach induces a prior on the joint. In particular, given a prior on $\mathcal{P}(\mathcal{B}_X) \times \mathcal{P}(\mathcal{B}_Y)^{\mathcal{X}}$, the map $(P_X, P_{Y|X}) \to \int_{(\cdot)} P_{Y|X}(\cdot|x) dP_X(x)$ induces a prior on $\mathcal{P}(\mathcal{B})$ if it is jointly measurable in $(P_X, P_{Y|X})$, which is not true in general. Fortunately, if the prior for the marginal concentrates on the set of discrete probability measures and independence assumptions hold, the prior on the marginal and conditionals can be restricted to a subspace of $\mathcal{P}(\mathcal{B}_X) \times \mathcal{P}(\mathcal{B}_Y)^{\mathcal{X}}$ that has measure one, and on this subspace, the mapping is measurable, which is shown after the following definition.

Definition 2. *Let α be a finite measure on $(\mathcal{X}, \mathcal{B}_X)$ and μ be a mapping from $(\mathcal{B}_Y \times \mathcal{X})$ to $\mathbb{R}_+$ such that as a function of $B \in \mathcal{B}_Y$ it is a finite measure on $(\mathcal{Y}, \mathcal{B}_Y)$ and as a function of $x \in \mathcal{X}$ it is α-integrable. Assume:*

1. *Law of Marginal, Q^X: $\mathbf{P}_X$ is a random probability measure on $(\mathcal{X}, \mathcal{B}_X)$ where $\mathbf{P}_X \sim DP(\alpha)$.*
2. *Law of Conditionals, $Q_x^{Y|X}$: $\forall x \in \mathcal{X}$, $\mathbf{P}_{Y|X}(\cdot|x)$ is a random probability measure on $(\mathcal{Y}, \mathcal{B}_Y)$ where $\mathbf{P}_{Y|X}(\cdot|x) \sim DP(\mu(\cdot, x))$.*
3. *Joint Law of Conditionals, $Q^{Y|X} = \prod_{x \in \mathcal{X}} Q_x^{Y|X}$: $\mathbf{P}_{Y|X}(\cdot|x), x \in \mathcal{X}$ are independent among themselves.*
4. *Joint Law of Marginal and Conditionals, $Q = Q^X \times Q^{Y|X}$: $\mathbf{P}_X$ is independent of $\{\mathbf{P}_{Y|X}(\cdot|x)\}_{x \in \mathcal{X}}$.*

The joint law of the marginal and conditionals, Q, induces the law, $\tilde{Q}$, of the stochastic process $\{\mathbf{P}(C)\}_{C \in \mathcal{B}}$ through the following reparametrization:

$$\mathbf{P}(A \times B) \stackrel{d}{=} \int_A \mathbf{P}_{Y|X}(B \mid x) d\mathbf{P}_X(x), \quad \text{for any set } A \times B \in \mathcal{B}_X \times \mathcal{B}_Y. \tag{5}$$

This process is called an Enriched Dirichlet Process *(EDP) with parameters* α *and* μ*, and is denoted* $\mathbf{P} \sim EDP(\alpha, \mu)$.

The following arguments verify that the four conditions in definition (2) induce a distribution for the random joint probability measure. In particular, we define a subspace of $\mathcal{P}(\mathcal{B}_X) \times \mathcal{P}(\mathcal{B}_Y)^{\mathcal{X}}$ that has measure one, such that on this subspace, the mapping $(P_X, P_{Y|X}) \to \int_{(\cdot)} P_{Y|X}(\cdot|x)dP_X(x)$ is measurable.

First note that in order for $\left\{\mathbf{P}_{Y|X}(\cdot|x), x \in \mathcal{X}\right\}$ to be a set of conditional random probability measures, the following two properties need to be satisfied:

1. $\forall x \in \mathcal{X}$, $\mathbf{P}_{Y|X}(\cdot|x)$ is a probability measure on $(\mathcal{Y}, \mathcal{B}_Y)$ a.s $Q_x^{Y|X}$.
2. $\forall B \in \mathcal{B}_Y$, as a function of x, $\mathbf{P}_{Y|X}(B|x)$ is $\mathcal{B}_X$ measurable a.s $Q^{Y|X}$.

The first item is satisfied since $\mathbf{P}_{Y|X}(\cdot|x) \sim DP(\mu(\cdot, x))$ implies $\mathbf{P}_{Y|X}(\cdot|x) \in \mathcal{P}(\mathcal{B}_Y)$ with probability one. The second property follows from results of Ramamoorthi and Sangalli (2006). In particular, letting Δ be the subset of $\mathcal{P}(\mathcal{B}_Y)^{\mathcal{X}}$ such that $P_{Y|X}$ is measurable as a function of x, they show that if $\mathbf{P}_{Y|X}(\cdot|x)$ are independent among $x \in \mathcal{X}$, then the product measure, $Q^{Y|X} = \prod_{x \in \mathcal{X}} Q_x^{Y|X}$, given by Kolmogorov's Extension Theorem, assigns outer measure one to Δ.

Let $\mathcal{P}_D(\mathcal{B}_X)$ denote the set of discrete probability measures on the measurable space $(\mathcal{X}, \mathcal{B}_X)$. From properties of the DP, $Q^X(\mathcal{P}_D(\mathcal{B}_X)) = 1$. Therefore, by independence of $\mathbf{P}_X$ and $\mathbf{P}_{Y|X}$, the set $\mathcal{P}_D(\mathcal{B}_X) \times \Delta$ has Q-measure one. Again, by results of Ramamoorthi and Sangalli (2006), on $\mathcal{P}_D(\mathcal{B}_X) \times \Delta$, for $A \times B \in \mathcal{B}_X \times \mathcal{B}_Y$, the function $(P_X, P_{Y|X}) \to \int_A P_{Y|X}(B|x)dP_X(x)$ is jointly measurable in $(P_X, P_{Y|X})$. These results imply that we can define a prior, $\tilde{Q}$, on $\mathcal{P}(\mathcal{B})$ induced from Q restricted to $\mathcal{P}_D(\mathcal{B}_X) \times \Delta$ via the map $(P_X, P_{Y|X}) \to \int_{(\cdot)} P_{Y|X}(\cdot|x)dP_X(x)$.

Obviously, this map is not $1-1$. In fact, the definition of the EDP states that the four conditions hold for the joint distribution of $(\mathbf{P}_X, \mathbf{P}_{Y|X})$ for a fixed version of the conditional, and this induces a prior on the joint. However, from the induced prior on the random joint probability measure, we can obtain the joint distribution of $\mathbf{P}_X$ and $\mathbf{P}_{Y|X}$ through the mapping $\mathbf{P} \to (\mathbf{P}_X, \mathbf{P}_{Y|X})$ defined from any version of the conditional. In the next section, we show that although the mapping is not 1-1, the joint law of $\mathbf{P}_X$ and $\mathbf{P}_{Y|X}$ defined from any version of the conditional and the induced law of the joint probability measure still satisfies the conditions in definition (2) through an extension of the enriched Pólya urn scheme to the infinite case.

4.1 Enriched Pólya Sequence

Similar to Blackwell and MacQueen (1973), we define an Enriched Pólya sequence which extends the enriched Pólya urn scheme to the case when $\mathcal{X}$ and $\mathcal{Y}$ are complete separable metric spaces.

Definition 3. *The sequence of random vectors* $\{(X_n, Y_n)\}_{n\in\mathbb{N}}$ *taking values in* $\mathcal{X}\times\mathcal{Y}$ *is an* Enriched Pólya sequence *with parameters* α *and* μ *if:*

1. *For* $A \in \mathcal{B}_X$ *and for all* $n \geq 1$,

$$Pr(X_1 \in A) = \frac{\alpha(A)}{\alpha(\mathcal{X})},$$
$$Pr(X_{n+1} \in A \mid X_1 = x_1, ..., X_n = x_n) = \frac{\alpha(A) + \sum_{i=1}^{n} \delta_{x_i}(A)}{\alpha(\mathcal{X}) + n}.$$

2. *For* $B \in \mathcal{B}_Y$ *and for all* $n \geq 1$,

$$Pr(Y_1 \in B \mid X_1 = x) = \frac{\mu(B, x)}{\mu(\mathcal{Y}, x)},$$
$$Pr(Y_{n+1} \in B \mid Y_1 = y_1, ..., Y_n = y_n, X_1 = x_1, ..., X_n = x_n, X_{n+1} = x) = \frac{\mu(B, x) + \sum_{j=1}^{n_x} \delta_{y_{x,j}}(B)}{\mu(\mathcal{Y}, x) + n_x},$$

where $n_x = \sum_{i=1}^{n} \delta_{x_i}(x)$ *and* $\{y_{x,j}\}_{j=1}^{n_x} = \{y_i : x_i = x, i = 1, ..., n\}$.

In words, the predictive distributions characterizing the Enriched Pólya sequence can be interpreted in terms of draws from urns as follows; initially, there is an X-urn containing $\alpha(\mathcal{X})$ balls of color 0. A ball is first drawn from the X-urn, and once drawn, its true color, x_1, is revealed (where x_1 is the realization of a draw from $P_{0X}(\cdot) = \frac{\alpha(\cdot)}{\alpha(\mathcal{X})}$). A ball of color x_1 is added to the urn along with a ball of color 0, so that the urn is now composed of $\alpha(\mathcal{X})$ balls of color 0 and one ball of color x_1. Once the true color x_1 of the X-ball is revealed, a $Y|x_1$-urn is created with $\mu(\mathcal{Y}, x_1)$ balls of color 0. Next, a ball is drawn from the $Y|x_1$-urn, and similarly, once drawn its true color is revealed to be y_1 (where y_1 is the realization of a draw from $P_{0Y|X}(\cdot|x_1) = \frac{\mu(\cdot, x_1)}{\mu(\mathcal{Y}, x_1)}$). This ball is then added to the $Y|x_1$-urn along with a ball of color 0, so that the urn contains $\mu(\mathcal{Y}, x_1)$ balls of color 0 and one ball of color y_1.

At the next stage, we again first draw a ball from the X-urn. We can either draw a 0 ball or an x_1 ball. If an x_1 ball is drawn, we replace it along with another ball of the same color and then draw a Y-ball from the $Y|x_1$ urn. If the X-ball drawn is of color 0, then once drawn its true color is revealed, x_2. We add a ball of color x_2 to the X-urn and create a $Y|x_2$ urn with $\mu(\mathcal{Y}, x_2)$ balls of color 0. This process is repeated, so that a new $Y|x$ urn is created for each new value of X that is observed.

Note that if $\mathbf{P} \sim EDP(\alpha, \mu)$ and the random vectors $(X_1, Y_1), ...(X_n, Y_n)$ given $\mathbf{P} = P$ are i.i.d and distributed according to P, then $\{(X_n, Y_n)\}_{n\in\mathbb{N}}$ is an enriched Pólya sequence. Conversely, the following theorem proves that if $\{(X_n, Y_n)\}_{n\in\mathbb{N}}$ is an Enriched Pólya sequence, then given a random probability measure $\mathbf{P} = P$, the random vectors $(X_1, Y_1), ...(X_n, Y_n)$ are i.i.d and distributed according to P where the joint distribution of $(\mathbf{P}_X, \mathbf{P}_{Y|X})$ defined from any fixed version of the conditional satisfies

the four conditions in defintion (2). Therefore, in addition to the fact that the de Finetti measure of an Enriched Pólya sequence is an Enriched Dirichlet Process, this theorem also shows that the induced law of the random joint from the four conditions in defintion (2) still maintains those properties even though the mapping is not $1-1$.

Theorem 4. *If $\{(X_n, Y_n)\}_{n\in\mathbb{N}}$ is an Enriched Pólya sequence with parameters α and μ, then $\{(X_n, Y_n)\}_{n\in\mathbb{N}}$ is an exchangeable sequence and its de Finetti measure is an Enriched Dirichlet Process with parameters (α, μ).*

For a quick sketch of the proof, we start by showing that the sequence $\{(X_n, Y_n)\}_{n\in\mathbb{N}}$ is exchangeable, and then apply de Finetti's Theorem. Next, after reparametrizing in terms of the marginal and conditonals, we verify the de Finetti measure satisfies the four conditions in the definition of the EDP. A detailed proof is given in the Appendix.

4.2 Properties

Define $P_{0X}(\cdot) = \frac{\alpha(\cdot)}{\alpha(\mathcal{X})}$ and for every $x \in \mathcal{X}$, $P_{0Y|X}(\cdot|x) = \frac{\mu(\cdot,x)}{\mu(\mathcal{Y},x)}$. From well-known properties of the Dirichlet distribution, we have:

Proposition 1. *If $\mathbf{P} \sim EDP(\alpha, \mu)$, for $A \in \mathcal{B}_X, B \in \mathcal{B}_Y$,*

- $E[\mathbf{P}_X(A)] = P_{0X}(A)$; $\quad Var(\mathbf{P}_X(A)) = \frac{P_{0X}(A)(1-P_{0X}(A))}{\alpha(\mathcal{X})+1}$.
- $\forall x \in \mathcal{X}$, $E[\mathbf{P}_{Y|X}(B \mid x)] = P_{0Y|X}(B|x)$;
  $Var(\mathbf{P}_{Y|X}(B|x)) = \frac{P_{0Y|X}(B|x)(1-P_{0Y|X}(B|x))}{\mu(\mathcal{Y},x)+1}$.
- $E[\mathbf{P}(A \times B)] = \int_A P_{0Y|X}(B|x) dP_{0X}(x) := P_0(A \times B)$.

Therefore, similar to the DP, the location of the EDP is determined by the base measure P_0, but the there are now many more parameters to control the precision, namely $\alpha(\mathcal{X})$ and $\mu(\mathcal{Y}, x)$ for every $x \in \mathcal{X}$. The following proposition states that the DP is in fact a special case of the EDP.

Proposition 2. *$\mathbf{P} \sim EDP(\alpha, \mu)$ with $\mu(\mathcal{Y}, x) = \alpha(x)$, $\forall x \in \mathcal{X}$ is equivalent to $\mathbf{P} \sim DP(\alpha(\mathcal{X})P_0)$.*

The proof relies on the urn characterization of both processes; we show that an Enriched Pólya sequence is equivalent to a Pólya sequence with parameter $\alpha(\mathcal{X})\, P_0(\cdot)$, if $\mu(\mathcal{Y}, x) = \alpha(x)$, $\forall x \in \mathcal{X}$. A more detailed proof is given in the Appendix.

As a by-product of this proposition, if $\mathbf{P} \sim DP(\alpha(\mathcal{X})P_0)$, the law of the random conditionals is $\mathbf{P}_{Y|X}(\cdot|x) \sim DP(\alpha(x)P_{0Y|X}(\cdot|x))$, where $\mathbf{P}_{Y|X}(\cdot|x)$ are independent among $x \in \mathcal{X}$. In general, the marginal base measure can assign positive mass to countably many locations. Any random conditional probability measure associated with x that has positive mass under the marginal base measure will be a DP with precision parameter equivalent to the mass under the marginal base measure times α. Since a

DP with precision parameter 0 is degenerate on a random location with probability one, the random conditional probability measures associated with all other x's will be degenerate at some $y \in \mathcal{Y}$ with probability one. Thus, in the case when P_0 is non-atomic, a DP implies assuming the conditionals are independent and degenerate a.s., which is consistent with results in Ramamoorthi and Sangalli (2006).

As noted by Ferguson (1973), a prior for nonparametric problems should have large topological support. The following theorem shows that the EDP has full weak support. Here, $\mathcal{X} = \mathbb{R}^{k_1}$ and $\mathcal{Y} = \mathbb{R}^{k_2}$, implying $\mathcal{X} \times \mathcal{Y} = \mathbb{R}^k$ where $k = k_1 + k_2$.

Theorem 5. *Let S_0 denote the topological support of P_0. If $\mathbf{P} \sim EDP(\alpha, \mu)$, then the topological support of $\mathbf{P}$ is*

$$M_0 = \{P \in \mathcal{P}(\mathcal{B}) : \textit{topological support}(P) \subseteq S_0\}.$$

4.3 Posterior

Just as the finite dimensional Enriched Dirichlet distribution is conjugate to the multinomial likelihood, the Enriched Dirichlet Process is also conjugate for estimating an unknown distribution from exchangeable data. More precisely,

Proposition 3. *If $(X_i, Y_i) \mid \mathbf{P} = P \overset{iid}{\sim} P$, where $\mathbf{P} \sim EDP(\alpha, \mu)$, then*

$$\mathbf{P} \mid x_1, y_1, ..., x_n, y_n \sim EDP(\alpha_n, \mu_n),$$

where $\alpha_n = \alpha + \sum_{i=1}^n \delta_{x_i}$,

$$\forall x \in \mathcal{X} \quad \mu_n(\cdot, x) = \mu(\cdot, x) + \sum_{j=1}^{n_x} \delta_{y_{x,j}}; \quad n_x = \sum_{i=1}^{n} \delta_{x_i}(x), \quad \{y_{x,j}\}_{j=1}^{n_x} = \{y_j : x_j = x\}.$$

The proof of conjugacy is straightforward; one simply has to demonstrate that given the random sample the four conditions in the definition of EDP hold with the updated parameters specified above. The first two conditions, the fact that the marginal and conditionals are DPs with updated parameters, follow from conjugacy of the DP. The last two conditions, independence of the marginal and conditionals and independence among the conditionals, follow by combining the fact that a priori independence holds with independence of the random vectors $(X_1, ..., X_n)$ and $(Y_1, ..., Y_n | X_1 = x_1, ..., X_n = x_n)$ and independence of the random vectors $\{Y_{x,j}\}_{j=1}^{n_x}$ among $x \in \mathcal{X}$.

Posterior consistency is a frequentist validation tool that is useful in Bayesian nonparametric inference where the infinite dimension of the parameter space can make specification of a prior challenging and cause the prior to strongly influence the posterior even with large amounts of data. One of the reasons that makes the Dirichlet Process so appealing is that the posterior is weakly consistent for any probability measure, Π, on the product space under the assumption that the sequence of random vectors are distributed according to the i.i.d. product measure Π^∞. Another important property that the EDP maintains is posterior consistency. The proof requires that for a set

$A \times B \in \mathcal{B}_X \times \mathcal{B}_Y$, the posterior expectation of $\mathbf{P}(A \times B)$ converges to $\Pi(A \times B)$ a.s. Π^∞ and its posterior variance goes to zero. In the following lemma, the variance of the probability over a set $A \times B \in \mathcal{B}_X \times \mathcal{B}_Y$ is specified.

Lemma 1. *If* $\mathbf{P} \sim EDP(\alpha, \mu)$, *for* $A \times B \in \mathcal{B}_X \times \mathcal{B}_Y$,

$$Var(\mathbf{P}(A \times B)) = \frac{1}{\alpha(\mathcal{X}) + 1} \int_A \frac{P_{0Y|X}(B|x)(1 + \mu(\mathcal{Y}, x)P_{0Y|X}(B|x))}{\mu(\mathcal{Y}, x) + 1} dP_{0X}(x) \quad (I_1)$$

$$+ \frac{\alpha(\mathcal{X})}{\alpha(\mathcal{X}) + 1} \int_A \int_{\{x\}} \frac{P_{0Y|X}(B|x)(1 - P_{0Y|X}(B|x))}{\mu(\mathcal{Y}, x) + 1} dP_{0X}(x') dP_{0X}(x) \quad (I_2)$$

$$- \frac{1}{\alpha(\mathcal{X}) + 1} \int_A \int_{\{x\}} P_{0Y|X}(B|x)^2 dP_{0X}(x') dP_{0X}(x) \quad (I_3)$$

$$- \frac{1}{\alpha(\mathcal{X}) + 1} \int_A \int_{A \setminus \{x\}} P_{0Y|X}(B|x') P_{0Y|X}(B|x) dP_{0X}(x') dP_{0X}(x). \quad (I_4)$$

Theorem 6. *If* $\mathbf{P} \sim EDP(\alpha, \mu)$, *then, for* $\Pi \in \mathcal{P}(\mathcal{B})$, *the posterior distribution,* Q_n, *of* $\mathbf{P}$ *converges weakly to* δ_Π *for* $n \to \infty$, *a.s.* Π^∞.

The proofs are given in the Appendix.

4.4 Square-Breaking Construction

The following square-breaking representation of the EDP is a direct result of Sethuraman's stick-breaking representation of the DP (Sethuraman (1994)).

Proposition 4. *If* $\mathbf{P} \sim EDP(\alpha, \mu)$, *it has the following square-breaking a.s. representation:*

$$\mathbf{P} = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \pi_i^X \pi_{j|i}^Y \delta_{X_i^*, Y_{j|i}^*},$$

where: $\pi_1^X = V_1^X$; $\pi_i^X = V_i^X \prod_{h=1}^{i-1}(1 - V_h^X)$, *with*

$$V_i^X \overset{i.i.d}{\sim} beta(1, \alpha(\mathcal{X})), \quad X_i^* \overset{i.i.d}{\sim} P_{0X},$$

and for $i = 1, 2, \ldots$: $\pi_{1|i}^Y = V_{1|i}^Y$; $\pi_{j|i}^Y = V_{j|i}^Y \prod_{h=1}^{j-1}(1 - V_{h|i}^Y)$, *with*

$$V_{j|i}^Y | X_i^* = x_i^* \overset{i.i.d}{\sim} beta(1, \mu(\mathcal{Y}, x_i^*)), \quad Y_{j|i}^* | X_i^* = x_i^* \overset{i.i.d}{\sim} P_{0Y|X}(\cdot | x_i^*),$$

and the sequences $\{V_i^X\}_{i=1}^\infty$, $\{X_i^*\}_{i=1}^\infty$, $\left\{V_{j|1}^Y | X_1^* = x_1^*\right\}_{j=1}^\infty$, $\left\{V_{j|2}^Y | X_2^* = x_2^*\right\}_{j=1}^\infty$, ... *and* $\left\{Y_{j|1}^* | X_1^* = x_1^*\right\}_{j=1}^\infty$, $\left\{Y_{j|2}^* | X_2^* = x_2^*\right\}_{j=1}^\infty$, ... *are independent.*

For an alternative view of this proposition, consider a square of area one; we break off rectangles of the square defined by a width of π_i^X and length of $\pi_{j|i}^Y$ and we assign the area of that rectangle, $\pi_i^X \pi_{j|i}^Y$, to a random location $(X_i^*, Y_{j|i}^*)$.

Note that while a closed form for the finite dimensional distributions of $\mathbf{P}_Y$ may not be available, we can obtain a square-breaking construction for the random marginal probability measure on $(\mathcal{Y}, \mathcal{B}_Y)$,

$$\mathbf{P}_Y = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \pi_i^X \pi_{j|i}^Y \delta_{Y^*_{j|i}},$$

where the distribution of $\{\pi_i^X\}_{i=1}^{\infty}$, $\left\{\pi_{j|i}^Y\right\}_{i,j=1}^{\infty}$, $\left\{Y^*_{j|i}\right\}_{i,j=1}^{\infty}$ is specified above.

4.5 Clustering Structure

The clustering structure in a sample from $\mathbf{P} \sim EDP$ is characterized by the predictive rule. In particular, the predictive rule states that if P_0 is non-atomic, for $A \times B \in \mathcal{B}_X \times \mathcal{B}_Y$:

$$\begin{aligned} &Pr(X_{n+1} \in A, Y_{n+1} \in B | x_1, y_1, ..., x_n, y_n) \\ &\quad = \frac{\alpha(\mathcal{X})}{\alpha(\mathcal{X}) + n} P_0(A \times B) + \sum_{x_i^* \in A} \frac{n_i}{\alpha(\mathcal{X}) + n} \left(\frac{\mu(B, x_i^*) + \sum_{j=1}^{n_i} \delta_{y_{ij}}(B)}{\mu(\mathcal{Y}, x_i^*) + n_i} \right). \end{aligned}$$

Thus, the pair (X_{n+1}, Y_{n+1}) is either a "new-new" , "old-new" , or "old-old" pair with probabilities obtained by replacing the set $A \times B$ with the sets $(\mathcal{X} \setminus \{x_1, ..., x_n\}) \times (\mathcal{Y} \setminus \{y_1, ..., y_n\})$, $\{x_1, ..., x_n\} \times (\mathcal{Y} \setminus \{y_1, ..., y_n\})$, or $\{x_1, ..., x_n\} \times \{y_1, ..., y_n\}$ respectively. Let $(x_1^*, ..., x_{d_n}^*)$ be the unique values of $(x_1, ..., x_n)$ where d_n is the number of unique values and $(y_{i,1}^*, ..., y_{i,d_{n_i}}^*)$ be the unique values of $(y_{i,1}, ..., y_{i,n_i})$ where d_{n_i} is the number of unique values in this set. Succinctly, the clustering structure is described as follows:

$$X_{n+1}, Y_{n+1} = \begin{cases} \text{new-new}, (X_{n+1}, Y_{n+1}) \sim P_0 & \text{wp } \frac{\alpha(\mathcal{X})}{\alpha(\mathcal{X})+n}; \\ \text{old-new}, x_i^*, i = 1, ..., d_n, Y_{n+1} \sim P_{0Y|X}(\cdot | x_i^*) & \text{wp } \frac{n_i}{\alpha(\mathcal{X})+n} \frac{\mu(\mathcal{Y}, x_i^*)}{\mu(\mathcal{Y}, x_i^*)+n_i}; \\ \text{old-old}, x_i^*, i = 1, ..., d_n, y_{i,j}^*, j = 1, ..., d_{n_i} & \text{wp } \frac{n_i}{\alpha(\mathcal{X})+n} \frac{n_{i,j}}{\mu(\mathcal{Y}, x_i^*)+n_i}. \end{cases}$$

This gives a "two-level" clustering which reduces to the global clustering of the DP if $\mu(\mathcal{Y}, x) = 0$ for all $x \in \mathcal{X}$.

4.6 Comparison with different approaches

In recent literature, there have been many proposals of generalizations of the Dirichlet process, particularly, dependent Dirichlet Processes. Such an approach exploits marginal conditional independence. One considers a collection of random variables $\{Y_j, j \in \mathcal{J}\}$ and assumes that they are conditionally independent, that is, for any $j_1, \ldots, j_m \in \mathcal{J}$, $Y_{j_1}, \ldots, Y_{j_m} \mid F_{j_1}, \ldots, F_{j_m} \sim \prod_{i=1}^m F_{j_i}(\cdot)$. Then, a prior is given on the family of random distributions $\{\mathbf{F}_j, j \in \mathcal{J}\}$, such that the $\mathbf{F}_j$'s are dependent.

A first proposal along these lines was given by Cifarelli and Regazzini (1978), who assumed that $\mathbf{F}_j \mid \lambda \overset{i.i.d}{\sim} DP(\alpha F_0(\cdot; \lambda))$, with $\lambda \sim H(\lambda)$. A very interesting development is the Hierarchical Dirichlet Process proposed by Teh et al. (2006), who model

the random base measure $\mathbf{F}_0$ nonparametrically, assuming $\mathbf{F}_0 \sim DP(\gamma H)$. A further development is the Nested Dirichlet Process (Rodriguez et al. (2006)) where the model is given as $\mathbf{F}_j \mid G \overset{i.i.d}{\sim} G$ and $\mathbf{G} \sim DP(\alpha DP(\gamma H))$. The general and clever scheme given by the Dependent Dirichlet Processes (DDP; MacEachern (1999)), induces dependence across the $\mathbf{F}_j$'s by exploiting the stick breaking representation of the Dirichlet process and by assuming dependent weights and atoms along j.

Dependent DPs define the law of a collection of distribution functions $\{\mathbf{F}_j, j \in \mathcal{J}\}$ indexed by a non random covariate. If we simply replace $\mathcal{J}$ with $\mathcal{X}$, this does not necessarily define the law of the conditionals. In particular, since the covariate is non random, no σ-algebra on $\mathcal{X}$ is considered, and thus, measurability with respect to $\mathcal{B}_X$ a.s. is not required. If measurability with respect $\mathcal{B}_X$ a.s. is satisfied, this is a model on the random conditionals and does not induce a prior on the random joint distribution of (X, Y).

Instead, our approach gives a prior on the marginal-conditional pair and induces a prior on the joint. For a Dirichlet Process with non atomic base measure, the random conditionals are independent and degenerate a.s. We are extending this by allowing for non degenerate conditionals, but we will assume independence. A further extension would allow for dependence among the random conditionals through a dependent Dirichlet Process if measurability with respect to $\mathcal{B}_X$ a.s. is satisfied. However, some properties will be lost. For example, for a DDP, we would lose conjugacy, and the model would become much more complex, and using the Hierarchical DP or the Nested DP would remove dependence on x in the base measures for the conditionals.

Notice that the distribution of the conditional also as a random function of X is $\mathbf{P}_{Y|X}(\cdot|X) \sim \sum_{i=1}^{\infty} \pi_i^X \delta_{\mathbf{P}_{Y|X}(\cdot|X_i^*)}$. This resembles the prior for the Nested Dirichlet Process, but is not directly comparable since $\mathbf{P}_{Y|X}(\cdot|X)$ is a different object than $\{\mathbf{F}_j, j \in \mathcal{J}\}$.

5 Example

We provide an illustration of the properties of the EDP prior in an application to mixture models. The problem we consider is comparing different schools based on national test scores. The dataset we analyze contains two different test scores for students in 65 inner-London schools. The first score is based on the London Reading Test (LRT), taken at age 11, and the second is a score derived from the Graduate Certificate of Secondary Education (GCSE) exams in a number of different subjects, taken at age 16. Taking into account earlier LRT scores can give a sense of the "value added" for each school. To answer the question of which schools are most effective, we consider modeling the relationship between LRT and GCSE for all schools. The data are available at `http:// www.stata-press.com/data/mlmus.html`. School number 48 is dropped from the dataset since only 2 students were observed.

Rabe-Hesketh and Skrondal (2005) (Chapter 4) study the following multilevel parametric model where Y_{ij} and X_{ij} represent, respectively, the GCSE and LRT score for

student i in school j:

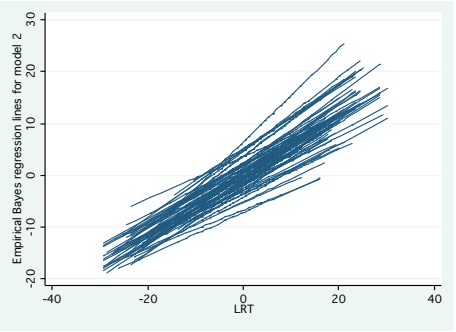
$$Y_{ij} \mid \beta_{0j}, \beta_{1j}, x_{ij} \overset{indep}{\sim} N(\beta_{0j} + \beta_{1j} x_{ij}, \sigma^2), \qquad (6)$$
$$\begin{bmatrix} \beta_{0j} \\ \beta_{1j} \end{bmatrix} \overset{i.i.d}{\sim} N_2 \left(\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \Sigma_\beta \right),$$

where β_{0j} and β_{1j} are independent of X_{ij}. The interest is in estimating the school specific coefficients $\beta_j = (\beta_{0j}, \beta_{1j})$. The intercept is interpreted as the school mean of GCSE scores for the students with the average LRT score of 0. The competitiveness of the school is captured by the school specific slope. Schools with greater slopes are competitive; more "value" is added for students with higher LRT scores. Schools with a slope of 0 are non-competitive; the performance of students is homogeneous regardless of how the students scored on the LRT. If parents are to choose the best school for their children, both average "value added" and competitiveness are important.

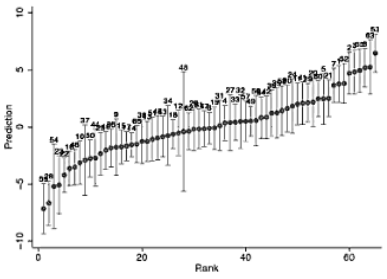
Maximum likelihood estimates of the parameters of the mixing distribution (Rabe-Hesketh and Skrondal (2005)) give $\hat{\beta}_0 = -.115$, with standard error $SE(\hat{\beta}_0) = .0199$, and $\hat{\beta}_1 = .55$, with $SE(\hat{\beta}_1) = .3978$, and estimated covariance matrix:

$$\hat{\Sigma}_\beta = \begin{bmatrix} 9.04 & .18 \\ .18 & .0145 \end{bmatrix}.$$

Empirical Bayes predictions of school specific intercept and slope were then obtained; figures (1a) and (1b) show the plots of estimated regression lines for each school and ranking of schools based on the intercept.



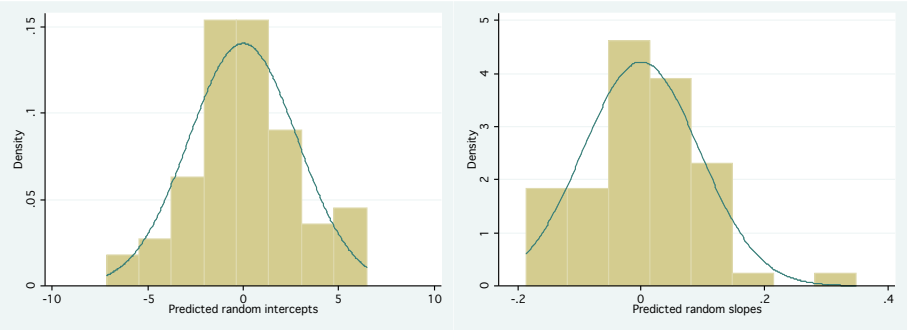
(a) Empirical Bayes Predictions of school-specific regression line



(b) Ranking of Schools

Figure 1: Results of Linear Mixed Effect model

By visual inspection of the histograms of the empirical Bayes estimates in figures (2a) and (2b), for the intercept and especially the slope, a normal distribution does not fit well. This may be due to the fact that there are only 65 schools, that the normality assumption does not hold or a combination of the two. To enlarge the class of models, we can consider modelling the mixing distribution of the intercept and slope



(a) Predicted random intercept (b) Predicted random slope

Figure 2: Assessing the model

nonparametrically. A pitfall of model (6) is that it assumes the same variability for all schools. In fact, the wide range of the naive OLS estimates of within school variance (not shown) supports a model which allows for school-specific variance.

Bayesian nonparametric extensions of this model would assign a DP prior on the mixing distribution of the (β_{0j}, β_{1j})'s (a DP-location mixture), assuming the same variance σ^2 for each school, or model school specific variances σ_j^2, with a DP prior for the latent distribution of $(\beta_{0j}, \beta_{1j}, \sigma_j^2)$ (DP scale-location mixture). The EDP is an intermediate choice. It may model clusters of schools that share the same variance, with different β's inside each cluster. We assume that

- $Y_{ij}|x_{ij}, \beta_{0j}, \beta_{1j}, \sigma_j^2 \overset{indep}{\sim} N(\beta_{0j} + \beta_{1j}x_{ij}, \sigma_j^2)$,
- $\beta_j, \sigma_j^2|\mathbf{P}_{\beta,\sigma^2} = P_{\beta,\sigma^2} \overset{i.i.d}{\sim} P_{\beta,\sigma^2}$ where $\beta_j = (\beta_{0j}, \beta_{1j})$,
- $\mathbf{P}_{\beta,\sigma^2} \sim EDP(\alpha, \mu)$ where $\alpha = \alpha_{\sigma^2} P_{0,\sigma^2}$ and, for all $\sigma^2 \in \mathbb{R}_+$, $\mu(\cdot, \sigma^2) = \mu_\beta(\sigma^2) P_{0,\beta|\sigma^2}(\cdot|\sigma^2)$.

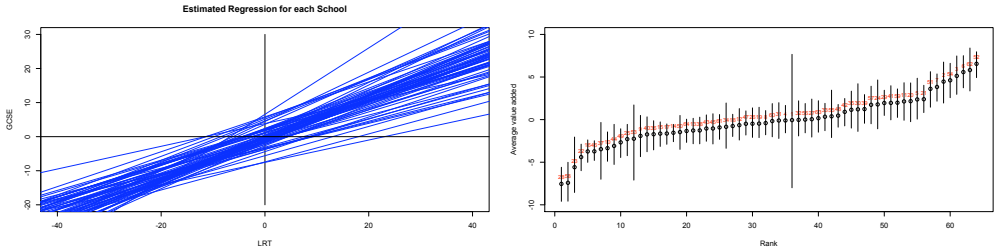
In the analysis reported below, we fixed the baseline measures $P_{0\sigma}$ as an Inverse-Gamma, with rate and shape parameters, respectively, 8 and 385, and $P_{0,\beta|\sigma^2}(\cdot|\sigma^2)$ as a bivariate Normal, $N_2(\mu_0, k_0\,\sigma^2\ \Sigma_0)$, with $\mu_0 = [0, .5]'$, $k_0 = 1/20$ and $\Sigma_0 = \begin{bmatrix} 9 & 3/16 \\ 3/16 & 1/64 \end{bmatrix}$.

Notice that if the precision parameter $\alpha_{\sigma^2} \approx 0$, we get back to a DP location mixture, and if the precision parameters $\mu_\beta(\sigma^2) \approx 0$ for all $\sigma^2 \in \mathbb{R}_+$, we get a DP scale-location mixture. Thus, with an EDP prior we can express uncertainty between homoskedasticity and heteroskedasticity.

We model uncertainty about α_{σ^2} and $\mu_\beta(\sigma^2)$ through Gamma hyperpriors: $\alpha_{\sigma^2} \sim Ga(u_\alpha, v_\alpha)$, where we choose $u_\alpha = 2$ and $v_\alpha = 1$, and for all $\sigma^2 \in \mathbb{R}_+$ $\mu_\beta(\sigma^2) \overset{i.i.d}{\sim}$

$Ga(u_{\mu_\beta}, v_{\mu_\beta})$, with $u_{\mu_\beta} = 2$ and $v_{\mu_\beta} = 1$.

The MCMC scheme to compute posterior distributions is based on the algorithm 6 described in Neal (2000), which is a Metropolis-Hastings algorithm with candidates drawn from the prior. Resampling the precision parameters is done by introducing a latent beta-distributed variable, as described in Escobar and West (1995). The number of iterations is set up to $20,000$ with 10% of burn-in. Looking at the trace and auto-correlation plots, convergence appears reached for the β's in all schools and for σ^2's in most schools. The results are summarized in Figures (3a) and (3b), which display the estimated regression line for each school and the ranking of schools based on average "value added " with empirical quantiles.



(a) Estimated regression line for each school

(b) Ranking of Schools based on average value added with empirical quantile

Figure 3: Results of EDP model

The MCMC posterior expectation of α_{σ^2} is 2.5, and Figure (4) depicts the estimated posterior values of $\mu_\beta(\sigma^2)$ for different values of σ^2.

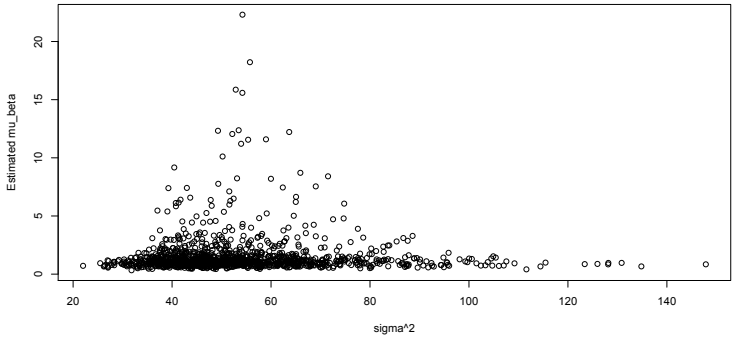


Figure 4: Estimated posterior values of $\mu_\beta(\sigma^2)$ for different values of σ^2

Neither $\alpha_{\sigma^2} \approx 0$ nor $\mu_\beta(\sigma^2) \approx 0$ for all σ^2, and interestingly, the estimated values of $\mu_\beta(\sigma^2)$ are high for values of σ^2 which are more likely a posteriori, and close to zero

for unlikely values of σ^2. Thus, the results favor a model which allows for homoskedacity among some schools with a more likely value σ^2 and some outlying schools with abnormally large or small variances.

6 Final remarks

We have proposed an *enrichment* of the DP starting from the idea of enriched conjugate priors. The advantages of this process are that it allows for more flexible specification of prior information, includes the DP as a special case, and retains some desirable properties including conjugacy and the fact that it can be constructed from an enriched urn scheme. The disadvantages include the difficulty in obtaining a closed form for the distribution of the joint probability over a given set and for the distribution of the marginal probability over a measurable subset of $\mathcal{Y}$. Using an EDP as the prior for the distribution of a random vector, Z, implies one has to determine a partition of Z into two groups and an ordering defining which group comes first. The "two-level" clustering resulting from the EDP introduces a clear asymmetry based on the partition and ordering chosen, and how to choose them depends on the application. There may be a natural ordering or partition and/or computational reasons, including decomposition of the base measure, for choosing the partition and ordering. In our example, we partitioned the random vector $(\beta_0, \beta_1, \sigma^2)$ into the two groups, (σ^2) and (β_0, β_1), with σ^2 chosen first due to uncertainty in homoskedasticity and decomposition of the conjugate normal-inverse gamma base measure. One may also examine all plausible and interesting partitions and orderings.

We have focused on the partition of the random vector into two groups, but most results could be extended to any finite partition of the random vector, although this would of course imply a further nested structure. Other future works include examining the implied clustering structure in regression settings when the joint model is an EDP mixture and exploring if other conjugate nonparametric priors whose finite dimensionals are standard conjugate priors can be generalized starting from enriched conjugate priors, such as extension of the enriched distribution, mentioned in the Remark 2, to an enriched bivariate Neutral to the Right Processes.

We hope that having explored these features can shed light on potentialities and limitations and encourage further developments in constructing more flexible priors for a random probability measure on $\mathbb{R}^k$.

Appendix

Proof of Theorem 1 From the predictive distribution, it follows that the joint distribution can be expressed as:

$$Pr(X_1 = i_1, Y_1 = j_1, ..., X_n = i_n, Y_n = j_n) = \prod_{l=1}^{n} \frac{\alpha(i_l) + \sum_{h=1}^{l-1} \delta_{i_h}(i_l)}{\alpha(\mathcal{X}) + l - 1} \frac{\mu(j_l, i_l) + \sum_{h=1}^{l-1} \delta_{j_h, i_h}(j_l, i_l)}{\mu(\mathcal{Y}, i_l) + \sum_{h=1}^{l-1} \delta_{i_h}(i_l)},$$

which can be equivalently expressed as:

$$\frac{\Gamma(\alpha(\mathcal{X}))}{\prod_{i=1}^{k}\Gamma(\alpha(i))}\frac{\prod_{i=1}^{k}\Gamma(\alpha(i)+n_{i+})}{\Gamma(\alpha(\mathcal{X})+n)}\prod_{i=1}^{k}\frac{\Gamma(\mu(\mathcal{Y},i))}{\prod_{j=1}^{m}\Gamma(\mu(j,i))}\prod_{i=1}^{k}\frac{\prod_{j=1}^{m}\Gamma(\mu(j,i)+n_{ij})}{\Gamma(\mu(\mathcal{Y},i)+n_{i+})}. \tag{7}$$

The joint distribution only depends on the number of unique pairs seen, not on the order in which they are observed. Thus, the pairs $\{X_n, Y_n\}_{n\in\mathbb{N}}$ form an exchangeable sequence. By de Finetti's Representation Theorem, there exists a probability measure $\tilde{Q}$ on the simplex $S_{k,m} = \{p_{1,1}, ..., p_{k,m} : p_{i,j} \geq 0 \text{ and } \sum_{i=1}^{k}\sum_{j=1}^{m} p_{i,j} = 1\}$ such that:

$$Pr(X_1 = i_1, Y_1 = j_1, ..., X_n = i_n, Y_n = j_n) = \int_{[0,1]^{km}} \prod_{i=1}^{k}\prod_{j=1}^{m} p_{i,j}^{n_{i,j}} \tilde{Q}(dp_{1,1}, ..., dp_{k,m}).$$

Define the simplexes $S_k = \{p_{1+}, ..., p_{k+} : p_{i+} \geq 0 \text{ and } \sum_{i=1}^{k} p_{i+} = 1\}$ and $S_m^{(i)} = \{p_{i|1}, ..., p_{i|k} : p_{j|i} \geq 0 \text{ and } \sum_{j=1}^{m} p_{j|i} = 1\}$ for $i = 1, ...k$. Let Q be the probability measure on the product of the simplexes $S_k \times \prod_{i=1}^{k} S_m^{(i)}$ obtained from $\tilde{Q}$ via a reparametrization in terms of $(\mathbf{p}_{1+}, ..., \mathbf{p}_{k+}, \mathbf{p}_{1|1}, ..., \mathbf{p}_{m|k})$. Then,

$$\begin{aligned} &Pr(X_1 = i_1, Y_1 = j_1, ..., X_n = i_n, Y_n = j_n) \\ &\quad = \int_{[0,1]^k \times [0,1]^{km}} \prod_{i=1}^{k} p_{i+}^{n_{i+}} \prod_{j=1}^{m} p_{j|i}^{n_{ij}} Q(dp_{1+}, ..., dp_{m|k}). \end{aligned} \tag{8}$$

Since the Dirichlet distribution is determined by its moments, combining equations (7) and (8) implies that

$$\begin{aligned} &\mathbf{p}_{1+}, ..., \mathbf{p}_{k+} \sim \text{Dir}(\alpha(1), ..., \alpha(k)), \\ &\mathbf{p}_{1|i}, ..., \mathbf{p}_{m|i} \sim \text{Dir}(\mu(1,i), ..., \mu(m,i)) \quad i = 1, ..., k, \end{aligned}$$

where $(\mathbf{p}_{1+}, ..., \mathbf{p}_{k+})$, $(\mathbf{p}_{1|1}, ..., \mathbf{p}_{m|1})$,..., and $(\mathbf{p}_{1|k}, ..., \mathbf{p}_{m|k})$ are independent.

The second part of the theorem follows from de Finetti's results on the asymptotic behavior of the predictive distributions for exchangeable sequences; see Cifarelli and Regazzini (1996). □

Proof of Theorem 4. We start by noting that the sequence $\{X_n\}_{n\in\mathbb{N}}$ is a Pólya sequence with parameter α. Recall that the predictive distribution of a Pólya sequence converges to a discrete random probability measure with positive mass at the countable number of unique values of the sequence almost surely with respect to the exchangeable law. Therefore, given $X_1 = x_1, ..., X_n = x_n$ and letting $U(x_1, ..., x_n)$ denote the set of the unique values of $\{x_1, ..., x_n\}$, we have that for $x^* \in U(x_1, ..., x_n)$, $n_{x^*} = \sum_{i=1}^{n} \delta_{x^*}(x_i) \to \infty$ as $n \to \infty$ almost surely with respect to the exchangeable law. This implies that given $\{X_n = x_n\}_{n\in\mathbb{N}}$, for any $x^* \in U(\{x_n\}_{n\in\mathbb{N}})$, the set of random variables, $\{Y_{x^*,j}\} = \{Y_i : X_i = x^*, i \in \mathbb{N} | \{X_n = x_n\}_{n\in\mathbb{N}}\}$ is a countable sequence. Furthermore, by assumption, for $x_1^* \neq x_2^* \in U(\{x_n\}_{n\in\mathbb{N}})$, the sequences $\{Y_{x_1^*,j}\}_{j\in\mathbb{N}}$ and $\{Y_{x_2^*,j}\}_{j\in\mathbb{N}}$ are independent Pólya sequences with parameters $\mu(\cdot, x_1^*)$ and $\mu(\cdot, x_2^*)$ respectively. These observations imply exchangeability of the sequence $\{X_n, Y_n\}_{n\in\mathbb{N}}$, as

shown in the following argument.

$$Pr(X_1 \in A_1, Y_1 \in B_1, ..., X_n \in A_n, Y_n \in B_n) \tag{9}$$
$$= \int_{A_1 \times ... \times A_n} Pr(Y_1 \in B_1, ..., Y_n \in B_n | x_1, ..., x_n) dPr(x_1, ..., x_n).$$

By independence of $\{Y_{x_1^*,j}\}_{j=1}^{n_{x_1^*}}$ and $\{Y_{x_2^*,j}\}_{j=1}^{n_{x_2^*}}$ for $x_1^* \neq x_2^* \in U(x_1, ..., x_n)$, we have that (9) is equal to:

$$\int_{A_1 \times ... \times A_n} \prod_{x^* \in U(x_1,...,x_n)} Pr(Y_{x^*,1} \in B_{x^*,1}, .., Y_{x^*,n_{x^*}} \in B_{x^*,n_{x^*}}) dPr(x_1, ., x_n). \tag{10}$$

A permutation, π, of the sets $(x_1 \times B_1), ..., (x_n \times B_n)$, is equivalent to the same permutation, π, of $(x_1, ..., x_n)$ and for $x^* \in U(x_{\pi(1)}, ..., x_{\pi(n)})$, a permutation, γ_{x^*}, of $(B_{x^*,1}, ..., B_{x^*,n_{x^*}})$. To keep notation concise, we will let $U_{\pi,n}$ represent $U(x_{\pi(1)}, ..., x_{\pi(n)})$ (and similarly, U_n represent $U(x_1, ..., x_n)$).The term inside the integral is invariant to the permutation, π, of $(x_1, ..., x_n)$, and due to exchangeability of Pólya sequences, the laws of the random vectors $\{X_i\}_{i=1}^n$ and $\{Y_{x^*,j}\}_{j=1}^{n_x^*}$ are invariant to the permutations π and γ_{x^*} respectively. Thus, (10) is equal to:

$$\int_{A_{\pi(1)} \times ... \times A_{\pi(n)}} \prod_{x^* \in U_{\pi,n}} Pr(Y_{x^*,1} \in B_{\gamma_{x^*}(1)}, .., Y_{x^*,n_{x^*}} \in B_{\gamma_{x^*}(n_{x^*})}) dPr(x_{\pi(1)}, ..., x_{\pi(n)})$$
$$= \int_{A_{\pi(1)} \times ... \times A_{\pi(n)}} Pr(Y_1 \in B_{\pi(1)}, ., Y_n \in B_{\pi(n)} | x_{\pi(1)}, ., x_{\pi(n)}) dPr(x_{\pi(1)}, ., x_{\pi(n)})$$
$$= Pr(X_1 \in A_{\pi(1)}, Y_1 \in B_{\pi(1)}, ., X_n \in A_{\pi(n)}, Y_n \in B_{\pi(n)}).$$

De Finetti's Representation Theorem states that there exists a random probability measure, $\mathbf{P}$, with distribution $\tilde{Q}$ on $\mathcal{P}(\mathcal{B})$ such that:

$$Pr(X_1 \in A_1, Y_1 \in B_1, ., X_n \in A_n, Y_n \in B_n) = \int_{\mathcal{P}(\mathcal{B})} \prod_{h=1}^{n} P(A_h \times B_h) d\tilde{Q}(P), \tag{11}$$

and $\frac{1}{n} \sum_{h=1}^{n} \delta_{A \times B}(X_h, Y_h) \xrightarrow{d} \mathbf{P}(A \times B)$ a.s. with respect to the exchangeable law as $n \to \infty$ where $\mathbf{P} \sim \tilde{Q}$. The distribution $\tilde{Q}$ determines the joint distribution, Q, of the marginal and a fixed version of the conditionals. Reparametrizing in terms of the marginal and conditionals implies:

$$Pr(X_1 \in A_1, Y_1 \in B_1, ., X_n \in A_n, Y_n \in B_n)$$
$$= \int_{\mathcal{P}(\mathcal{B}_X) \times \mathcal{P}(\mathcal{B}_Y)^{\mathcal{X}}} \prod_{h=1}^{n} \int_{A_h} P_{Y|X}(B_h|x) dP_X(x) dQ(P_X, \prod_{x \in \mathcal{X}} P_{Y|X}(\cdot|x)). \tag{12}$$

A simple application of the results of Blackwell and MacQueen (1973) for Pólya urn sequences, verifies that the first two conditions in the definition of the EDP hold. In particular, for any finite partition $A_1, ..., A_k \subseteq \mathcal{B}_{\mathcal{X}}$, define the simple measurable function, $\phi(x) = i$ if $x \in A_i$ for $i = 1, ..., k$. Noting that $\{\phi(X_n)\}_{n \in \mathbb{N}}$, is a Pólya sequence with parameter $\alpha \circ (\phi)^{-1}$ taking values in the finite space $\{1, ..., k\}$, implies:

$$\mathbf{P}_X(\phi^{-1}(1), ..., \mathbf{P}_X(\phi^{-1}(k))) \sim \text{Dir}(\alpha(\phi^{-1}(1)), ..., \alpha(\phi^{-1}(k)))$$
$$\Leftrightarrow \mathbf{P}_X(A_1), .., \mathbf{P}_X(A_k) \sim \text{Dir}(\alpha(A_1), .., \alpha(A_k)).$$

Similarly, for any finite partition $B_1, .., B_m \subseteq \mathcal{B}_\mathcal{Y}$, define the simple measurable function $\varphi(y) = j$ if $y \in B_j$. For any $x^* \in U(\{x_n\}_{n\in\mathbb{N}})$, the sequence $\{\varphi(Y_{x^*,j})\}_{j\in\mathbb{N}}$ is a Pólya sequence taking values in the finite space $\{1, ..., m\}$ with parameter $\mu(\varphi^{-1}(\cdot), x^*)$. Again, it follows that:

$$\begin{aligned}&\mathbf{P}_{Y|X}(\varphi^{-1}(1)|x^*), ..., \mathbf{P}_{Y|X}(\varphi^{-1}(m)|x^*) \sim \text{Dir}(\mu(\varphi^{-1}(1), x^*), ..., \mu(\varphi^{-1}(m), x^*))\\ &\Leftrightarrow \mathbf{P}_{Y|X}(B_1|x^*), ..., \mathbf{P}_{Y|X}(B_m|x^*) \sim \text{Dir}(\mu(B_1, x^*), ..., \mu(B_m, x^*)).\end{aligned} \tag{13}$$

The unique values of the Pólya sequence are actually draws from $P_{0X}(\cdot) = \frac{\alpha(\cdot)}{\alpha(\mathcal{X})}$ and can therefore take any value in $\mathcal{X}$. Thus, (13) holds for any $x \in \mathcal{X}$. Finally, we need to show the last two conditions in the defintion of the EDP hold. Exchangeability of the pairs implies exchangeability of the sequence $\{Y_i|X_i = x_i\}_{i\in\mathbb{N}}$. Therefore, by de Finetti's theorem:

$$Pr(Y_1 \in B_1, ..., Y_n \in B_n|x_1, ..., x_n) \tag{14}$$

$$= \int_{\mathcal{P}(\mathcal{B}_Y)^{U_n}} \prod_{x^*\in U_n} \prod_{j=1}^{n_{x^*}} P_{Y|X}(B_{x^*,j}|x^*) dQ_{U_n}^{Y|X}(\prod_{x^*\in U_n} P_{Y|X}(\cdot|x^*)). \tag{15}$$

Independence of the exchangeable sequences $\left\{Y_{x_1^*,j}\right\}_{j\in\mathbb{N}}$ and $\left\{Y_{x_2^*,j}\right\}_{j\in\mathbb{N}}$ for $x_1^* \neq x_2^*$ implies:

$$\begin{aligned}Pr(Y_1 \in B_1, ..., Y_n \in B_n|x_1, ..., x_n) &= \prod_{x^*\in U_n} Pr(Y_{x^*,1} \in B_{x^*,1}, ..., Y_{x^*,n_{x^*}} \in B_{x^*,n_{x^*}})\\ &= \prod_{x^*\in U_n} \int_{\mathcal{P}(\mathcal{B}_Y)} \prod_{j=1}^{n_{x^*}} P_{Y|X}(B_{x^*,j}|x^*) dQ_{x^*}^{Y|X}(P_{Y|X}(\cdot|x^*)).\end{aligned} \tag{16}$$

Comparing (15) and (16) shows that $Q_{U_n}^{Y|X} = \prod_{x^*\in U_n} Q_{x^*}^{Y|X}$. Since the unique values of $\{x_1, ...x_n\}$ are realizations of P_{0X} and can take any value in $\mathcal{X}$, independence of $\left\{\mathbf{P}_{Y|X}(\cdot|x)\right\}_{x\in\mathcal{X}}$ among $x \in \mathcal{X}$ follows. Therefore, (14) can be equivalently written as:

$$Pr(Y_1 \in B_1, ..., Y_n \in B_n|x_1, ..., x_n) = \int_{\mathcal{P}(\mathcal{B}_Y)^\mathcal{X}} \prod_{h=1}^{n} P_{Y|X}(B_h|x_h) d(\prod_{x\in\mathcal{X}} Q_x^{Y|X}(P_{Y|X}(\cdot|x))).$$

Now combining this result with the fact that $\{X_n\}_{n\in\mathbb{N}}$ is an exchangeable sequence implies:

$$\begin{aligned}&Pr(X_1 \in A_1, Y_1 \in B_1, ., X_n \in A_n, Y_n \in B_n)\\ &= \int_{A_1\times\cdot\times A_n} Pr(Y_1 \in B_1, ., Y_n \in B_n|x_1, ., x_n) dPr(x_1, ., x_n)\\ &= \int_{\mathcal{P}(\mathcal{B}_X)} \int_{A_1\times\cdot\times A_n} \int_{\mathcal{P}(\mathcal{B}_Y)^\mathcal{X}} \prod_{h=1}^{n} P_{Y|X}(B_h|x_h) d(\prod_{x\in\mathcal{X}} Q_x^{Y|X}(P_{Y|X}(\cdot|x))) d(\prod_{h=1}^{n} P_X(x_h)) dQ^X(P_X)\\ &= \int_{\mathcal{P}(\mathcal{B}_X)} \int_{\mathcal{P}(\mathcal{B}_Y)^\mathcal{X}} \prod_{h=1}^{n} \int_{A_h} P_{Y|X}(B_h \mid x_h) dP_X(x_h) d(\prod_{x\in\mathcal{X}} Q_x^{Y|X}(P_{Y|X}(\cdot \mid x))) dQ^X(P_X).\end{aligned} \tag{17}$$

Comparing (12) with (17) implies that $Q = Q^X \times \prod_{x\in\mathcal{X}} Q_x^{Y|X}$, i.e independence of $\mathbf{P}_X$ and $\{\mathbf{P}_{Y|X}(\cdot|x)\}_{x\in\mathcal{X}}$. □

Proof of Proposition 2 We show that an Enriched Pólya sequence is equivalent to a Pólya sequence with parameter $\alpha(\mathcal{X})\,P_0(\cdot)$, if $\mu(\mathcal{Y},x)=\alpha(x)$, $\forall x\in\mathcal{X}$. For an Enriched Pólya sequence with parameters α,μ and for $A\in\mathcal{B}_X, B\in\mathcal{B}_Y$, since $\lim_{\mu(\mathcal{Y},x)\to\alpha(x)} Pr(Y_1\in B\mid X_1=x)=P_{0Y|X}(B|x)$, then if $\mu(\mathcal{Y},x)=\alpha(x)$, $\forall x\in\mathcal{X}$, $Pr(X_1\in A, Y_1\in B)=P_0(A\times B)$. The joint predictive distribution is given by,

$$\begin{aligned} &Pr(X_{n+1}\in A, Y_{n+1}\in B|X_1=x_1,Y_1=y_1,...,X_n=x_n,Y_n=y_n)\\ &\quad=\int_A \frac{\mu(B,x)+\sum_{j=1}^{n_x}\delta_{y_{x,j}}(B)}{\mu(\mathcal{Y},x)+n_x}\,d\left(\frac{\alpha(x)+\sum_{i=1}^{n}\delta_{x_i}(x)}{\alpha(\mathcal{X})+n}\right). \end{aligned} \tag{18}$$

Rewriting this as the sum of the integrals over the sets $A\backslash\{x_1,...,x_n\}$ and $A\cap\{x_1,...,x_n\}$ and replacing $\mu(\mathcal{Y},x)$ with $\alpha(x)$, we get (18) is equal to,

$$\begin{aligned} &\frac{\alpha(\mathcal{X})}{\alpha(\mathcal{X})+n}P_0(A\setminus\{x_1,...,x_n\}\times B)+\sum_{x\in A\cap\{x_1,...,x_n\}}\frac{\alpha(x)P_{0Y|X}(B|x)+\sum_{j=1}^{n_x}\delta_{y_{x,j}}(B)}{\alpha(x)+n_x}\frac{\alpha(x)+n_x}{\alpha(\mathcal{X})+n}\\ &=\frac{\alpha(\mathcal{X})}{\alpha(\mathcal{X})+n}P_0(A\times B)+\frac{n}{\alpha(\mathcal{X})+n}\sum_{i=1}^{n}\frac{\delta_{x_i,y_i}(A,B)}{n}. \end{aligned}$$

□

Proof of Theorem 5 This proof is based on the proof of Theorem 3.2.4 in Ghosh and Ramamoorthi (2003). To show M_0 is the topological support - the smallest closed set of measure one - it is enough to show that M_0 is a closed set of measure one, such that for every $\Pi\in M_0$, $Q(U)>0$ for any neighborhood U of Π.

First, we show M_0 is closed. If $P_n\in M_0$, then $P_n(S_0)=1$ for all n and if $P_n \overset{weakly}{\to} P$, then for any closed set $C\in\mathcal{B}$, $\limsup_n P_n(C)\le P(C)$. Together these imply $P(S_0)=1$, or equivalently, $P\in M_0$.

Secondly, the set M_0 has measure one. This follows from the square breaking construction of $\mathbf{P}$. Since $X_i^*, Y_{j|i}^*\sim P_0$ implies $\delta_{X_i^*,Y_{j|i}^*}(S_0)=1$ a.s., $\sum_{i=1}^{\infty}\pi_i^X=1$ a.s., and for all i, $\sum_{j=1}^{\infty}\pi_{j|i}^Y=1$ a.s, then $\mathbf{P}(S_0)=1$ a.s. $(\Leftrightarrow Q(M_0)=1)$.

Lastly, our theorem will be proved if we show that for any $\Pi\in M_0$ and any neighborhood U of Π, $Q(U)>0$. By extension of Proposition 2.5.2 in Ghosh and Ramamoorthi (2003), there exists points $q_{1,j}<...<q_{n_j,j}$ in $\mathbb{R}$ for $j=1,..,k$, and $\delta>0$, such that

$$\begin{aligned} U^*=\Bigg\{&P\in\mathcal{P}(\mathcal{B}):|P(\prod_{j=1}^{k}[q_{i_j,j},q_{i_j+1,j}))-\Pi(\prod_{j=1}^{k}[q_{i_j,j},q_{i_j+1,j}))|<\delta \text{ and}\\ &\Pi(\partial\prod_{j=1}^{k}[q_{i_j,j},q_{i_j+1,j}))=0 \text{ for } i=1,...,n_j, j=1,...,k\Bigg\}\subseteq U. \end{aligned}$$

Define $A_{i_1,..,i_{k_1}} = \prod_{j=1}^{k_1}[q_{i_j,j}, q_{i_j+1,j})$ and $B_{i_1,..,i_{k_2}} = \prod_{j=k_1+1}^{k_2}[q_{i_j,j}, q_{i_j+1,j})$ and without loss of generality, we denote these sets as $A_1, ..., A_N$ and $B_1, ..., B_M$. If $P_0(A_n \times B_m) = 0$, then $\delta_{X_i^*, Y_{j|i}^*}(S_0) = 0$ a.s. and $\mathbf{P}(A_n \times B_m)$ is degenerate 0. In addition, $P_0(A_n \times B_m) = 0$ combined with the facts that $\Pi(\partial A_n \times B_m) = 0$ and $\Pi(S_0) = 1$, imply that $\Pi(A_n \times B_n) = 0$. Therefore, $|\mathbf{P}(A_n \times B_m) - \Pi(A_n \times B_m)| = 0$ a.s.. If $P_0(A_n \times B_m) > 0$, then $\delta_{X_i^*, Y_{j|i}^*}(A_n \times B_m) = 1$ with positive probability. Thus, the square breaking construction implies that $Q(U^*) > 0$. □

Proof of Lemma 1

$$E[\mathbf{P}(A \times B)^2] = E[\sum_{i=1}^{\infty} \pi_i^2 \mathbf{P}_{Y|X}(B|X_i^*)^2 \delta_{X_i^*}(A)] \tag{J_1}$$

$$+ E[\sum_{i=1}^{\infty} \sum_{j \neq i} \pi_i \pi_j \mathbf{P}_{Y|X}(B|X_i^*)^2 \delta_{X_i^*}(A) \delta_{X_j^*}(\{X_i^*\})] \tag{J_2}$$

$$+ E[\sum_{i=1}^{\infty} \sum_{j \neq i} \pi_i \pi_j \mathbf{P}_{Y|X}(B|X_i^*) \mathbf{P}_{Y|X}(B|X_j^*) \delta_{X_i^*}(A) \delta_{X_j^*}(A \setminus \{X_i^*\})]. \tag{J_3}$$

Using the fact that $E_\pi[\sum_{i=1}^{\infty} \pi_i^2] = \frac{1}{\alpha(\mathcal{X})+1}$ and properties of the Dirichlet distribution,

$$\begin{aligned}(J_1) &= E_\pi[\sum_{i=1}^{\infty} \pi_i^2 E_{X^*}[E_{Q^{Y|X}}[\mathbf{P}_{Y|X}(B|X_i^*)^2 | X_i^*] \delta_{X_i^*}(A)]] \\ &= \frac{1}{\alpha(\mathcal{X})+1} \int_A \frac{P_{0Y|X}(B|x)(1 + \mu(\mathcal{Y}, x) P_{0Y|X}(B|x))}{\mu(\mathcal{Y}, x) + 1} dP_{0X}(x).\end{aligned}$$

Now, using the fact that $E_\pi[\sum_{i=1}^{\infty} \sum_{i \neq j} \pi_i \pi_j] = \frac{\alpha(\mathcal{X})}{\alpha(\mathcal{X})+1}$ and, again, properties of the Dirichlet distribution,

$$\begin{aligned}(J_2) &= E_\pi[\sum_{i=1}^{\infty} \sum_{i \neq j} \pi_i \pi_j E_{X^*}[E_{Q^{Y|X}}[\mathbf{P}_{Y|X}(B|X_i^*)^2 | X_i^*] \delta_{X_i^*}(A) \delta_{X_j^*}(\{X_i^*\})]] \\ &= \frac{\alpha(\mathcal{X})}{\alpha(\mathcal{X})+1} \int_A \int_{\{x\}} \frac{P_{0Y|X}(B|x)(1 + \mu(\mathcal{Y}, x) P_{0Y|X}(B|x))}{\mu(\mathcal{Y}, x) + 1} dP_{0X}(x') dP_{0X}(x), \\ (J_3) &= E_\pi[\sum_{i=1}^{\infty} \sum_{i \neq j} \pi_i \pi_j E_{X^*}[E_{Q^{Y|X}}[\mathbf{P}_{Y|X}(B|X_i^*) \mathbf{P}_{Y|X}(B|X_j^*) | X_i^*, X_j^*] \delta_{X_i^*}(A) \delta_{X_j^*}(A \setminus \{X_i^*\})]] \\ &= \frac{\alpha(\mathcal{X})}{\alpha(\mathcal{X})+1} \int_A \int_{A \setminus \{x\}} P_{0Y|X}(B|x') P_{0Y|X}(B|x) dP_{0X}(x') dP_{0X}(x).\end{aligned}$$

The result is obtained following some algebra. □

Proof of Theorem 6 First, we show that $E[\mathbf{P}(A \times B)|X_1 = x_1, Y_1 = y_1, ..., X_n = x_n, Y_n =$

$y_n] \to \Pi(A \times B)$ a.s. Π^∞.

$$
\begin{aligned}
&E[\mathbf{P}(A \times B) | X_1 = x_1, Y_1 = y_1, ..., X_n = x_n, Y_n = y_n] \\
&= \frac{\alpha(\mathcal{X})}{\alpha(\mathcal{X}) + n} P_0(A \setminus \{x_1, ..., x_n\} \times B) + \sum_{x \in A \cap \{x_1, ..., x_n\}} \frac{\mu(\mathcal{Y}, x) + \sum_{j=1}^{n_x} \delta_{y_{x,j}}(B)}{\alpha(\mathcal{X}) + n} \frac{\alpha(x) + n_x}{\mu(\mathcal{Y}, x) + n_x} \\
&\sim \frac{1}{n} \sum_{x \in A \cap \{x_1, ..., x_n\}} \sum_{j=1}^{n_x} \delta_{y_{x,j}}(B) = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i, y_i}(A, B) \\
&\to \Pi(A \times B) \text{ a.s } \Pi^\infty.
\end{aligned}
$$

Using lemma (1), we show the posterior variance of $\mathbf{P}(A \times B)$ goes to 0, by showing each of the four terms in (1) goes to 0. Since $\frac{\alpha_n(A)}{\alpha_n(\mathcal{X})} \sim \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}(A)$ and for, $x \in \{x_1, ..., x_n\}$, $\frac{\mu_n(B,x)}{\mu_n(\mathcal{Y},x)} \sim \frac{1}{n_x} \sum_{i=1}^{n_x} \delta_{y_{x,j}}(B)$,

$$
\begin{aligned}
(I_1) &\sim \frac{1}{n} \int_A (\frac{1}{n_x} \sum_{i=1}^{n_x} \delta_{y_{x,j}}(B))(\frac{1}{n_x} + \frac{1}{n_x} \sum_{i=1}^{n_x} \delta_{y_{x,j}}(B)) d(\frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}(x)) \\
&\to 0, \\
(I_2) &\sim \int_A \int_{\{x\}} \frac{1}{n_x} (\frac{1}{n_x} \sum_{i=1}^{n_x} \delta_{y_{x,j}}(B))(\frac{1}{n_x} \sum_{i=1}^{n_x} \delta_{y_{x,j}}(B^c)) d(\frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}(x')) d(\frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}(x)) \\
&\to 0, \\
(I_3) &\sim -\frac{1}{n} \int_A \int_{\{x\}} (\frac{1}{n_x} \sum_{i=1}^{n_x} \delta_{y_{x,j}}(B))^2 d(\frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}(x')) d(\frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}(x)) \\
&\to 0, \\
(I_4) &\sim -\frac{1}{n} \int_A \int_{A \setminus \{x\}} (\frac{1}{n_x} \sum_{i=1}^{n_x} \delta_{y_{x,j}}(B))(\frac{1}{n_{x'}} \sum_{i=1}^{n_{x'}} \delta_{y_{x',j}}(B)) d(\frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}(x')) d(\frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}(x)) \\
&\to 0.
\end{aligned}
$$

This holds for any finite collection of sets. By a straightforward extension of Theorem 2.5.2 of Ghosh and Ramamoorthi (2003), this implies weak convergence of Q_n to δ_Π a.s. Π^∞. □

References

Blackwell, D. and MacQueen, J. B. (1973). "Ferguson Distributions Via Pólya Urn Schemes." *Annals of Statistics*, 1: 353–355. 367, 379

Cifarelli, D. and Regazzini, E. (1978). "Problemi statistici nonparametrici in condizioni di scambiabilità parziale e impiego di medie associative." *Quaderni Istituto di Matematica Finanziaria, Università di Torino*, 12: 1–36. English translation available at `www.unibocconi.it/wps/allegatiCTP/CR-Scamb-parz[1].20080528.135739.pdf`. 372

— (1996). "De Finetti's contribution to probability and statistics." *Statistical Science*, 11: 253–282. 378

Connor, R. and Mosimann, J. E. (1969). "Concepts of Independence for Proportions with a Generalization of the Dirichlet Distribution." *Journal of the American Statistical Association*, 64: 194–206. 361, 363

Consonni, G. and Veronese, P. (2001). "Conditionally reducible natural exponential families and enriched conjugate priors." *Scandinavian Journal of Statistics*, 28: 377–406. 360, 361

Diaconis, P. and Ylvisaker, D. (1979). "Conjugate priors for exponential families." *Annals of Statistics*, 7: 269–281. 361

Doksum, K. A. (1974). "Tailfree and Neutral random probabilities and their posterior distributions." *Annals of Probability*, 2: 183–201. 360

Escobar, M. and West, M. (1995). "Bayesian density estimation and inference using mixtures." *Journal of the American Statistical Association*, 90: 577–588. 376

Ferguson, T. (1973). "A Bayesian analysis of some nonparametric problems." *Annals of Statistics*, 1: 209–230. 370

Geiger, D. and Heckerman, D. (1997). "A characterization of the Dirichlet distribution through global and local parameter independence." *Annals of Statistics*, 25: 1344–1369. 363

Ghosh, J. and Ramamoorthi, R. (2003). *Bayesian nonparametrics*. New York: Springer-Verlag, Springer Series in Statistics. 365, 381, 383

MacEachern, S. (1999). "Dependent nonparametric processes." In *ASA Proceedings of the Section on Bayesian Statistical Science*, 50–55. Alexandria, VA: American Statistical Association. 373

Neal, R. (2000). "Markov Chain Sampling Methods for Dirichlet Process Mixture Models." *Journal of Computational and Graphical Statistcs*, 9: 249–265. 376

Rabe-Hesketh, S. and Skrondal, A. (2005). *Multilevel and longitudinal modeling using stata*. College Station, Texas: Stata Press. 373, 374

Ramamoorthi, R. and Sangalli, L. (2006). "On a Characterization of Dirichlet Distribution." In Upadhyay, S., Singh, U., and Dey, D. (eds.), *Proceedings of the International Conference on Bayesian Statistics and its Applications, Jan. 6-8, 2005*, 385–397. Varanasi, India: Banaras Hindu University. 367, 370

Rodriguez, A., Dunson, D., and Gelfand, A. (2006). "The nested Dirichlet Process." *Journal of the American Statistical Association*, 103: 1131–1154. 373

Sethuraman, J. (1994). "A Constructive Definition of Dirichlet Priors." *Statistica Sinica*, 4: 639–650. 371

Springer, M. and Thompson, W. (1970). "The distribution of Products of Beta, Gamma and Gaussian Random Variables." *Journal on Applied Mathematics*, 18: 721–737. 364

Teh, Y., Jordan, M., Beal, M., and Blei, D. (2006). "Hierarchical Dirichlet Process." *Journal of the American Statistical Association*, 101: 1566–1581. 372

Acknowledgments

The authors are grateful to R.V. Ramamoorthi for his guidance, and to the two referees and the Associate Editor for their very careful and helpful comments. This work has been partially funded by grants from Bocconi University and by PRIN from the Italian Ministry of Education, University, and Research.