Bayesian nonparametrics

5 Feature models

5.1 Motivation

A feature allocation table is a binary matrix which characterizes a set of objects (rows), with different features (columns). For example, consider Table 1.

	Female protagonist	Action	Comedy	Biopic
Gattaca	0	0	0	0
Side effects	1	1	0	0
The Iron Lady	1	0	0	1
Skyfall	0	1	0	0
Zero Dark Thirty	1	1	0	0

Table 1: Example of a feature allocation table for movies

Given a feature allocation table $(Z_{i\ell}; 1 \leq i \leq n, 1 \leq \ell \leq m)$, we can model a variety of data associated to the objects. The simplest example would be a linear regression model with normal random effects for each feature.

Example (Gaussian linear model). Suppose there is a data point X_i for each object i. We assume

$$X_i \mid Z_i, (\mu_\ell) \overset{iid}{\sim} \mathcal{N}\left(\sum_{\ell=1}^{m} Z_{i\ell}\mu_\ell, 1\right) \quad i = 1, \ldots, n$$
$$\mu_\ell \overset{iid}{\sim} \mathcal{N}(0,1) \quad \ell = 1, \ldots, m.$$

For example, X_i could be the amount of money grossed by movie i, and μ_ℓ an additive effect for each feature.

We will be concerned with the situation where the features are not known and neither is the feature allocation table. We will assign a prior to $(Z_{i\ell})$ and infer the latent features that best describe the data. Thus, our task has more in common with Principal Components Analysis than with linear regression, even though the resulting models can be useful for prediction.

Clustering vs. feature allocation. We have discussed the Dirichlet Process as a tool for clustering data. Like clusterings, feature allocations also encode similarity between samples, but a latent feature model could express the variations between subsets of data more parsimoniously.

For example, suppose that movie earnings are really explained by three latent features. There are 2^3 possible feature vectors, so if we clustered the movies we might obtain 8 distinct clusters and a mixture model would estimate a different mean for each one. The feature model explains the same variation with just 3 random effects.

Feature allocations can be used to model a variety of data, by themselves or in combination with latent clusterings. We illustrate this with a couple of examples in which the data are indexed by pairs of objects.

Example (Protein-protein interactions). An experiment measures the interaction between pairs of proteins, recorded in a matrix (X_{ij}). To each protein we assign a vector of latent features which could represent, for example, complexes to which the protein belongs to. This will be encoded in a feature matrix $(Z_{i\ell})$. Each complex that two proteins have in common has a positive effect on the strength of the interaction. This is expressed by the model

$$X_{ij} \mid Z_i, Z_j, (\mu_\ell) \overset{iid}{\sim} \mathcal{N}\left(\sum_{\ell=1}^{m} Z_{i\ell} Z_{j\ell} \mu_\ell, 1\right) \quad 1 \le i, j \le n$$
$$\mu_\ell \overset{iid}{\sim} \text{Gamma}(1,1) \quad 1 \le \ell \le m.$$

By assigning a suitable prior to Z, we can infer whether there are a few complexes that explain the data well.

Example (Social graph). A recent paper models a network of coauthorship using a latent feature allocation[1]. The data are binary variables X_{ij} which indicate whether authors i and j have coauthored a paper. The model includes a feature matrix $(Z_{i\ell})$ where the latent features might represent, for example, different disciplines or communities of authors. In addition, each feature or community is assumed to be partitioned into clusters, where members of different clusters might have stronger or weaker interactions. Let c_i^ℓ be an index for the cluster within feature ℓ that author i belongs to, with the understanding that $c_i^\ell = 0$ if author i does not have feature ℓ. The

[1] Palla, Knowles, and Ghahramani. *An infinite latent attribute model for network data*, ICML (2012).

model for the data is

$$X_{ij} \mid Z \sim \text{Bernoulli}\left(\sigma\left(\mathcal{N}\left(\sum_{\ell=1}^{m} Z_{i\ell}Z_{j\ell}W^{\ell}_{c_i^\ell, c_j^\ell}, 1\right)\right)\right),$$

where σ is the logistic function, and W^ℓ is an interaction matrix for feature ℓ. The authors show that assigning nonparametric priors to the feature allocation Z and the clusterings c^ℓ, and a suitable prior to the interactions W^ℓ yields a highly predictive model for the network.

5.2 A prior for a finite feature model

If we assume that the objects in a feature model are exchangeable, de Finetti's theorem suggests a prior in which each feature ℓ has a latent probability π_ℓ and, given this probability, the sequence $Z_{1\ell}, Z_{2\ell}, \ldots$ is i.i.d. Bernoulli(π_ℓ). We will assume the feature probabilities are independent and assign to them the conjugate Beta prior. This leads to the model

$$\begin{aligned}\pi_1, \ldots, \pi_k &\overset{iid}{\sim} \text{Beta}(\alpha/k, 1) \\ Z_{i\ell} \mid \pi_\ell &\overset{iid}{\sim} \text{Bernoulli}(\pi_\ell) \quad i \geq 1, 1 \leq \ell \leq k.\end{aligned}$$

The parameter α is the expected number of features that any given object will have. Note that we have parametrized the model such that this does not depend on k.

When discussing mixture models, we found it useful to label each cluster with a vector of parameters or labels. We can do the same in a feature allocation model. Let $\theta_1, \ldots, \theta_k$ be labels or parameters for each feature, drawn independently from some distribution G. The feature allocation for object i, can be represented by a measure $\phi_i = \sum_{\ell; Z_{i\ell}=1} \delta_{\theta_\ell}$.

Finally, consider the following algorithm for sampling the feature allocations sequentially. Given $\phi_1, \ldots, \phi_{n-1}$ we sample ϕ_n in two steps:

1. For each feature that has been observed before, $\theta \in \text{Supp}(\sum_{i=1}^{n-1} \phi_i)$, the sequence $\phi_1(\theta), \ldots, \phi_n(\theta)$ is a Polya urn which starts with one ball of weight α/k and one ball of weight 1. Therefore, the probability that ϕ_n has an atom on θ given the first $n-1$ feature allocations is

$$\frac{\alpha/k + \sum_{i=1}^{n-1} \phi_i(\theta)}{\alpha/k + n}.$$

2. Among the features that have not been observed, each one has a probability
$$\frac{\alpha/k}{\alpha/k+n} = \frac{\alpha}{\alpha+kn}$$
of being in ϕ_n. Since the features are independent, if $k_{n-1} = |\text{Supp}(\sum_{i=1}^{n-1}\phi_i)|$ is the number of features observed, we include a
$$\text{Binomial}\left(k - k_{n-1}, \frac{\alpha}{\alpha+kn}\right)$$
number of new atoms in ϕ_n and draw the locations independently from G.

5.3 Nonparametric model

Now, we will take the limit of the finite feature model as $k \to \infty$. In particular, the sampling algorithm above simplifies in the following way. In the first step, the probability of including a feature θ, which has already been observed in $\phi_1, \ldots, \phi_{n-1}$, in the next allocation ϕ_n is simply $\sum_{i=1}^{n-1}\phi_i(\theta)/n$. In the second step, the distribution of the number of new features introduced converges to a Poisson(α/n) distribution as $k \to \infty$.

This procedure in the limit $k \to \infty$ is known as the *Indian Buffet Process*. The objects are customers walking through a buffet, and the features are the dishes they sample. The nth customer samples a dish that has been sampled i times previously with probability i/n, and independently, he samples a Poisson(α/n) number of new dishes.

Lemma 5.1. *There exists a random measure $H = \sum_{i=1}^{\infty}\pi_i\delta_{\theta_i}$ such that given H, the sequence $\phi_1, \phi_2, \ldots$ is i.i.d. and each ϕ_i is a Bernoulli Process with base measure H, i.e.*
$$\phi_i \overset{d}{=} \sum_{i=1}^{\infty} Y_i\delta_{\theta_i},$$
where $Y_i \sim$ Bernoulli(π_i) are independent for $i \geq 1$.

Proof. The feature allocations $\phi_1, \phi_2, \ldots$ are exchangeable in the finite mixture model, so they are also exchangeable in the limit. By de Finetti's theorem, they are independent given the tail σ-field of the process. Finally, we note that as $n \to \infty$, $\phi_n \mid \phi_1, \ldots, \phi_{n-1}$ converges in distribution to a Bernoulli Process with base measure $\sum_{i=1}^{n-1}\phi_i/n$. □

The distribution of the latent measure H is known as the Beta Process with parameter α and base distribution G. We will provide two constructions of the Beta Process which are more instructive than the existence proof above. The first one is a stick breaking construction which is a direct consequence of the following result.

Theorem 5.2. *If $\pi_1, \ldots, \pi_k$ are i.i.d. Beta($\alpha/k, 1$) random variables, their order statistics $\pi_{(1)} \geq \pi_{(2)} \geq \cdots \geq \pi_{(k)}$ satisfy*

$$\pi_i \stackrel{d}{=} \prod_{j=1}^{i} \eta_j,$$

where $\eta_1, \ldots, \eta_k$ are independent and $\eta_i \sim$ Beta($\alpha(k-i+1)/k, 1$) random variables.

In words, to sample the probability of the most likely feature, we break a Beta$(\alpha, 1)$ fraction of a unit stick. The probability of the second most likely feature is sampled by breaking a Beta$(\alpha(k-1)/k, 1)$ fraction of the first probability, and so on. The proof of this will be left as a homework exercise.

Corollary 5.3. *Taking the limit of the finite feature model as $k \to \infty$, we obtain the following stick breaking representation for the Beta Process H. If $\pi_{(i)}$ is the ith largest weight in H, then $\pi_{(i)} \stackrel{d}{=} \prod_{j=1}^{i} \eta_j$ where η_j are i.i.d. Beta($\alpha, 1$) random variables.*

What makes this result remarkable is that distribution of the stick breaking variables converges to a limit as $k \to \infty$. For our initial definition of the Beta Process, we used the limit of the predictive distribution of a finite feature model (the Indian Bufffet Process) instead of taking the limit of the latent feature probabilities $\pi_1, \pi_2, \ldots, \pi_k$. This is because the limit of a Beta$(\alpha/k, 1)$ distribution as $k \to \infty$ is improper. On the other hand, the distribution of the order statistic $\pi_{(i)}$ does have a limit, given by the stick breaking representation. This directly proves the existence of the latent measure H and yields the distribution of the weights (π_i) sorted in decreasing order.

5.4 Gibbs sampling

There are several approaches to sample the posterior distribution of a nonparametric feature allocation given data. The basic approaches to Gibbs

sampling can be divided into marginal and conditional samplers, depending on whether or not the state space of the Markov chain includes the latent probabilities of each feature (π_j).

Marginal Gibbs sampler. Iterate the following steps:

1. Sample $Z_{i\ell} \mid Z_{-(i\ell)}, X, (\theta_j)$ for all (i, ℓ), where $Z_{-(i\ell)}$ is the feature allocation matrix without the (i, ℓ) entry.
2. Sample $(\theta_j) \mid X, Z$.

The second step depends very much on the model $F(X \mid Z, (\theta_j))$, and the prior G on the parameters (θ_j). This step could be performed exactly if F and G have conjugacy; otherwise, we can perform a step of a Metropolis-Hastings Markov chain which preserves the distribution $(\theta_j) \mid X, Z$.

The first step utilizes the fact that the distribution $Z_{i\ell} \mid Z_{-(i\ell)}$ is given by the Indian Buffet Process, because exchangeability allows us to make object i the last in a feature allocation sequence. To obtain $\mathbb{P}(Z_{i\ell} = 1 \mid Z_{-(i\ell)}, X)$, we multiply the prior given by the Indian Buffet Process times the likelihood $F(X \mid Z, (\theta_j))$. If F and G are a conjugate pair, we could marginalize out the parameters (θ_j) from the sampler and instead use the marginal likelihood $F(X \mid Z)$.

The step described works for all features ℓ which have been observed in $Z_{-(i\ell)}$; in addition, we must sample a number K of new features for each object i. Given Z_{-i}, K has a Poisson distribution by the Indian Buffet Process. To obtain the posterior of K conditioned on Z_{-i}, X, and (θ_j), we multiply this Poisson prior by the likelihood $F(X \mid Z, (\theta_j))$. Necessarily, we must truncate K at a finite maximum which makes the algorithm only approximate.

Conditional Gibbs sampler. Iterate the following steps:

1. Sample $Z_{i\ell} \mid X, (\pi_j), (\theta_j)$ for all (i, ℓ).
2. Sample $(\theta_j) \mid X, Z$.
3. Sample $(\pi_j) \mid Z$.

The first step is simple because $\mathbb{P}(Z_{i\ell} = 1 \mid (\pi_j)) = \pi_\ell$, therefore $\mathbb{P}(Z_{i\ell} = 1 \mid (\pi_j), (\theta_j), X) \propto \pi_\ell F(X \mid Z, (\theta_j))$. There is an important difference between conditional Gibbs samplers for feature models and mixture models. Note that in a mixture model, the likelihood of the data given the cluster

memberships is a product of independent likelihoods, which allows us to sample all of the cluster memberships simultaneously given the data and the latent probability of each cluster. In a feature model, the likelihood of the data X given the feature allocations is not a product of factors depending only on a single entry in Z, as illustrated in the examples of Section 5.1.

As in the conditional sampler, we must sample a number of new features for each object. This can be done by ordering the latent probabilities (π_j) in decreasing order as in the stick breaking representation and applying a slice sampling trick, which effectively truncates the number of features under consideration in each Gibbs update. For details, see the paper by Teh et al.[2]

The second step in the sampler is the same as in the marginal algorithm. Finally, to sample $(\pi_j) \mid Z$, we can take advantage of the stick breaking representation and the slice sampling trick mentioned above.

6 Completely Random Measures

Definition. Let H be a random measure on $(\Omega, \mathcal{F})$. We say H is *completely random* if for any set of disjoint measurable sets $F_1, \ldots, F_m$, the random variables $H(F_1), \ldots, H(F_m)$ are mutually independent.

Completely random measures were introduced by Kingman in 1963 as a generalization of a stochastic process with independent increments to arbitrary index sets. Kingman showed that any completely random measure can be represented as the sum of three components:

1. A deterministic measure.
2. A discrete measure with deterministic locations for each atom.
3. A discrete measure $\sum_{i=1}^{\infty} \pi_i \delta_{\theta_i}$ where the set of points $(\pi_i, \theta_i)_{i \geq 1}$ is distributed as an inhomogeneous Poisson Point Process (PPP) on $\mathbb{R}^+ \times \Omega$.

We will be interested in completely random measures consisting only of the third component. In particular, we will focus on the case where the intensity measure of the PPP is a product measure $\mu(d\xi)G(d\theta)$ on $\mathbb{R}^+ \times \Omega$. Our first example is the Beta Process.

[2] Teh, Görür, and Ghahramani. Stick breaking construction for the Indian Buffet Process. *AISTATS* (2007).

Theorem 6.1. *The Beta Process with mass parameter α and base distribution G is distributed as $\sum_{i=1}^{\infty} \pi_i \delta_{\theta_i}$ where $(\pi_i, \theta_i)_{i \geq 1}$ is a PPP with intensity measure $(\alpha \xi^{-1} d\xi) \times G(d\theta)$.*

Proof. We will prove that if H is derived from a PPP as above, and

$$\phi_1, \phi_2, \cdots \mid H \overset{iid}{\sim} \mathrm{BeP}(H),$$

then, the predictive distribution of $\phi_1, \phi_2, \ldots$ is the Indian Buffet Process. The theorem would follow from the uniqueness of the de Finetti measure in Lemma 5.1. Our main tool for proving this claim is the following property of Poisson processes.

Let the measure $\mu = \sum_{i=1}^{\infty} \delta_{X_i}$ be a PPP with intensity measure $\lambda(dx)$ and define $\nu = \sum_{i=1}^{\infty} Y_i \delta_{X_i}$ where, given μ, Y_i is a Bernoulli$(h(X_i))$ random variable independent for $i \geq 1$. We say ν is a *thinning* of μ with thinning rate h.

Theorem 6.2. *The measure ν is also a PPP with intensity measure $\lambda'(dx)$ defined by $\lambda'(A) = \int_A h(x)\lambda(dx)$ for all measurable A.*

To prove the claim, we induct on the following hypothesis. Given the feature allocations $\phi_1, \ldots, \phi_n$, the posterior of H is the sum of two independent components:

1. A process $\sum_{\theta \in S_n} \pi_\theta \delta_\theta$, where S_n is the support of $\sum_{i=1}^n \phi_i$ and $\pi_\theta \sim \mathrm{Beta}(\sum_{i=1}^n \phi_i(\theta), n + 1 - \sum_{i=1}^n \phi_i(\theta))$ for $\theta \in S_n$ are independent.
2. A process $H_n = \sum_{i=1}^{\infty} \pi'_i \delta_{\theta'_i}$ where $(\pi'_i, \theta'_i)_{i \geq 1}$ is a PPP with intensity measure $\alpha(1-\xi)^n \xi^{-1} d\xi \; G(d\theta)$.

The base case $n = 0$ is trivial. Assuming the inductive hypothesis, for every $\theta \in S_n$, we have $\phi_{n+1}(\theta) = 1$ with probability π_θ. By the Beta-Bernoulli conjugacy, in the posterior given $\phi_1, \ldots, \phi_{n+1}$, H has an atom at θ with an independent $\mathrm{Beta}(\sum_{i=1}^{n+1} \phi_i(\theta), n + 2 - \sum_{i=1}^{n+1} \phi_i(\theta))$ weight.

Then, observe that in sampling ϕ_{n+1} we make an independent Bernoulli choice for each atom in H_n. If the atoms chosen are in locations $\{\theta'_i; i \in \mathcal{I}\}$, then the process $(\pi'_i, \theta'_i)_{i \in \mathcal{I}}$ is clearly a thinning with rate $h(\pi, \theta) = \pi$ of the PPP $(\pi'_i, \theta'_i)_{i \geq 1}$. By Theorem 6.2, $(\pi'_i, \theta'_i)_{i \in \mathcal{I}}$ is also a PPP with intensity measure $h(\xi, \theta)\alpha(1-\xi)^n \xi^{-1} d\xi \; G(d\theta) = \alpha(1-\xi)^n d\xi \; G(d\theta)$. The intensity measure has

$$\int_{\mathbb{R}^+ \times \Omega} \alpha(1-\xi)^n d\xi \; G(d\theta) = \frac{\alpha}{n+1};$$

therefore, the number of new points included in ϕ_{n+1} is Poisson$(\alpha/(n+1))$. Furthermore, given the number of new points $|\mathcal{I}|$, the locations $(\pi_i, \theta_i)_{i\in\mathcal{I}}$ are drawn independently from $\alpha(1-\xi)^n d\xi\ G(d\theta)$; i.e. for each new feature in ϕ_{n+1} we pick a location from G and a weight from a Beta$(1, n+1)$ distribution, which is consistent with the induction hypothesis for $n+1$.

Finally, by the same argument, the points in H_n that are not included in ϕ_{n+1} are generated from a PPP with intensity measure $(1-h(\xi,\theta))\alpha(1-\xi)^n\xi^{-1}d\xi\ G(d\theta) = \alpha(1-\xi)^{n+1}\xi^{-1}d\xi\ G(dx)$. These constitute the process H_{n+1}.

□

This proof yields a result stronger than the theorem. We derive a characterization of the posterior of H given $\phi_1, \ldots, \phi_n$. It is then easy to verify that the predictive distribution of the feature allocations (ϕ_i) matches the Indian Buffet Process.

The method of Poisson thinning can be extended to characterize the posterior of more general completely random measures. In Homework 2, you are asked to extend Theorem 6.1 to a completely random measure H with the following Lévy intensity measure:

$$\alpha\frac{\Gamma(1+c)}{\Gamma(1-\sigma)\Gamma(c+\sigma)}\xi^{-\sigma-1}(1-\xi)^{c+\sigma-1}d\xi G(d\theta)$$

on $\mathbb{R}^+ \times \Omega$, for some $\alpha > 0$, $c > -\sigma$, and $\sigma \in [0, 1)$. This is known as the *Stable Beta Process*.

6.1 Stick breaking representation for CRMs

If $(\pi_i, \theta_i)_{i\geq 1}$ is a PPP with intensity $\mu(d\xi)G(d\theta)$ on $[0,1] \times \Omega$, then the process $(\pi_i)_{i\geq 1}$ is a PPP on $[0,1]$ with intensity measure $\mu(d\xi)$. Define the mapping $T : (0,1] \to \mathbb{R}^+$ such that

$$T(u) = \int_u^1 \mu(d\xi)$$

where we assume the right hand side is finite for all $u > 0$ and strictly monotone. The mapping theorem for Poisson processes states that $(T(\pi_i))_{i\geq 1}$ is a Poisson process on $\mathbb{R}^+$ with Lebesgue intensity. It follows that if $\pi_{(1)} \geq \pi_{(2)} \geq \ldots$ are the jumps of the process in decreasing order, the waiting times $t_i := T(\pi_{(i)}) - T(\pi_{(i-1)})$ for $i > 1$ are i.i.d. Exponential(1)

random variables, where we define $\pi_{(0)} = 1$. Therefore,

$$\pi_{(i)} = T^{-1}\left(\sum_{j=1}^{i} t_j\right).$$

This yields a very useful approach to sampling the sorted weights $(\pi_{(i)})_{i\geq 1}$ even when there is no closed-form expression for T or its inverse.

Furthermore, we can rewrite the last expression

$$\pi_{(i)} = T^{-1}\left(t_i + T(\pi_{(i-1)})\right).$$

Since $\pi_{(i-1)}$ only depends on $t_1, \ldots, t_{i-1}$, it is clear from the last identity that $(\pi_{(i)})_{i\geq 1}$ is a Markov chain. This observation is enough to represent the sequence in terms of stick breaking. The distribution of the stick breaking variables is the distribution of the ratio

$$\frac{\pi_{(i)}}{\pi_{(i-1)}} = \frac{T^{-1}\left(\sum_{j=1}^{i} t_j\right)}{T^{-1}\left(\sum_{j=1}^{i-1} t_j\right)}.$$

This distribution is not always nice. In the case of the Beta Process, we have

$$T(u) = \int_u^1 \alpha\xi^{-1} d\xi = -\alpha \log(u)$$

and $T^{-1}(x) = e^{-x/\alpha}$. This leads to

$$\frac{\pi_{(i)}}{\pi_{(i-1)}} = \exp(-t_i/\alpha).$$

Since t_i is Exponential(1) for all $i \geq 1$, the stick breaking variables are i.i.d. and a simple change of variables shows that they are Beta$(\alpha, 1)$ distributed.

6.2 The Gamma Process

Definition. The *Gamma Process* with concentration α and base distribution G is a completely random measure with Lévy intensity

$$\alpha\xi^{-1}e^{-\xi} d\xi \; G(d\theta).$$

The Gamma Process is a fundamental process in Bayesian nonparametrics. It is distinguished from the Beta Process in that it is almost surely a finite and positive measure, which makes it possible to normalize the measure to obtain a probability measure. We will use the following result repeatedly.

Theorem 6.3 (Campbell)**.** *Let ν be a Poisson process on $(\Omega, \mathcal{F})$ with intensity measure $\lambda(dx)$. Then, for any function $f : \Omega \to \mathbb{R}^+$, $\nu(f) < \infty$ if and only if*

$$\int_\Omega \min(1, f(x))\lambda(dx) < \infty.$$

In this case, the Laplace functional of ν

$$E(e^{-\nu(f)}) = \exp\left(-\int (1 - e^{-f(x)})\lambda(dx)\right).$$

Since $3(1 - e^{-f(x)}) \geq \min(1, f(x))$, in order to verify that $\nu(f)$ is finite, it suffices to prove that $\int(1 - e^{-f(x)})\lambda(dx) < \infty$; i.e. the right hand side in the last expression is finite.

Lemma 6.4. *If μ is a Gamma Process with parameters (α, G), for any set of disjoint measurable sets $A_1, \ldots, A_m$, the random variables $\mu(A_1), \ldots, \mu(A_m)$ are independent and satisfy $\mu(A_i) \sim$ Gamma$(\alpha G(A_i), 1)$.*

Proof. The variables are clearly independent because μ is completely random. To prove the distributional result, note that $\mu(A) = \int \xi \mathbb{1}_A(\theta)\mu'(d\xi, d\theta)$ where we use μ' to denote the PPP on $\mathbb{R}^+ \times \Omega$ which gives rise to μ. Then, for $t > 0$,

$$\begin{aligned}
E(e^{-t\mu(A)}) &= \exp\left(-\int_{\mathbb{R}^+\times\Omega} (1 - e^{-t\xi\mathbb{1}_A(\theta)})\alpha\xi^{-1}e^{-\xi}d\xi\ G(d\theta)\right) \\
&= \exp\left(-\alpha G(A)\int_{\mathbb{R}^+} (1 - e^{-t\xi})\xi^{-1}e^{-\xi}d\xi\right) \\
&= \exp\left(-\alpha G(A)\log(1 - t)\right) \\
&= (1 - t)^{-\alpha G(A)}.
\end{aligned}$$

Since the right hand side is finite for $t = 1$, Campbell's theorem implies that $\mu(A)$ is a.s. finite, furthermore, the moment generating function of $\mu(A)$ is that of a Gamma$(\alpha G(A), 1)$ random variable.

□

Note in particular that if μ is a Gamma Process, $\mu(\Omega)$ is a Gamma$(\alpha, 1)$ random variable, which is finite and greater than zero a.s. This allows us to define the *normalized completely random measure* $\tilde{\mu} := \mu/\mu(\Omega)$.

Lemma 6.5. *The measure $\tilde{\mu}$ is a Dirichlet Process with base measure αG.*

Proof. For any disjoint $A_1, \ldots, A_m$, the marginals $\mu(A_1), \ldots, \mu(A_1)$ are independent gamma random variables. It is well known that

$$\left(\frac{\mu(A_1)}{\sum_i \mu(A_i)}, \ldots, \frac{\mu(A_m)}{\sum_i \mu(A_i)} \right) = (\tilde{\mu}(A_1), \ldots, \tilde{\mu}(A_m))$$

has a Dirichlet$(\alpha G(A_1), \ldots, \alpha G(A_m))$ distribution. This characterizes a Dirichlet Process with base measure αG. □

The Gamma Process can be used in models that require an infinite number of atoms with summable weights. We give two examples from the literature.

Example (Gamma-Poisson Process). Let μ be a Gamma Process on $(\Omega, \mathcal{F})$, and given μ, let $\phi_1, \phi_2, \ldots$ be independent Poisson processes with intensity measure μ:

$$\mu \sim \text{GP}(\alpha, G)$$
$$\phi_1, \phi_2, \cdots \mid \mu \overset{iid}{\sim} \text{Poisson}(\mu).$$

Since μ is a discrete and finite measure, each process ϕ_i is defined on the same set of atoms and $\phi_i(\Omega) \sim \text{Poisson}(\mu(\Omega))$ is almost surely finite.

If we construct a table where each row corresponds to a different ϕ_i, and each column corresponds to a different atom in μ, this is similar to a feature allocation matrix, where each entry is a positive integer instead of a binary variable. As in a Beta-Bernoulli process, each row has a finite number of nonzero entries. This is a useful structure to model data where each latent feature can have multiplicities in any given object.

Example (Reversible HMM). We define a Hidden Markov Model in which the latent Markov chain has an infinite number of states and is reversible with probability 1 in the prior. The prior on the transition probability matrix is the following:

$$\mu \sim \text{GP}(\alpha, G)$$
$$\forall \theta_1, \theta_2 \in \text{Supp}(\mu) \quad P(\theta_1, \theta_2) = P(\theta_2, \theta_1) \sim \text{Gamma}(\beta \mu(\theta_1) \mu(\theta_2), 1).$$

This defines a symmetric matrix P, which is a gamma process on the set of pairs of atoms in μ. Note that the rows of P are a.s. finite, therefore, we can define transition probabilities

$$\tilde{P}(\theta_1, \theta_2) = \frac{P(\theta_1, \theta_2)}{\sum_{\theta \in \text{Supp}(\mu)} P(\theta_1, \theta)}$$

on the support of μ. The Markov chain is reversible with respect to the measure $\sum_{\theta \in \text{Supp}(\mu)} P(\cdot, \theta)$.

7 Normalized completely random measures

Of course, the Gamma Process is not the only CRM which is almost surely finite and nonzero. We highlight a CRM which extends the Gamma Process and has three parameters.

Definition. The *Generalized Gamma Process* with parameters $\alpha > 0$, $\sigma \in (0, 1)$ and $\tau \geq 0$ is a completely random measure with Lévy intensity

$$\frac{\alpha}{\Gamma(1-\sigma)} \xi^{-\sigma-1} e^{-\tau\xi} d\xi \; G(d\theta).$$

Several special cases of this process have names in the literature:

- The Gamma Process has $\sigma = 0$ and $\tau = 1$.
- The *Stable Process* has $\tau = 0$.
- The *Inverse Gaussian Process* has $\sigma = 0.5$.

The family of random probability measures obtained by normalizing a CRM has properties that are convenient for Bayesian analysis and have been studied extensively.

James, Lijoi, and Prünster[3] proved that these measures satisfy a conditional conjugacy in the following model:

$$\begin{aligned} \mu &\sim \text{CRM with intensity } \rho(dw)G(dx) \\ T &:= \mu(\Omega) \\ \tilde{\mu} &:= \mu/T \\ X_1, \ldots, X_n \mid \tilde{\mu} &\overset{iid}{\sim} \tilde{\mu} \\ U \mid T &\sim \text{Gamma}(n, T). \end{aligned}$$

This is a standard model for sampling from a random measure with a normalized CRM prior, where we introduce an auxiliary variable U which does not affect the marginal distribution of other variables in the model. Let $X_1^*, \ldots, X_k^*$ be the unique values observed in $X_1, \ldots, X_n$ and let $n_1, \ldots, n_k$ be the size of each cluster.

[3] James, Lijoi, and Prünster. Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics*, (2009).

Theorem 7.1 (James et al.)**.** *The posterior of μ given $X_1, \ldots, X_n$ and U is the distribution of*

$$\mu_p + \sum_{j=1}^{k} W_j \delta_{X_j^*},$$

where W_j has density proportional to $e^{-Uw} w^{n_j}$ with respect to $\rho(dw)$, and is independent for each $j = 1, \ldots, k$ and independent from μ_p, a CRM with an exponentially tilted intensity measure

$$e^{-wU} \rho(dw) G(dx).$$

Proof. We claim that the characteristic functional of μ given $X_1, \ldots, X_n, U$ is

$$E\left[e^{-\mu(f)} \mid X_1, \ldots, X_n, U\right] =$$
$$\exp\left(-\int (1 - e^{wf(x)}) e^{-wU} \rho(dw) G(dx)\right) \prod_{j=1}^{k} \int e^{-f(X_j^*)w} \frac{e^{-Uw} w^{n_j} \rho(dw)}{B(U, n_j)},$$

where $B(u, j) = \int e^{-uw} w^j \rho(dw)$. This functional characterizes the posterior, so to prove the theorem it suffices to show that this is equal to $E_p[e^{-\mu(f)}]$ where E_p denotes the expectation with respect to the posterior in the statement of the theorem. Indeed,

$$E_p\left[e^{-\mu(f)}\right] =$$
$$E_p\left[e^{-\mu_p(f) - \sum_{j=1}^{k} f(X_j^*) W_j}\right] =$$
$$E_p\left[e^{-\mu_p(f)}\right] \prod_{j=1}^{k} E_p\left[e^{-f(X_j^*)W_j}\right] =$$
$$\exp\left(-\int (1 - e^{-wf(x)}) e^{-wU} \rho(dw) G(dx)\right) \prod_{j=1}^{k} \int e^{-f(X_j^*)w} \frac{e^{-Uw} w^{n_j} \rho(dw)}{B(U, n_j)}.$$

To prove the claim, we will take the expectation of $e^{-\mu(f)}$ given that $X_i \in dx_i$ and $U \in du$ for infinitesimal sets $dx_1, \ldots, dx_n$ and du. Given these events, μ has conditional density

$$\frac{\prod_{i=1}^{k} \frac{\mu(dx_i^*)^{n_i}}{T^{n_i}} \frac{u^{n-1}}{\Gamma(n)} e^{-Tu} T^n du}{E\left[\prod_{i=1}^{k} \frac{\mu(dx_i^*)^{n_i}}{T^{n_i}} \frac{u^{n-1}}{\Gamma(n)} e^{-Tu} T^n du\right]}$$

with respect to the CRM prior on μ, where E is the expectation with respect to the prior. Then,

$$E\left[e^{-\mu(f)} \mid X_i \in dx_i \ \forall i, U \in du\right] = \frac{E\left[e^{-\mu(f)} \prod_{i=1}^k \mu(dx_i^*)^{n_i} \frac{u^{n-1}}{\Gamma(n)} e^{-Tu} du\right]}{E\left[\prod_{i=1}^k \mu(dx_i^*)^{n_i} \frac{u^{n-1}}{\Gamma(n)} e^{-Tu} du\right]}$$
$$= \frac{E\left[e^{-\mu(f)} \prod_{i=1}^k \mu(dx_i^*)^{n_i} e^{-Tu}\right]}{E\left[\prod_{i=1}^k \mu(dx_i^*)^{n_i} e^{-Tu}\right]}. \quad (1)$$

The denominator is equal to the numerator when $f = 0$. The numerator can be simplified using the Palm property of Poisson processes. This property states that if $N(dx)$ is a Poisson process with intensity $\lambda(dx)$, we have

$$E\left[\int H(N) g(x) N(dx)\right] = \int E[H(N + \delta'_x)] g(x') \lambda(dx'),$$

where $N + \delta_{x'}$ is the Poisson process with an extra point at x'. Letting μ' be the Poisson Process on $\mathbb{R}^+ \times \Omega$ that generates μ, $H(\mu') = e^{-\mu(f+u)} \prod_{i=2}^k \mu(dx_i^*)^{n_i}$ and $g(w, x) = \mathbb{1}(x \in dx_1^*) w^{n_1}$, we can write the numerator of Eq. 1

$$E\left[\int_{\mathbb{R}^+ \times \Omega} H(\mu') g(w, x) \mu'(dw, dx)\right],$$

where we use the fact that dx_1^* is infinitesimal. By the Palm property, this is equal to

$$\int_{\mathbb{R}^+ \times \Omega} E\left[e^{-\mu(f+u) - (f(x')+u)w'} \prod_{i=2}^k \mu(dx_i^*)^{n_i}\right] g(w', x') \rho(dw') G(dx'),$$

where we use the fact that $dx_2^*, \ldots, dx_k^*$ are infinitesimal. This can be simplified to,

$$\int_{\mathbb{R}^+ \times \Omega} E\left[e^{-\mu(f+u)} \prod_{i=2}^k \mu(dx_i^*)^{n_i}\right] e^{-(f(x')+u)w'} \mathbb{1}(x \in dx_1^*) w'^{n_1} \rho(dw') G(dx')$$
$$= G(dx_1^*) E\left[e^{-\mu(f+u)} \prod_{i=2}^k \mu(dx_i^*)^{n_i}\right] \int_{\mathbb{R}^+} e^{-w' f(x_1^*)} e^{-uw'} w'^{n_1} \rho(dw').$$

Repeating the argument for $i = 2, \ldots, k$, we arrive at

$$E\left[e^{-\mu(f+u)}\right] \prod_{i=1}^k G(dx_i^*) \int_{\mathbb{R}^+} e^{-w' f(x_i^*)} e^{-uw'} w'^{n_i} \rho(dw').$$

Setting $f = 0$ we derive the denominator of Eq. 1

$$E\left[e^{-Tu}\right] \prod_{i=1}^{k} G(dx_i^*) \int_{\mathbb{R}^+} e^{-uw'} w'^{n_i} \rho(dw').$$

Taking the ratio of the numerator and denominator, we obtain the desired functional, which is only a function of u and $x_1, \ldots, x_n$,

$$\frac{E[e^{-\mu(f+u)}]}{E[e^{-Tu}]} \prod_{j=1}^{k} \int e^{-f(x_j^*)w} \frac{e^{-uw} w^{n_j} \rho(dw)}{B(u, n_j)}.$$

We apply Campbell's theorem to complete the proof of the claim,

$$\frac{\exp\left(-\int (1 - e^{-(f(x)+u)w}) \rho(dw) G(dx)\right)}{\exp\left(-\int (1 - e^{-uw}) \rho(dw) G(dx)\right)} \prod_{j=1}^{k} \int e^{-f(x_j^*)w} \frac{e^{-uw} w^{n_j} \rho(dw)}{B(u, n_j)}$$

$$= \exp\left(-\int (1 - e^{wf(x)}) e^{-wu} \rho(dw) G(dx)\right) \prod_{j=1}^{k} \int e^{-f(x_j^*)w} \frac{e^{-uw} w^{n_j} \rho(dw)}{B(u, n_j)}.$$

□

This result has been applied to define MCMC methods for Bayesian analysis of mixture models with normalized CRM priors. For details, see the paper by Favaro and Teh[4].

From the proof of Theorem 7.1, we derive an expression for the probability of a sequence $X_1, \ldots, X_n$. First, conditional on U,

$$\begin{aligned}
\mathbb{P}(X_i \in dx_i \ \forall i \mid U \in du) &= E\left[\prod_{i=1}^{k} \frac{\mu(dx_i^*)^{n_i}}{T^{n_i}} \frac{\Gamma(n)^{-1} u^{n-1} T^n e^{-Tu} du}{E[\Gamma(n)^{-1} u^{n-1} T^n e^{-Tu} du]}\right] \\
&= \frac{1}{E[T^n e^{-Tu}]} E\left[\prod_{i=1}^{k} \mu(dx_i^*)^{n_i} e^{-Tu}\right],
\end{aligned}$$

where the second factor is the denominator in Eq. 1. This leads to

$$\mathbb{P}(X_i \in dx_i \ \forall i \mid U \in du) = \prod_{i=1}^{k} G(dx_i^*) \frac{E\left[e^{-Tu}\right]}{E[T^n e^{-Tu}]} \prod_{i=1}^{k} B(u, n_i).$$

[4]Favaro and Teh. MCMC for Normalized Random Measure Mixture Models. *Statistical Science*, (2013)

The first product is the probability of the labels falling in $dx_1^*, \ldots, dx_k^*$, the important part of this formula is

$$\frac{E\left[e^{-Tu}\right]}{E[T^n e^{-Tu}]} \prod_{i=1}^{k} B(u, n_i),$$

the probability of a sequence which has k clusters of sizes $n_1, \ldots, n_k$, also known as the *exchangeable partition probability function or EPPF*, which we denote $p(n_1, \ldots, n_k \mid U)$[5]. To obtain the unconditional EPPF, we integrate with respect to the marginal density of U,

$$\begin{aligned} p(n_1, \ldots, n_k) &= \int_0^\infty \frac{E\left[e^{-Tu}\right]}{E[T^n e^{-Tu}]} \prod_{i=1}^{k} B(u, n_i) E\left[\Gamma(n)^{-1} u^{n-1} T^n e^{-Tu}\right] du \\ &= \Gamma(n)^{-1} \int_0^\infty E\left[e^{-Tu}\right] u^{n-1} \prod_{i=1}^{k} B(u, n_i) du. \end{aligned} \tag{2}$$

In Homework 3, you will use Eq. 2 to derive the EPPF of a normalized Generalized Gamma Process.

8 Poisson-Kingman partitions

Poisson-Kingman models are a class of random probability measures which generalize normalized CRMs. They were introduced by Pitman[6].

Definition. A *Poisson-Kingman Process*, $\mathrm{PKP}(\rho G, \gamma)$, is a random probability on $(\Omega, \mathcal{F})$ parametrized by measures ρ and γ on $\mathbb{R}^+$ and G on Ω. Let μ be a CRM with Lévy intensity $\rho(dw)G(dx)$, then,

$$\frac{\mu}{\mu(\Omega)} \mid \mu(\Omega) = t \sim \mathrm{PKP}(\rho G, \delta_t).$$

In words, $\mathrm{PKP}(\rho G, \delta_t)$ is the distribution of a normalized CRM $\mu/\mu(\Omega)$ given that $\mu(\Omega) = t$. For a general measure γ on $\mathbb{R}^+$, we define $\mathrm{PKP}(\rho G, \gamma)$ to be the mixture of $\mathrm{PKP}(\rho G, \delta_t)$ with measure γ on t,

$$\mathrm{PKP}(\rho G, \gamma) = \int \mathrm{PKP}(\rho G, \delta_t) \gamma(dt).$$

[5] Note, this is the probability of a single sequence with these cluster sizes, not the overall probability that $X_1, \ldots, X_n$ is partitioned into clusters of size $n_1, \ldots, n_k$.

[6] Pitman. Poisson-Kingman partitions. *Institute of Mathematical Statistics Lecture Notes – Monograph Series*, (2003).

The most famous example of a Poisson-Kingman Process is the Pitman-Yor process, which is derived from a Stable CRM, μ, with Lévy intensity

$$\rho_\sigma(dw)G(dx) = \frac{\sigma}{\Gamma(1-\sigma)} w^{-\sigma-1} G(dx).$$

Under this model the distribution of the total mass $\mu(\Omega)$ is $g_\sigma(dT)$. The Pitman-Yor process with concentration parameter α and discount parameter σ is equivalent to $\text{PKP}(\rho_\sigma G, \gamma_{\alpha,\sigma})$, where

$$\gamma_{\alpha,\sigma}(dT) = \frac{\Gamma(\alpha+1)}{\Gamma(\alpha/\sigma+1)} T^{-\alpha} g_\sigma(dT).$$

9 Gibbs-type priors

Gibbs type priors are a family of random probability measures which include the Pitman-Yor process, the normalized Generalized Gamma Process, and therefore the Dirichlet Process. They are defined through the random partitions they induce.

Definition. Let μ be a random probability measure, and let $X_1, \ldots, X_n$ be random variables which are i.i.d. from μ, conditional on μ. We say μ is of *Gibbs type* it has an EPPF of the form

$$p(n_1, \ldots, n_k) = V_{n,k} \prod_{i=1}^{k} W_{n_i}.$$

As noted previously, the EPPF is the probability of a specific sequence $X_1, \ldots, X_n$ which is partitioned into k clusters of sizes $n_1, \ldots, n_k$, so $p(\cdot)$ is by definition symmetric in its arguments.

Example. The Pitman-Yor process with parameters (α, σ) is a Gibbs type prior, because its EPPF is

$$p(n_1, \ldots, n_k) = \frac{(\alpha)_{\sigma\uparrow k}}{(\alpha)_{1\uparrow n}} \prod_{i=1}^{k} (1-\sigma)_{1\uparrow n_i - 1}.$$

Pitman and Gnedin characterized these priors as Poisson-Kingman processes with a stable Lévy measure[7]. In the following, we summarize their results.

[7] Pitman and Gnedin. Exchangeable partitions and Stirling triangles. *Journal of Mathematical Sciences*, (2008)

Theorem 9.1. *Let p be the EPPF of an exchangeable partition induced by a Gibbs-type random probability; that is,*

$$p(n_1, \ldots, n_k) = V_{n,k} \prod_{i=1}^{k} W_{n_i}.$$

There exists $\sigma < 1$, such that $W_j = (1-\sigma)_{1\uparrow j-1}$, where we define $W_1 = 1$.

Proof. The EPPF is invariant to the following transformation of the parameters: $W_i \to cW_i$ for all $i \geq 1$, and $V_{n,k} \to c^{-k}V_{n,k}$ for all $1 \leq k \leq n$. Therefore, without loss of generality we can choose $W_1 = 1$.

Choose any $n_1, n_2 \geq 1$. Every EPPF satisfies

$$p(n_1, n_2) = p(n_1, n_2, 1) + p(n_1 + 1, n_2) + p(n_1, n_2 + 1),$$

because the finite dimensional distributions of $X_1, X_2, \ldots$ are consistent. Letting $n = n_1 + n_2$, this can be rewritten

$$V_{n,2}W_{n_1}W_{n_2} = V_{n+1,3}W_{n_1}W_{n_2} + V_{n+1,2}W_{n_1+1}W_{n_2} + V_{n+1,2}W_{n_1}W_{n_2+1},$$

rearranging terms yields

$$\frac{W_{n_1+1}}{W_{n_1}} + \frac{W_{n_2+1}}{W_{n_2}} = \frac{V_{n,2} - V_{n+1,3}}{V_{n+1,2}}.$$

The right hand side only depends on $n = n_1 + n_2$, therefore

$$\frac{W_{n_1+1}}{W_{n_1}} + \frac{W_{n_2+1}}{W_{n_2}} = \frac{W_{n_1+2}}{W_{n_1+1}} + \frac{W_{n_2}}{W_{n_2-1}},$$

This implies that

$$\frac{W_{n_1+2}}{W_{n_1+1}} - \frac{W_{n_1+1}}{W_{n_1}}$$

is constant, so we can write

$$\frac{W_{j+1}}{W_j} = jb - \sigma$$

for some constants σ and b. The EPPF is also invariant to the following transformation of the parameters: $W_i \to d^i W_i$ for all $i \geq 1$, and $V_{n,k} \to d^{-n}V_{n,k}$ for all $1 \leq k \leq n$. Therefore, in addition to setting $W_1 = 1$, we can choose $b = 1$. We conclude

$$W_j = (1-\sigma)_{1\uparrow j-1} \quad \forall j \geq 2,$$

where $1 - \sigma = W_2 > 0$ implies $\sigma < 1$.

□

This theorem implies that every Gibbs type prior is fully specified by the parameter σ and a triangular array of numbers $(V_{n,k}; 1 \leq k \leq n, n \geq 1)$. Since the EPPF is associated to a consistent stochastic process $X_1, X_2, \ldots$, we have

$$V_{n,k} = V_{n+1,k+1} + V_{n+1,k} \sum_{i=1}^{k} \frac{W_{n_i+1}}{W_{n_i}}$$

for any $1 \geq k \leq n$ and any partition $n_1, \ldots, n_k$ of n. Simplifying,

$$\begin{aligned} V_{n,k} &= V_{n+1,k+1} + V_{n+1,k} \sum_{i=1}^{k} (1 - \sigma + n_i - 1) \\ V_{n,k} &= V_{n+1,k+1} + V_{n+1,k}(n - \sigma k), \end{aligned} \tag{3}$$

we obtain a recursion for the triangular array. One can verify that any positive triangular array satisfying this recursion parametrizes a valid Gibbs-type prior.

Pitman and Gnedin observed that the set $\mathcal{V}_\sigma$ of triangular arrays satisfying recursion 3 for a specific value of σ is convex since all the constraints are linear. Furthermore, there is a bijection between $\mathcal{V}_\sigma$ and the set $\mathcal{P}_\sigma$ of Gibbs-type priors with a specific value of σ. There is also a bijection between the extreme points of these two convex sets.

The number of clusters K_n in a sample of size n is a sufficient statistic of the sequence of partitions, in the sense that the probability of a specific partition given K_n, or

$$\begin{aligned} p(n_1, \ldots, n_k \mid K_n = k) &= \frac{V_{n,k} \prod_{i=1}^{k} W_{n_i}}{\sum_{\substack{n'_1, \ldots, n'_k \\ n'_i \geq 1, \sum n'_i = n}} \binom{n}{n_1, \ldots, n_k} V_{n,k} \prod_{i=1}^{k} W_{n'_i}} \\ &= \frac{\prod_{i=1}^{k} W_{n_i}}{\sum_{\substack{n'_1, \ldots, n'_k \\ n'_i \geq 1, \sum n'_i = n}} \binom{n}{n_1, \ldots, n_k} \prod_{i=1}^{k} W_{n'_i}} \end{aligned}$$

does not depend on the parameters $(V_{n,k})$, but only on σ. Pitman and Gnedin appeal to a result from Choquet theory[8] which implies that any array $(V_{n,k}) \in \mathcal{V}_\sigma$ is a *unique* convex combination or mixture of the extreme points in $\mathcal{V}_\sigma$. The extreme points are characterized in the following theorem.

Theorem 9.2. *A Gibbs-type prior with parameter σ can be classified in the following way:*

[8] For a statistical perspective, see: Dianonis and Freedman. Partial exchangeability and sufficiency. *Statistics: Applications and new directions*, (1984).

- *If $\sigma = 0$, it is a unique mixture of Dirichlet Processes with a random concentration α.*
- *If $\sigma < 0$, it is a unique mixture of finite, symmetric Dirichlet distributions with pseudo counts $-\sigma$ on each class and a random number of classes k.*
- *If $\sigma \in (0, 1)$, it is a Poisson-Kingman Process PKP($\rho_\sigma G, \gamma$) where ρ_σ is a Stable Lévy intensity, for some unique measure γ. In other words, the extreme points of $\mathcal{P}_\sigma$ are the processes PKP($\rho_\sigma G, \delta_t$).*

The proof of Gnedin and Pitman implies a related result about the asymptotic behavior of Gibbs-type priors. Define the *σ-diversity* as

$$S_\sigma = \lim_{n\to\infty} \frac{K_n}{c(n)},$$

where

$$c(n) = \begin{cases} 1 & \text{if } \sigma < 0, \\ \log(n) & \text{if } \sigma = 0, \\ n^\sigma & \text{if } \sigma \in (0, 1). \end{cases}$$

It holds that S_σ is an a.s. finite random variable, and conditional on $S_\sigma = s$, the process $X_1, X_2, \ldots$ is generated by an extreme Gibbs-type random measure which has σ-diversity s a.s. Therefore, the distribution of S_σ is in one-to-one correspondence with the mixing distributions in Theorem 9.2. For example, when $\sigma = 0$, the σ-diversity S_0 has the same distribution as the concentration parameter α of the Dirichlet Process in the mixture. When $\sigma < 0$, S_σ has the same distribution as the number of classes k.