1

Colouring and breaking sticks: random distributions and heterogeneous clustering

Peter J. Green[a]

Abstract

We begin by reviewing some probabilistic results about the Dirichlet Process and its close relatives, focussing on their implications for statistical modelling and analysis. We then introduce a class of simple mixture models in which clusters are of different 'colours', with statistical characteristics that are constant within colours, but different between colours. Thus cluster identities are exchangeable only within colours. The basic form of our model is a variant on the familiar Dirichlet process, and we find that much of the standard modelling and computational machinery associated with the Dirichlet process may be readily adapted to our generalisation. The methodology is illustrated with an application to the partially-parametric clustering of gene expression profiles.

Some key words: Bayesian nonparametrics, gene expression profiles, hierarchical models, loss functions, MCMC samplers, optimal clustering, partition models, Pólya urn, stick breaking.

[a] School of Mathematics, University of Bristol, Bristol BS8 1TW, UK.
Email: P.J.Green@bristol.ac.uk.

Edited by

1.1 Introduction

The purpose of this note is four-fold: to remind some Bayesian nonparametricians gently that closer study of some probabilistic literature might be rewarded, to encourage probabilists to think that there are statistical modelling problems worth of their attention, to point out to all another important connection between the work of John Kingman and modern statistical methodology (the role of the coalescent in population genetics approaches to statistical genomics being the most important example; see papers by Donnelly, Ewens and Griffiths in this volume), and finally to introduce a modest generalisation of the Dirichlet process.

The most satisfying basis for statistical clustering of items of data is a probabilistic model, which usually takes the form of a mixture model, broadly interpreted. In most cases, the statistical characteristics of each cluster or mixture component are the same, so that cluster identities are *a priori* exchangeable. In Section 1.5 we will introduce a class of simple mixture models in which clusters are of different categories, or colours as we shall call them, with statistical characteristics that are constant within colours, but different between colours. Thus cluster identities are exchangeable only within colours.

1.2 Mixture models and the Dirichlet process

Many statistical models have the following character. Data $\{Y_i\}$ are available on n units that we shall call *items*, indexed $i = 1, 2, \ldots, n$. There may be item-specific covariates, and other information, and the distribution of each Y_i is determined by an unknown parameter $\phi_i \in \Omega$, where we will take Ω here to be a subset of a euclidean space. Apart from the covariates, the items are considered to be exchangeable, so we assume the $\{Y_i\}$ are conditionally independent given $\{\phi_i\}$, and model the $\{\phi_i\}$ as exchangeable random variables. Omitting covariates for simplicity, we write $Y_i|\phi_i \sim f(\cdot|\phi_i)$.

It is natural to take $\{\phi_i\}$ to be independent and identically distributed random variables, with common distribution G, where G itself is unknown, and treated as random. We might be led to this assumption whether we are thinking of a de Finetti-style representation theorem (de Finetti (1931, 1937); see also Kingman (1978), Kallenberg (2005)), or by following hierarchical modelling principles (Gelman et al., 1995;

Green et al., 2003), Thus, unconditionally, $Y_i|G \sim \int f(\cdot|\phi)G(d\phi)$, independently given G.

This kind of formulation enables us to borrow strength across the units in inference about unknown parameters, with the aim of controlling the degrees of freedom, capturing the idea that while the $\{\phi_i\}$ may be different from item to item, we nevertheless understand that, through exchangeability, knowing the value of one of them would tell us something about the others.

There are still several options. One is to follow a standard parametric formulation, and to assume a specific parametric form for G, with parameters, or rather 'hyperparameters', in turn given a hyperprior distribution. However, many would argue that in most practical contexts, we would have little information to build such a model for G, which represents variation in the population of possible items of the parameter ϕ that determines the distribution of the data Y.

Thus we would be led to consider more flexible models, and one of several approaches might occur to us:

- a nonparametric approach, modelling uncertainty about G without making parametric assumptions;
- a mixture model representation for G;
- a partition model, where the $\{\phi_i\}$ are grouped together, in a way determined *a posteriori* by the data.

One of the things we will find, below, is that taking natural choices in each of these approaches can lead to closely related formulations in the end, so long as both modelling and inference depend solely on the $\{\phi_i\}$. These connections, not novel but not entirely well-known either, shed some light on the nature and implications of the different modelling approaches.

1.2.1 Ferguson definition of the Dirichlet process

Much Bayesian nonparametric distributional modelling (Walker et al., 1999) begins with the Dirichlet process (Ferguson, 1973). Building on earlier work by Dubins, Freedman and Fabius, Ferguson intended this model to provide a nonparametric prior model for G with a large support, yet one remaining capable of tractable prior–to–posterior analysis.

Given a probability distribution G_0 on an arbitrary measure space Ω, and a positive real θ, we say the random distribution G on Ω follows a

Dirichlet process,

$$G \sim DP(\theta, G_0)$$

if for all partitions $\Omega = \bigcup_{j=1}^{m} B_j$ ($B_j \cap B_k = \emptyset$ if $j \neq k$), and for all m,

$$(G(B_1), \dots, G(B_m)) \sim \text{Dirichlet}(\theta G_0(B_1), \dots, \theta G_0(B_m)),$$

where $\text{Dirichlet}(\alpha_1, \alpha_2, \dots, \alpha_m)$ denotes the distribution on the m-dimensional simplex with density at $(x_1, x_2, \dots, x_m)$ proportional to $\prod_{j=1}^{m} x_j^{\alpha_j - 1}$.

The base measure G_0 gives the *expectation* of G:

$$E(G(B)) = G_0(B)$$

Even if G_0 is continuous, G is a.s. discrete (Kingman, 1967; Ferguson, 1973; Blackwell, 1973; Kingman, 1975), so i.i.d. draws $\{\phi_i, i = 1, 2, \dots, n\}$ from G exhibit ties. The parameter θ measures (inverse) *concentration*: given i.i.d. draws $\{\phi_i, i = 1, 2, \dots, n\}$ from G,

- As $\theta \to 0$, all ϕ_i are equal, a single draw from G_0.
- As $\theta \to \infty$, the ϕ_i are drawn i.i.d. from G_0.

1.2.2 The stick-breaking construction

A draw G from a Dirichlet process is a discrete distribution on Ω, so an alternative way to define the Dirichlet process would be via a construction of such a random distribution, through specification of the joint distribution of the locations of the atoms, and their probabilities. Such a construction was given by Ferguson (1973): in this, the locations are i.i.d. draws from G_0, with probabilities forming a decreasing sequence constructed from increments of a gamma process.

This is not the explicit construction that is most commonly used today, which is that known in the Bayesian nonparametric community as Sethuraman's stick-breaking model (Sethuraman and Tiwari, 1982; Sethuraman, 1994). This leads to this algorithm for generating the $\{\phi_i\}$:

1. draw $\phi_j^\star \sim G_0$, i.i.d., $j = 1, 2, \dots$
2. draw $V_j \sim \text{Beta}(1, \theta)$, i.i.d., $j = 1, 2, \dots$
3. define G to be the discrete distribution putting probability $(1-V_1)(1-V_2)\dots(1-V_{j-1})V_j$ on $\phi_j^\star$
4. draw ϕ_i i.i.d from G, $i = 1, 2, \dots, n$.

This construction can be found considerably earlier in the probability literature, especially in connection with models for species sampling.

The earliest reference seems to be in McCloskey (1965); for more readily accessible sources, see Patil and Taillie (1977) and Donnelly and Joyce (1989), where it is described in the context of size-biased sampling and the GEM (Generalised Engen–McCloskey) distributions. See also Section 1.3.3 below.

1.2.3 Limits of finite mixtures

A more direct, classical approach to modelling the distribution of Y in a flexible way would be to use a finite mixture model. Suppose that Y_i are i.i.d. with density $\sum_j w_j f_0(\cdot|\phi_j^\star)$ for a prescribed parametric density family $f_0(\cdot|\phi)$, and consider a Bayesian formulation with priors on the component weights $\{w_j\}$ and the component-specific parameters $\{\phi_j^\star\}$. The simplest formulation (e.g. Richardson and Green (1997)) uses a Dirichlet prior on the weights, and takes the $\{\phi_j^\star\}$ to be i.i.d. *a priori*, but with arbitrary distribution, so in algorithmic form:

1. Draw $(w_1, w_2, \dots, w_k) \sim \text{Dirichlet}(\delta, \dots, \delta)$
2. Draw $c_i \in \{1, 2, \dots, k\}$ with $P\{c_i = j\} = w_j$, i.i.d., $i = 1, \dots, n$
3. Draw $\phi_j^\star \sim G_0$, i.i.d., $j = 1, \dots, k$
4. Set $\phi_i = \phi_{c_i}^\star$

It is well known that if we take the limit $k \to \infty$, $\delta \to 0$ such that $k\delta \to \theta$, then the joint distribution of the $\{\phi_i\}$ is the same as that obtained via the Dirichlet process formulation in the previous subsections (see for example Green and Richardson (2001)). This result is actually a corollary of a much stronger statement due to Kingman (1975), about the convergence of discrete probability measures. For more recent results in this direction see Muliere and Secchi (2003) and Ishwaran and Zarepour (2002).

We are still using the formulation $Y_i|G \sim \int f(\cdot|\phi)G(d\phi)$, independently given G, but note that G is invisible in this view; it has implicitly been integrated out.

1.2.4 Partition distribution

Suppose that, as above, G is drawn from $DP(\theta, G_0)$, and then $\{\phi_i : i = 1, 2, \dots, n\}$ drawn i.i.d. from G. We can exploit the conjugacy of the Dirichlet with respect to multinomial sampling to integrate out G. For a fixed partition $\{B_j\}_{j=1}^m$ of Ω, and integers $c_i \in \{1, 2, \dots, m\}$, we can write

$$P\{\phi_i \in B_{c_i}, i = 1, 2, \ldots, n\} = \frac{\Gamma(\theta)}{\Gamma(\theta+n)} \prod_{j=1}^{m} \frac{\Gamma(\theta G_0(B_j) + n_j)}{\Gamma(\theta G_0(B_j))},$$

where $n_j = \#\{i : c_i = j\}$. The jth factor in the product above is 1 if $n_j = 0$, and otherwise $\theta G_0(B_j)(\theta G_0(B_j)+1)(\theta G_0(B_j)+2)\ldots(\theta G_0(B_j)+n_j-1)$, so we find that if the partition becomes increasingly refined, and G_0 is non-atomic, then the joint distribution of the $\{\phi_i\}$ can equivalently be described by

1. partitioning $\{1, 2, \ldots, n\} = \bigcup_{j=1}^{d} C_j$ at random, so that
$$p(C_1, C_2, \ldots, C_d) = \frac{\Gamma(\theta)}{\Gamma(\theta+n)} \theta^d \prod_{j=1}^{d} (n_j - 1)! \tag{1.1}$$
where $n_j = \#C_j$.
2. drawing $\phi_j^\star \sim G_0$, i.i.d., $j = 1, \ldots, d$, and then
3. setting $\phi_i = \phi_j^\star$ if $i \in C_j$.

Note that the partition model (1.1) shows extreme preference for unequal cluster sizes. If we let $a_r = \#\{j : n_j = r\}$, then the joint distribution of $(a_1, a_2, \ldots)$ is

$$\frac{n!}{n_1! n_2! \cdots n_d!} \times \frac{1}{\prod_r a_r!} \times p(C_1, C_2, \ldots, C_d) \tag{1.2}$$

This is equation (A3) of Ewens (1972), derived in a context where n_j is the number of genes in a sample of the jth allelic type, in sampling from a selectively neutral population process. The first factor in (1.2) is the multinomial coefficient accounting for the number of ways the n items can be allocated to clusters of the required sizes, and the second factor accounts for the different sets of $\{n_1, n_2 \ldots, n_d\}$ leading to the same $(a_1, a_2, \ldots)$. Multiplying all this together, a little manipulation leads to the familiar Ewens' sampling formula:

$$p(a_1, a_2, \ldots) = \frac{n!\Gamma(\theta)}{\Gamma(\theta+n)} \prod_r \frac{\theta^{a_r}}{r^{a_r} a_r!}. \tag{1.3}$$

See also Kingman (1993), page 97.

This representation of the partition structure implied by the Dirichlet process was derived by Antoniak (1974), in the form (1.3). He noted that a consequence of this representation is that the joint distribution of the $\{\phi_i\}$ given d is independent for θ; thus given observed $\{\phi_i\}$, d is sufficient for θ. A similar observation was also made by Ewens (1972) in the genetics context of his work.

Note that as in the previous section, G has been integrated out, and so is invisible in this view of the Dirichlet process model.

1.2.5 Reprise

Whichever of the points of view is taken, items are clustered, according to a tractable distribution parameterised by $\theta > 0$, and for each cluster the cluster-specific parameter ϕ is an independent draw from G_0. Much statistical methodology built on the Dirichlet process model only uses this joint distribution of the $\{\phi_i\}$, and so should hardly be called 'non-parametric'. Of course, even though G itself is invisible in two of the derivations above, the Dirichlet process model does support inference about G, but this is seldom exploited in applications.

1.2.6 Multiple notations for partitions

In what follows, we will need to make use of different notations for the random partition induced by the Dirichlet process model, or its relatives. We will variously use

- $\boldsymbol{c}$ is a partition of $\{1, 2, \ldots, n\}$
- clusters of partition are $C_1, C_2, \ldots, C_d$ (d is the *degree* of the partition): $\bigcup_{j=1}^{d} C_j = \{1, 2, \ldots, n\}$, $C_j \cap C_{j'} = \emptyset$ if $j \neq j'$
- c is the allocation vector: $c_i = j$ if and only if $i \in C_j$

Note that the first of these makes no use of the (arbitrary) labelling of the clusters used in the second and third. We have to take care with multiplicities, and the distinction between (labelled) allocations and (unlabelled) partitions.

1.3 Applications and generalisations

1.3.1 Some applications of the Dirichlet process in Bayesian nonparametrics

Lack of space precludes a thorough discussion of the huge statistical methodology literature exploiting the Dirichlet process in Bayesian nonparametric procedures, so we will only review a few highlights.

Lo (1984) proposed density estimation procedures devised by mixing a user-defined kernel function $K(y, u)$ with respect to a Dirichlet process, thus i.i.d. data $\{Y_i\}$ are assumed distributed as $\int K(\cdot, u)G(du)$ with G drawn from a Dirichlet process; this is now known as the Dirichlet process mixture model (a better terminology that the formerly-used 'mixture of Dirichlet processes'). The formulation is identical to that we started with in Section 1.2, but for the implicit assumption that y and u lie in the same space, and that the kernel $K(\cdot, u)$ is a unimodal density located near u.

In the 1990's there was a notable flourishing of applied Bayesian nonparametrics, stimulated by interest in the Dirichlet process, and the rapid increase computational power available to researchers, allowing almost routine use of the Pólya urn sampler approach (see Section 1.4) to posterior computation. For example, Escobar (1994) re-visited the Normal Means problem, West et al. (1994) discussed regression and density estimation, and Escobar and West (1995) further developed Bayesian density estimation. Müller et al. (1996) ingeniously exploited multivariate density estimation using Dirichlet process mixtures to perform Bayesian curve fitting of one margin on the others.

1.3.2 Example: clustered linear models for gene expression profiles

Let us consider a substantial and more specific application in some detail, to motivate the Dirichlet process (DP) set-up as a natural elaboration of a standard parametric Bayesian hierarchical model approach.

A remarkable aspect of modern microbiology has been the dramatic development of novel high-throughput assays, capable of delivering very high dimensional quantitative data on the genetic characteristics of organisms from biological samples. One such technology is the measurement of gene expression using Affymetrix gene chips. In Lau and Green (2007), we work with possibly replicated gene expression measures. The data are $\{Y_{isr}\}$, indexed by

- genes $i = 1, 2, \ldots, n$
- conditions $s = 1, 2, \ldots, S$, and
- replicates $r = 1, 2, \ldots, R_s$

Typically R_s is very small, S is much smaller than n, and the 'conditions' represent different subjects, different treatments, or different experimental settings.

We suppose there is a k-dimensional ($k \leq S$) covariate vector x_s describing each condition, and model parametric dependence of Y on x; the focus of interest is on the pattern of variation in these gene-specific parameters across the assayed genes.

Although other variants are easily envisaged, we suppose here that

$$Y_{isr} \sim N(x_s'\beta_i, \tau_i^{-1}), \quad \text{independently.}$$

Here $\phi_i = (\beta_i, \tau_i) \in \mathcal{R}^{k+1}$ is a gene-specific parameter vector characterising the dependence of gene expression on the condition-specific covariates. *A priori*, the genes can be considered exchangeable, and a standard hierarchical formulation would model the $\{\phi_i\}$ as i.i.d. draws from a parametric prior distribution G, say, whose (hyper)parameters have unknown values. This set-up allows borrowing of strength across genes in the interest of stability and efficiency of inference.

The natural nonparametric counterpart to this would be to suppose instead that G, the distribution describing variation of ϕ across the population of genes, does not have prescribed parametric form, but is modelled as a random distribution from a 'nonparametric' prior such as the Dirichlet process, specifically

$$G \sim DP(\theta, G_0)$$

A consequence of this assumption, as we have seen, is that G is atomic, so that the genes will be clustered together into groups sharing a common value of ϕ. *A posteriori* we obtain have a probabilistic clustering of the gene expression profiles.

Lau and Green (2007) take a standard normal–inverse Gamma model, so that $\phi = (\beta, \tau) \sim G_0$ means

$$\tau \sim \Gamma(a_0, b_0) \quad \text{and} \quad \beta|\tau \sim \mathrm{N}_k(m_0, (\tau t_0)^{-1} I)$$

This is a conjugate set-up, so that (β, τ) can be integrated out *in each cluster*. This leads easily to explicit within-cluster parameter posteriors:

$$\begin{aligned} \tau_j^\star | Y &\sim \Gamma(a_j, b_j) \\ \beta_j^\star | \tau_j^\star, Y &\sim \mathrm{N}_k(m_j, (\tau_j^\star t_j)^{-1}) \end{aligned}$$

where

$$\begin{aligned} a_j &= a_0 + 1/2\#\{isr : c_i = j\} \\ b_j &= b_0 + 1/2(Y_{C_j} - X_{C_j} m_0)'(X_{C_j} t_0^{-1} X'_{C_j})^{-1}(Y_{C_j} - X_{C_j} m_0) \\ m_j &= (X'_{C_j} X_{C_j} + t_0 I)^{-1}(X'_{C_j} Y_{C_j} + t_0 m_0) \\ t_j &= X'_{C_j} X_{C_j} + t_0 I. \end{aligned}$$

The marginal likelihoods $p(Y_{C_j})$ are multivariate t distributions.

We continue this example later, in Sections 1.5.4 and 1.5.5.

1.3.3 Generalisations of the Dirichlet process, and related models

Viewed as a nonparametric model or as a basis for probabilistic clustering, the Dirichlet process is simple but inflexible – a single real parameter θ controls both variation and concentration, for example. And although the space Ω where the base measure G_0 lies and in which ϕ lives can be rather general, it is essentially a model for 'univariate' variation and unable to handle in a flexible way, for example, time series data.

Driven both by such considerations of statistical modelling (Walker et al., 1999), or curious pursuit of more general mathematical results, the Dirichlet process has proved a fertile starting point for numerous generalisations, and we touch on just a few of these here.

The Poisson–Dirichlet distribution and its two-parameter generalisation. Kingman (1975) observed and exploited the fact that the limiting behaviour of random discrete distributions could become non-trivial and accessible through permutation of the components to be in ranked (decreasing) order. The limit law is the Poisson–Dirichlet distribution, implicitly defined and later described (Kingman, 1993, page 98) as 'rather less than user-friendly'.

Donnelly and Joyce (1989) elucidated the role of both ranking and size-biased sampling in establishing limit laws for random distributions; see also Holst (2001) and Arratia et al. (2003, page 107). The two-parameter generalisation of the Poisson–Dirichlet model was discovered by Pitman and co-workers, see for example Pitman and Yor (1997). This has been a rich topic for probabilistic study to the present day; see chapters by Gnedin, Haulk and Pitman, and by Aldous in this volume. The simplest view to take of the two-parameter Poisson–Dirichlet model

PD(α, θ) is to go back to stick-breaking (Section 1.2.2) and replace the Beta($1, \theta$) distribution for the variables V_j there by Beta($1 - \alpha, \theta + j\alpha$).

Ishwaran and James (2001) have considered Bayesian statistical applications of stick-breaking priors defined in this way, and implementation of Gibbs sampling for computing posterior distributions.

Dirichlet process relations in structured dependent models. Motivated by the need to build statistical models for structured data of various kinds, there has been a huge effort in generalising Dirichlet process models for such situations – indeed, there is now an 'xDP' for nearly every letter of the alphabet.

This has become a rich and sometimes confusing area; perhaps the most important current models are Dependent Dirichlet processes (MacEachern, 1999; MacEachern et al., 2001), Order-based dependent Dirichlet processes (Griffin and Steel, 2006), Hierarchical Dirichlet processes (Teh et al., 2006), and Kernel stick breaking processes (Dunson and Park, 2007). Many of the models are based on stick-breaking representations, but in which the atoms and/or the weights for the representations of different components of the process are made dependent on each other, or on covariates. The new book by Hjort et al. (2010) provides an excellent introduction and review of these developments.

Pólya trees. Ferguson's definition of the Dirichlet process focussed on the (random) probabilities to be assigned to arbitrary partitions (Section 1.2.1). As we have seen, the resulting distributions G are almost surely discrete. An effective way to modify this process to control continuity properties is to limit the partitions to which elementary probabilities are assigned, and in the case of Pólya tree processes this is achieved by imposed a fixed binary partition of Ω, and assigning probabilities to successive branches in the tree through independent Beta distributions. The parameters of these distributions can be set to obtain various degrees of smoothness of the resulting G. This approach, essentially beginning with Ferguson himself, has been pursued by Lavine (1992, 1994); see also Walker et al. (1999).

1.4 Pólya urn schemes and MCMC samplers

There is a huge literature on Markov chain Monte Carlo methods for posterior sampling in Dirichlet mixture models (MacEachern, 1994; Escobar and West, 1995; Müller et al., 1996; MacEachern and Müller, 1998; Neal, 2000; Green and Richardson, 2001). Although these models have 'variable dimension', the posteriors can be sampled without necessarily using reversible jump methods (Green, 1995).

Cases where G_0 is not conjugate to the data model $f(\cdot|\phi)$ demand keeping $\{\phi_i\}$ in the state vector, to be handled through various augmentation or reversible jump schemes. In the conjugate case, however, it is obviously appealing to integrate ϕ out, and target Markov chain on the posterior solely of the partition, generating ϕ values subsequently as needed. To discuss this, we first go back to probability theory.

1.4.1 The Pólya urn representation of the Dirichlet process

The Pólya urn is a simple and well-known discrete probability model for a reinforcement process: coloured balls are drawn sequentially from an urn; after each is drawn it is replaced, together with a new ball of the same colour. This idea can be seen in a generalised form, in a recursive definition of the joint distribution of the $\{\phi_i\}$.

Suppose that for each $n = 0, 1, 2, \ldots$,

$$\phi_{n+1}|\phi_1, \phi_2, \ldots, \phi_n \sim \frac{1}{n+\theta}\sum_{i=1}^{n}\delta_{\phi_i} + \frac{\theta}{n+\theta}G_0, \tag{1.4}$$

where $\theta > 0$, G_0 is an arbitrary probability distribution, and δ_ϕ is a point probability mass at ϕ. Blackwell and MacQueen (1973) termed such a sequence a Pólya sequence; they showed that the conditional distribution on the right hand side of (1.4) converges to a random probability distribution G distributed as $DP(\theta, G_0)$, and that, given G, $\phi_1, \phi_2, \ldots$ are i.i.d. distributed as G. See also Antoniak (1974) and Pitman (1995).

Thus we have yet another approach to defining the Dirichlet process, at least in so far as specifying the joint distribution of the $\{\phi_i\}$ is concerned. This representation has a particular role, of central importance in computing inferences in DP models. This arises directly from (1.4) and the exchangeability of the $\{\phi_i\}$, for it follows that

$$\phi_i|\phi_{-i} \sim \frac{1}{n-1+\theta}\sum_{j\neq i}\delta_{\phi_j} + \frac{\theta}{n-1+\theta}G_0, \tag{1.5}$$

where ϕ_{-i} means $\{\phi_j : j = 1, 2, \ldots, n, j \neq i\}$. In this form, the statement has an immediate role as the *full conditional* distribution for each component of $(\phi_i)_{i=1}^n$, and hence defines a Gibbs sampler update in a

Markov chain Monte Carlo method aimed at this target distribution. By conjugacy this remains true, with obvious changes set out in the next section, for posterior sampling as well.

The Pólya urn representation of the Dirichlet process has been the point of departure for yet another class of probability models, namely species sampling models (Pitman, 1995, 1996), that are beginning to find a use in statistical methodology (Ishwaran and James, 2003).

1.4.2 The Gibbs sampler for posterior sampling of allocation variables

We will consider posterior sampling in the conjugate case in a more general setting, specialising back to the Dirichlet process mixture case later. The set-up we will assume is based on a partition model: it consists of a prior distribution $p(\boldsymbol{c}|\theta)$ on partitions $\boldsymbol{c}$ of $\{1, 2, \ldots, n\}$ with hyperparameter θ, together with a conjugate model within each cluster. The prior on the cluster-specific parameter ϕ_j has hyperparameter ψ, and is conjugate to the likelihood, so that for any subset $C \subseteq \{1, 2, \ldots, n\}$, $p(Y_C|\psi)$ is known explicitly, where Y_C is the subvector of $(Y_i)_{i=1}^n$ with indices in C. We have

$$p(Y_C|\psi) = \int \prod_{i \in C} p(Y_i|\phi)p(\phi|\psi)d\psi$$

We first consider only re-allocating a single item at a time (the single-variable Gibbs sampler for c_i). Then repeatedly we withdraw an item, say i, from the model, and reallocate it to a cluster according to the full conditional for c_i, which is proportional to $p(\boldsymbol{c}|Y, \theta, \psi)$. It is easy to see that we have two choices:

- allocate Y_i to a new cluster $C_\star$, with probability
$$\propto p(\boldsymbol{c}^{i\to\star}|\theta) \times p(Y_i|\psi),$$
where $\boldsymbol{c}^{i\to\star}$ denotes the current partition $\boldsymbol{c}$ with i moved to $C_\star$, or
- allocate Y_i to cluster C_j^{-i}, with probability
$$\propto p(\boldsymbol{c}^{i\to j}|\theta) \times p(Y_{C_j^{-i}\cup\{i\}}|\psi)/p(Y_{C_j^{-i}}|\psi).$$
where $\boldsymbol{c}^{i\to j}$ denotes the partition $\boldsymbol{c}$, with i moved to cluster C_j.

The ratio of marginal likelihoods $p(Y|\psi)$ in the second expression can be interpreted as the posterior predictive distribution of Y_i given those observations already allocated to the cluster, i.e. $p(Y_i|Y_{C_j^{-i}}, \psi)$ (a multivariate t for the Normal–inverse gamma set-up from Section 1.3.2).

For Dirichlet mixtures we have, from (1.1),

$$p(\boldsymbol{c}|\theta) = \frac{\Gamma(\theta)}{\Gamma(\theta+n)}\theta^d \prod_{j=1}^{d}(n_j - 1)!$$

where $n_j = \#C_j$ and $\boldsymbol{c} = (C_1, C_2, \ldots, C_d)$, so the re-allocation probabilities are explicit and simple in form.

But the same sampler can be used for many other partition models, and the idea is not limited to moving one item at a time.

1.4.3 When the Pólya urn sampler applies

All we require of the model for the Pólya urn sampler to be available for posterior simulation are that

1. A partition $\boldsymbol{c}$ of $\{1, 2, \ldots, n\}$ is drawn from a prior distribution with parameter θ
2. Conditionally on $\boldsymbol{c}$, parameters $(\phi_1, \phi_2, \ldots, \phi_d)$ are drawn independently from a distribution G_0 (possibly with a hyperparameter ψ)
3. Conditional on $\boldsymbol{c}$ and on $\phi = (\phi_1, \phi_2, \ldots, \phi_d)$, $\{y_1, y_2, \ldots, y_n\}$ are drawn independently, from not necessarily identical distributions $p(y_i|\boldsymbol{c}, \phi) = f_i(y_i|\phi_j)$ for $i \in C_j$, for which G_0 is conjugate.

If these all hold, then the Pólya urn sampler can be used; we see from Section 1.4.2 that it will involve computing only marginal likelihoods, and ratios of the partition prior, up to a multiplicative constant. The first factor depends only on G_0 and the likelihood, the second only on the partition model.

Examples $p(\boldsymbol{c}^{i\to\star}|\theta)$ and $p(\boldsymbol{c}^{i\to j}|\theta)$ are proportional simply to

- θ and $\#C_j^{-i}$ for the DP mixture model
- $(k{-}d(\boldsymbol{c}^{-i}))\delta$ and $\#C_j^{-i}{+}\delta$ for the Dirichlet–multinomial finite mixture model
- $\theta{+}\alpha d(\boldsymbol{c}^{-i})$ and $\#C_j^{-i}{-}\alpha$ for the Kingman–Pitman–Yor two-parameter Poisson–Dirichlet process (Section 1.3.3)

It is curious that the ease of using the Pólya urn sampler has often been cited as motivation to use Dirichlet process mixture models, when the class of models for which it is equally readily used is so wide.

1.4.4 Simultaneous re-allocation

There is no need to restrict to updating only one c_i at a time: the idea extends to simultaneously re-allocating any subset of items *currently in the same cluster*.

The notation can be rather cumbersome, but again the subset forms a new cluster, or moves to an existing cluster, with relative probabilities that are each products of two terms:

- the relative (new) partition prior probabilities, and
- the predictive density of the moved set of item data, given those already in the receiving cluster

A more sophisticated variant on this scheme has been proposed by Nobile and Fearnside (2007), and studied in the case of finite mixture models.

1.5 A Coloured Dirichlet process

For the remainder of this note, we focus on the use of these models for clustering, rather than density estimation or other kinds of inference. There needs to be a small caveat – mixture models are commonly used either for clustering, or for fitting non-standard distributions; in a problem demanding *both*, we cannot expect to be able meaningfully to identify clusters with the components of the mixture, since multiple components may be needed to fit the non-standard distributional shape within each cluster. Clustered Dirichlet process methodology in which there is clustering at two levels that can be used for such a purpose is under development by Dan Merl and Mike West at Duke (personal communication).

Here we will not pursue this complication, and simply consider a mixture model used for clustering in the obvious way.

In many domains of application, practical considerations suggest that the clusters in the data do not have equal standing; the most common such situation is where there is believed to be a 'background' cluster, and one or several 'foreground' clusters, but more generally, we can imagine there being several classes of cluster, and our prior beliefs are represented by the idea that cluster labels are exchangeable within these classes, but not overall. It would be common, also, to have different beliefs about cluster-specific parameters within each of these classes.

In this section, we present a variant on standard mixture/cluster models of the kinds we have already discussed, aimed at modelling this situation of partial exchangeability of cluster labels. We stress that it will remain true that, *a priori*, item labels are exchangeable, and that we have no prior information that particular items are drawn to particular classes of cluster; the analysis is to be based purely on the data $\{Y_i\}$.

We will describe the class of a cluster henceforth as its 'colour'. To define a variant on the DP in which not all clusters are exchangeable:

1. for each 'colour' $k = 1, 2, \ldots$, draw G_k from a Dirichlet process $\mathrm{DP}(\theta_k, G_{0k})$, independently for each k
2. draw weights (w_k) from the Dirichlet distribution $\mathrm{Dir}(\gamma_1, \gamma_2, \ldots)$, independently of the G_k.
3. define G on $\{k\} \times \Omega$ by $G(k, B) = w_k G_k(B)$.
4. draw colour–parameter pairs (k_i, ϕ_i) i.i.d from G, $i = 1, 2, \ldots, n$

This process, denoted $\mathrm{CDP}(\{(\gamma_k, \theta_k, G_{0k})\})$, is a Dirichlet mixture of Dirichlet processes (with different base measures), $\sum_k w_k \mathrm{DP}(\theta_k, G_{0k})$, with the added feature that the the colour of each cluster is identified (and indirectly observed), while labelling of clusters within colours is arbitrary.

It can be defined by a 'stick-breaking-and-colouring' construction:

1. colour segments of the stick using the Dirichlet($\{\gamma_k\}$)-distributed weights
2. break each coloured segment using an infinite sequence of independent $\mathrm{Beta}(1, \theta_k)$ variables V_{jk}
3. draw $\phi^\star_{jk} \sim G_{0k}$, i.i.d., $j = 1, 2, \ldots; k = 1, 2, \ldots$
4. define G_k to be the discrete distribution putting probability $(1 - V_{1k})(1 - V_{2k}) \ldots (1 - V_{j-1,k})V_{jk}$ on $\phi^\star_{jk}$

Note that in contrast to other elaborations to more structured data of the Dirichlet process model, in which the focus is on nonparametric analysis and more sharing of information would be desirable, here, where the focus is on clustering, we are content to leave the atoms and weights within each colour completely uncoupled *a priori*.

1.5.1 Coloured partition distribution

The coloured Dirichlet process (CDP) generates the following partition model: partition $\{1, 2, \ldots, n\} = \bigcup_k \bigcup_{j=1}^{d_k} C_{kj}$ at random, where C_{kj} is the jth cluster of colour k, so that

$$p(C_{11}, C_{12}, \ldots, C_{1d_1}; C_{21}, \ldots, C_{2d_2}; C_{31}, \ldots) =$$

$$\frac{\Gamma(\sum_k \gamma_k)}{\Gamma(n + \sum_k \gamma_k)} \prod_k \left(\frac{\Gamma(\theta_k)\Gamma(n_k + \gamma_k)}{\Gamma(n_k + \theta_k)\Gamma(\gamma_k)} \theta_k^{d_k} \prod_{j=1}^{d_k} (n_{kj} - 1)! \right)$$

where $n_{kj} = \#C_{kj}$, $n_k = \sum_j n_{kj}$.

It is curious to note that this expression simplifies when $\theta_k \equiv \gamma_k$, although such a choice seems to have no particular significance in the probabilistic construction of the model. Only when it is also true that the θ_k are independent of k (and the colours are ignored) does the model degenerate to an ordinary Dirichlet process.

The clustering remains exchangeable over items. To complete the construction of the model, analogously to Section 1.2.4, for $i \in C_{kj}$, we set $k_i = k$ and $\phi_i = \phi_j^\star$, where $\phi_j^\star$ are drawn i.i.d. from G_{0k}.

1.5.2 Pólya urn sampler for the CDP

The explicit availability of the (coloured) partition distribution immediately allows generalisation of the Pólya urn Gibbs sampler to the CDP.

In reallocating item i, let n_{kj}^{-i} denote the number *among the remaining items* currently allocated to C_{kj}, and define n_k^{-i} accordingly. Then reallocate i to

- a new cluster of colour k, with probability $\propto \theta_k \times (\gamma_k + n_k^{-i})/(\theta_k + n_k^{-i}) \times p(Y_i|\psi)$, for $k = 1, 2, \ldots$
- the existing cluster C_{kj}, with probability $\propto n_{kj}^{-i} \times (\gamma_k + n_k^{-i})/(\theta_k + n_k^{-i}) \times p(Y_i|Y_{C_{kj}^{-i}}, \psi)$, for $j = 1, 2, \ldots, n_k^{-i}; k = 1, 2, \ldots$

Again, the expressions simplify when $\theta_k \equiv \gamma_k$.

1.5.3 A Dirichlet process mixture with a background cluster

In many applications of probabilistic clustering, including the gene expression example from Section 1.3.2, it is natural to suppose a 'background' cluster that is not *a priori* exchangeable with the others. One way to think about this is to adapt the 'limit of finite mixtures' view from Section 1.2.3:

1. Draw $(w_0, w_1, w_2, \ldots, w_k) \sim \text{Dirichlet}(\gamma, \delta, \ldots, \delta)$

2. Draw $c_i \in \{0, 1, \ldots, k\}$ with $P\{c_i = j\} = w_j$, i.i.d., $i = 1, \ldots, n$
3. Draw $\phi_0^\star \sim H_0$, $\phi_j^\star \sim G_0$, i.i.d., $j = 1, \ldots, k$
4. Set $\phi_i = \phi_{c_i}^\star$

Now let $k \to \infty$, $\delta \to 0$ such that $k\delta \to \theta$, but leave γ fixed. The cluster labelled 0 represents the 'background'.

The background cluster model is a special case of the CDP, specifically CDP$(\{(\gamma, 0, H_0), (\theta, \theta, G_0)\})$. The two colours correspond to the background and regular clusters. The limiting-case DP$(0, H_0)$ is a point mass, randomly drawn from H_0. We can go a little further in a regression setting, and allow different regression models for each colour.

The Pólya urn sampler for prior or posterior simulation is readily adapted. When re-allocating item i, there are three kinds of choice: a new cluster $C_\star$, the 'top table' C_0, or a regular cluster $C_j, j \neq 0$: the corresponding prior probabilities $p(\boldsymbol{c}^{i\to\star}|\theta)$, $p(\boldsymbol{c}^{i\to 0}|\theta)$ and $p(\boldsymbol{c}^{i\to j}|\theta)$ are proportional to θ, $(\gamma + \#C_0^{-i})$ and $\#C_j^{-i}$ for the background cluster CDP model.

1.5.4 Using the CDP in a clustered regression model

As a practical illustration of the use of the CDP background cluster model, we discuss a regression set-up that expresses a vector of measurements $\mathbf{y}_i = (y_{i1}, \ldots, y_{iS})$ for $i = 1 \ldots, n$, where S is the number of samples, as a linear combination of known covariates, $(\mathbf{z}_1 \cdots \mathbf{z}_S)$ with dimension K' and $(\mathbf{x}_1 \cdots \mathbf{x}_S)$ with dimension K. These two collections of covariates, and the corresponding regression coefficients $\boldsymbol{\delta}_j$ and $\boldsymbol{\beta}_j$. are distinguished since we wish to hold one set of regression coefficients fixed in the background cluster. We assume

$$\mathbf{y}_i = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{iS} \end{bmatrix} = \sum_{k'=1}^{K'} \delta_{jk'} \begin{bmatrix} z_{1k'} \\ \vdots \\ z_{Sk'} \end{bmatrix} + \sum_{k=1}^{K} \beta_{jk} \begin{bmatrix} x_{1k} \\ \vdots \\ x_{Sk} \end{bmatrix} + \begin{bmatrix} \epsilon_{j1} \\ \vdots \\ \epsilon_{jS} \end{bmatrix}$$
$$= [\mathbf{z}_1 \cdots \mathbf{z}_S]'\boldsymbol{\delta}_j + [\mathbf{x}_1 \cdots \mathbf{x}_S]'\boldsymbol{\beta}_j + \boldsymbol{\epsilon}_j \qquad (1.6)$$

where $\boldsymbol{\epsilon}_j \sim N(\mathbf{0}_{S\times 1}, \tau_j^{-1}\mathbf{I}_{S\times S})$ and $\mathbf{0}_{S\times 1}$ is the S–dimension zero vector and $\mathbf{I}_{S\times S}$ is the order–S identity matrix. Here, $\boldsymbol{\delta}_j$, $\boldsymbol{\beta}_j$ and τ_j are cluster-specific parameters. The profile of measurements for individual i is $\mathbf{y}_i = [y_{i1} \cdots y_{iS}]'$ for $i = 1, \ldots, n$. Given the covariates $\mathbf{z}_s = [z_{s1} \cdots z_{sK'}]'$, $\mathbf{x}_s = [x_{s1} \cdots x_{sK}]'$, and the cluster j, the parameters/latent variables are $\boldsymbol{\delta}_j = [\delta_{j1} \cdots \delta_{jK'}]'$, $\boldsymbol{\beta}_j = [\beta_{j1} \cdots \beta_{jK}]'$ and τ_j. The kernel is now represented as $k(\mathbf{y}_i|\boldsymbol{\delta}_j, \boldsymbol{\beta}_j, \tau_j)$ and which is a multivariate Normal density, $N([\mathbf{z}_1 \cdots \mathbf{z}_S]'\boldsymbol{\delta}_j + [\mathbf{x}_1 \cdots \mathbf{x}_S]'\boldsymbol{\beta}_j, \tau_j^{-1}\mathbf{I}_{S\times S})$. In particular, we take different probability measures, the parameters of heterogeneous DP, for the background and regular clusters,

$$\begin{aligned}\mathbf{u}_0 = (\boldsymbol{\delta}_0, \boldsymbol{\beta}_0, \tau_0) &\sim H_0(d\boldsymbol{\delta}_0, d\boldsymbol{\beta}_0, d\tau_0)\\ &= \delta_{\boldsymbol{\delta}_0}(d\boldsymbol{\delta}_0) \times \text{Normal–Gamma}(d\boldsymbol{\beta}_0, d\tau_0^{-1})\end{aligned}$$

$$\begin{aligned}\mathbf{u}_j = (\boldsymbol{\delta}_j, \boldsymbol{\beta}_j, \tau_j) &\sim G_0(d\boldsymbol{\delta}_j, d\boldsymbol{\beta}_j, d\tau_j)\\ &= \text{Normal–Gamma}(d(\boldsymbol{\delta}_j', \boldsymbol{\beta}_j')', d\tau_j^{-1})\\ &\qquad \text{for } j = 1, \ldots, n(\mathbf{p}) - 1\end{aligned}$$

Here H_0 is a probability measure that includes a point mass at $\boldsymbol{\delta}_0$ and a Normal–Gamma density for $\boldsymbol{\beta}_0$ and τ_0^{-1}. On the other hand, we take G_0 to be a probability measure that is a Normal–Gamma density for $(\boldsymbol{\delta}_j', \boldsymbol{\beta}_j')'$ and τ_j^{-1}. Thus the regression parameters corresponding to the z covariates are held fixed at $\boldsymbol{\delta}_0$ in the background cluster, but not in the others.

We will first discuss the marginal distribution for the regular clusters. Given τ_j, $(\boldsymbol{\delta}_j', \boldsymbol{\beta}_j')'$ follows the $(K' + K)$–dimensional multivariate Normal with mean $\widetilde{\mathbf{m}}$ and variance $(\underset{\sim}{\tau_j}\widetilde{\mathbf{t}})^{-1}$ and τ_j follows the univariate Gamma with shape $\widetilde{a}$ and scale $\widetilde{b}$. We denote the joint distribution $G_0(d(\boldsymbol{\delta}_j', \boldsymbol{\beta}_j')', d\tau_j)$ as a joint Gamma and Normal distribution, Normal–Gamma$(\widetilde{a}, \widetilde{b}, \widetilde{\mathbf{m}}, \widetilde{\mathbf{t}})$ and further we take

$$\widetilde{\mathbf{m}} = \begin{bmatrix}\widetilde{\mathbf{m}}_\delta\\ \widetilde{\mathbf{m}}_\beta\end{bmatrix} \text{ and } \widetilde{\mathbf{t}} = \begin{bmatrix}\widetilde{\mathbf{t}}_\delta & 0\\ 0 & \widetilde{\mathbf{t}}_\beta\end{bmatrix} \tag{1.7}$$

Based on this set-up, we have

$$\begin{aligned}&m_{G_0}(\mathbf{y}_{C_j}) =\\ &t_{2\widetilde{a}}(\mathbf{Y}_{C_j}|\mathbf{Z}_{C_j}\widetilde{\mathbf{m}}_\delta + \mathbf{X}_{C_j}\widetilde{\mathbf{m}}_\beta, \frac{\widetilde{b}}{\widetilde{a}}(\mathbf{Z}_{C_j}\widetilde{\mathbf{t}}_\delta^{-1}\mathbf{Z}_{C_j}' + \mathbf{X}_{C_j}\widetilde{\mathbf{t}}_\beta^{-1}\mathbf{X}_{C_j}' + \mathbf{I}_{e_jS\times e_jS}))\end{aligned} \tag{1.8}$$

where $\mathbf{Y}_{C_j} = [\mathbf{y}_{i_1}' \cdots \mathbf{y}_{i_{e_j}}']'$, $\mathbf{X}_{C_j} = [[\mathbf{x}_1 \cdots \mathbf{x}_S] \cdots [\mathbf{x}_1 \cdots \mathbf{x}_S]]'$ and $\mathbf{Z}_{C_j} = [[\mathbf{z}_1 \cdots \mathbf{z}_S] \cdots [\mathbf{z}_1 \cdots \mathbf{z}_S]]'$ for $C_j = \{i_1, \ldots, i_{e_j}\}$. Note that $\mathbf{Y}_{C_j}$ is a e_jS vector, $\mathbf{Z}_{C_j}$ is a $e_jS \times K'$ matrix and $\mathbf{X}_{C_j}$ is a $e_jS \times K$ matrix. Moreover, $m_{G_0}(\mathbf{y}_{C_j})$ is a multivariate t density with mean $\mathbf{Z}_{C_j}\widetilde{\mathbf{m}}_\delta + \mathbf{X}_{C_j}\widetilde{\mathbf{m}}_\beta$, scale $\frac{\widetilde{b}}{\widetilde{a}}(\mathbf{Z}_{C_j}\widetilde{\mathbf{t}}_\delta^{-1}\mathbf{Z}_{C_j}' + \mathbf{X}_{C_j}\widetilde{\mathbf{t}}_\beta^{-1}\mathbf{X}_{C_j}' + \mathbf{I}_{e_jS\times e_jS}))$and degree of freedom $2\widetilde{a}$.

For the background cluster, we take H_0 to be a joint Gamma and Normal distribution, Normal–Gamma$(\overline{a}, \overline{b}, \overline{\mathbf{m}}_\beta, \overline{\mathbf{t}}_\beta)$. The precision τ_0 follows the univariate Gamma with shape $\overline{a}$ and scale $\overline{b}$. Given τ_0, $\boldsymbol{\beta}_0$ follows the K–dimension multivariate Normal with mean $\overline{\mathbf{m}}_\beta$ and variance $(\tau_0\overline{\mathbf{t}}_\beta)^{-1}$ and τ_0 follows the univariate Gamma with shape $\overline{a}$ and scale $\overline{b}$. The marginal distribution becomes

$$m_{H_0}(\mathbf{y}_{C_0}) = t_{2\overline{a}}(\mathbf{Y}_{C_j}|\mathbf{Z}_{C_j}\boldsymbol{\delta}_0 + \mathbf{X}_{C_j}\overline{\mathbf{m}}_\beta, \frac{\overline{b}}{\overline{a}}(\mathbf{X}_{C_j}\overline{\mathbf{t}}_\beta^{-1}\mathbf{X}'_{C_j} + \mathbf{I}_{e_jS\times e_jS})) \tag{1.9}$$

So, $m_{H_0}(\mathbf{y}_{C_0})$ is a multivariate t density with mean $\mathbf{Z}_{C_j}\boldsymbol{\delta}_0 + \mathbf{X}_{C_j}\overline{\mathbf{m}}_\beta$, scale $\frac{\overline{b}}{\overline{a}}(\mathbf{X}_{C_j}\overline{\mathbf{t}}_\beta^{-1}\mathbf{X}'_{C_j} + \mathbf{I}_{e_jS\times e_jS})$ and degree of freedom $2\overline{a}$.

In some applications, the xs and βs are not needed and so can be omitted, and we consider the following model,

$$\mathbf{y}_i = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{iS} \end{bmatrix} = \sum_{k'=1}^{K'} \delta_{jk'} \begin{bmatrix} z_{1k'} \\ \vdots \\ z_{Sk'} \end{bmatrix} + \begin{bmatrix} \epsilon_{j1} \\ \vdots \\ \epsilon_{jS} \end{bmatrix} = [\mathbf{z}_1 \cdots \mathbf{z}_S]'\boldsymbol{\delta}_j + \boldsymbol{\epsilon}_j \tag{1.10}$$

here we assume that $K = 0$ or $[\mathbf{x}_1 \cdots \mathbf{x}_S]' = \mathbf{0}_{S\times K}$ where $\mathbf{0}_{S\times K}$ is the $S \times K$ matrix with all zero entries of the model (1.6). We can derive the marginal distributions analogous to (1.8) and (1.9),

$$m_{G_0}(\mathbf{y}_{C_j}) = t_{2\widetilde{a}}(\mathbf{Y}_{C_j}|\mathbf{Z}_{C_j}\widetilde{\mathbf{m}}_\delta, \frac{\widetilde{b}}{\widetilde{a}}(\mathbf{Z}_{C_j}\widetilde{\mathbf{t}}_\delta^{-1}\mathbf{Z}'_{C_j} + \mathbf{I}_{e_jS\times e_jS})) \tag{1.11}$$

$$m_{H_0}(\mathbf{y}_{C_0}) = t_{2\overline{a}}(\mathbf{Y}_{C_j}|\mathbf{Z}_{C_j}\boldsymbol{\delta}_0, \frac{\overline{b}}{\overline{a}}\mathbf{I}_{e_jS\times e_jS}) \tag{1.12}$$

Here $t_\nu\,(\mathbf{x}\,|\boldsymbol{\mu}, \boldsymbol{\Sigma}\,)$ is a multivariate t density in d dimensions with mean $\boldsymbol{\mu}$ and scale $\boldsymbol{\Sigma}$ with degrees of freedom ν,

$$t_\nu(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\Gamma((\nu+d)/2)}{\Gamma((\nu)/2)}\frac{|\boldsymbol{\Sigma}|^{-1/2}}{(\nu\pi)^{d/2}}(1 + \frac{1}{\nu}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}))^{-(\nu+d)/2} \tag{1.13}$$

1.5.5 Time course gene expression data

We consider the application of this methodology to data from a gene expression time course experiment. Wen et al. (1998) studied the central nervous system development of the rat; see also Yeung et al. (2001). The mRNA expression levels of 112 genes were recorded over the period of development of the central nervous system development. In the dataset, there are 9 records for each gene over 9 time points, they are from

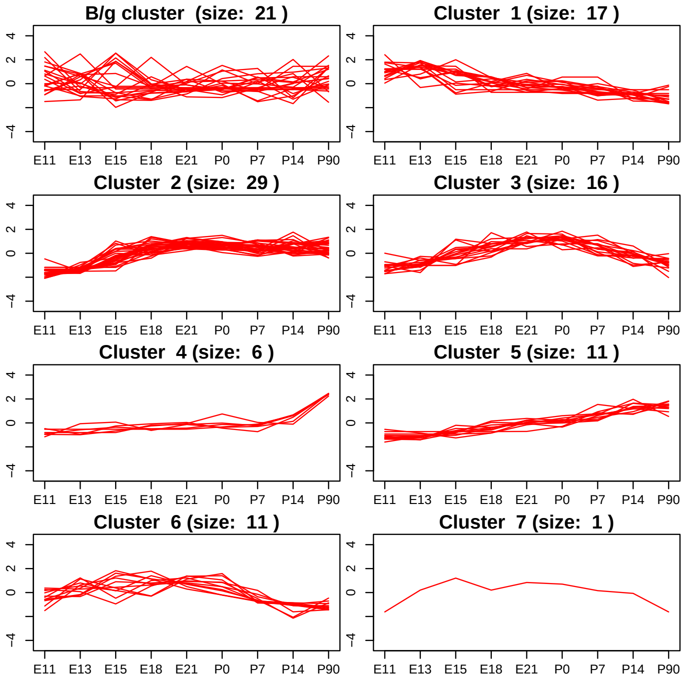

Figure 1.1 Profile plot of our partition estimate for the Rat data set of Wen et al. (1998).

embryonic days 11, 13, 15, 18, 21, postnatal days 0, 7, 14, and the 'adult' stage (postnatal day 90).

In their analysis, Wen et al. (1998) obtained 5 clusters/waves (totally 6 clusters), taken to characterize distinct phases of development. The data set is available at `http://faculty.washington.edu/kayee/cluster/GEMraw.txt`. We take $S = 9$ and $K' = 5$. The design matrix of covariates is taken to be

$$[\mathbf{z}_1 \cdots \mathbf{z}_S]' = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 11 & 13 & 15 & 18 & 21 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 7 & 14 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}'$$

representing piecewise linear dependence on time, within three separate phases (embryonic, postnatal and adult).

In our analysis of these data, we take $\theta = 1$, $\gamma = 5$, $\widetilde{a} = \overline{a} = 0.01$,

$\widetilde{b} = \overline{b} = 0.01$, $\widetilde{\mathbf{m}}_\delta = \overline{\mathbf{m}}_\delta = [0 \cdots 0]'$, $\widetilde{\mathbf{t}}_\delta = \overline{\mathbf{t}}_\delta = 0.01\mathbf{I}$, $\widetilde{\mathbf{m}}_\beta = [0 \cdots 0]'$, $\widetilde{\mathbf{t}}_\beta = 0.01\mathbf{I}$ and $\boldsymbol{\delta}_0 = [0 \cdots 0]'$. The Pólya urn sampler was implemented, and run for 20000 sweeps starting from the partition consisting of all singleton clusters, 10000 being discarded as burn-in. We then use the last 10000 partitions sampled as in Lau and Green (2007), to estimate the optimal Bayesian partition on a decision-theoretic basis, using a pairwise coincidence loss function that equally weights false 'positives' and 'negatives'.

We present some views of the resulting posterior analysis of this data set.

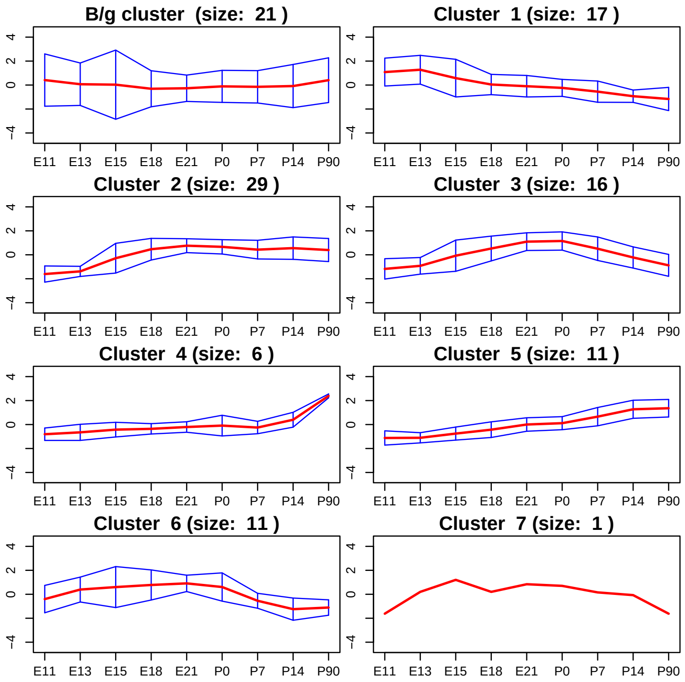

Figure 1.2 Mean and 95% CI of genes across clusters of our partition estimate.

Figure 1.1 shows the profiles in the inferred clusters plotted, and Figure 1.2 the mean and the 95% CI of the clusters. Figure 1.3 cross-tabulates the clusters with the biological functions attributed to the relevant genes by Wen et al. (1998).

	General gene class				Neurotransmitter receptors					
					Ligand class				Sequence class	
	% peptide signaling	% neurotr. receptors	% neuroglial markers	% diverse	% ACh	% GABA	% Glu	% 5HT	% ion channel	% G protein coupled
1	**41% (7)**	6% (1)	24% (4)	29% (5)	**100% (1)**				**100% (1)**	
2	3% (1)	**62% (18)**	31% (9)	3% (1)	17% (3)	**39% (7)**	33% (6)	11% (2)	**61% (11)**	39% (7)
3	6% (1)	**63% (10)**	19% (3)	13% (2)	20% (2)	20% (2)	**40% (4)**	20% (2)	**50% (5)**	**50% (5)**
4		33% (2)	17% (1)	**50% (3)**			**100% (2)**		**50% (1)**	**50% (1)**
5	18% (2)	27% (3)	**45% (5)**	9% (1)	33% (1)	**67% (2)**			**67% (2)**	33% (1)
6	27% (3)	18% (2)	18% (2)	**36% (4)**	**50% (1)**		**50% (1)**		**100% (2)**	
7		**100% (1)**			**100% (1)**				**100% (1)**	
B/g	**62% (13)**	10% (2)	5% (1)	24% (5)	**100% (2)**				**100% (2)**	

Figure 1.3 Biological functions of our Bayesian partition estimate for the genes in the data set of Wen et al. (1998), showing correspondence between inferred clusters and the functional categories of the genes. All genes are classified into 4 general gene classes. Additionally, the Neurotransmitter genes have been further categorised by ligand class and functional sequence class. Boldface type represents the dominant class in the cluster, in each categorisation.

Acknowledgements

I am grateful to Sonia Petrone and Simon Tavaré for some pointers to the literature, John Lau for the analysis of the gene expression data, and John Kingman for his undergraduate lectures in Measure and Probability.

References

Antoniak, C. E. 1974. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, **2**, 1152–1174.

Arratia, R., Barbour, A. D., and Tavaré, S. 2003. *Logarithmic combinatorial structures: a probabilistic approach.* Monographs in Mathematics. European Mathematical Society.

Blackwell, D. 1973. Discreteness of Ferguson selections. *The Annals of Statistics*, **1**, 356–358.

Blackwell, D., and MacQueen, J. B. 1973. Ferguson distributions via Pólya Urn Schemes. *The Annals of Statistics*, **1**, 353–355.

de Finetti, Bruno. 1931. Funzione caratteristica di un fenomeno aleatorio. *Atti della R. Academia Nazionale dei Lincei, ser. 6*, **4**, 251–299. Memorie, Classe di Scienze Fisiche, Mathematiche e Narurali.

de Finetti, Bruno. 1937. La prévision: ses lois logiques, ses sources subjectives. *Ann. Inst. H. Poincaré*, **7**, 1–68.

Donnelly, P., and Joyce, P. 1989. Continuity and weak convergence of ranked and size-biased permutations on the infinite simplex. *Stochastic Processes and their Applications*, **31**, 89–103.

Dunson, D.B., and Park, J-H. 2007. Kernel stick breaking processes. *Biometrika*, **95**, 307–323.

Escobar, M. D. 1994. Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association*, **89**, 268–277.

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

Ewens, W. 1972. The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, **3**, 87–112.

Ferguson, T. S. 1973. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, **1**, 209–230.

Gelman, Andrew, Carlin, John B., Stern, Hal S., and Rubin, Donald B. 1995. *Bayesian data analysis.* Chapman and Hall, London.

Green, P. J. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Green, P. J., and Richardson, S. 2001. Modelling heterogeneity with and without the Dirichlet process. *Scandinavian Journal of Statistics*, **28**, 355–375.

Green, Peter J., Hjort, Nils Lid, and Richardson, Sylvia (eds). 2003. *Highly Structured Stochastic Systems.* Oxford University Press, Oxford.

Griffin, J.E., and Steel, M.F.J. 2006. Order-Based Dependent Dirichlet Processes. *Journal of the American Statistical Association*, **101**, 179–194.

Hjort, Nils Lid, Holmes, Chris, Müller, Peter, and Walker, Stephen G. (eds). 2010. *Bayesian Nonparametrics.* Cambridge Series in Statistical and Probabilistic Mathematics (No. 28). Cambride University Press.

Holst, L. 2001. *The Poisson-Dirichlet distribution and its relatives revisited.* Tech. rept. Department of Mathematics, Royal Institute of Technology, Stockholm.

Ishwaran, H., and James, L. F. 2001. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, **96**, 161–173.

Ishwaran, H., and James, L. F. 2003. Generalized weighted chinese restaurant processes for species sampling mixture models. *Statistica Sinica*, **13**, 1211–1235.

Ishwaran, H., and Zarepour, M. 2002. Dirichlet prior sieves in finite Normal mixtures. *Statistica Sinica*, **12**, 941–963.

Kallenberg, Olav. 2005. *Probabilistic Symmetries and Invariance Principles.* Springer, New York.

Kingman, J. F. C. 1967. Completely random measures. *Pacific Journal of Mathematics*, **21**, 59–78.

Kingman, J. F. C. 1975. Random discrete distributions (with discussion). *Journal of the Royal Statistical Society, B*, **37**, 1–22.

Kingman, J. F. C. 1978. Uses of exchangeability. *Annals of Probability*, **6**, 183–197.

Kingman, J. F. C. 1993. *Poisson Processes.* Oxford University Press.

Lau, J. W., and Green, P. J. 2007. Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics*, **16**, 526–558.

Lavine, M. 1992. Some Aspects of Pólya Tree Distributions for Statistical Modelling. *Annals of Statistics*, **20**, 1222–1235.

Lavine, M. 1994. More Aspects of Pólya Tree Distributions for Statistical Modelling. *Annals of Statistics*, **22**, 1161–1176.

Lo, A. Y. 1984. On a class of Bayesian nonparametric estimates: (I) Density estimates. *The Annals of Statistics*, **12**, 351–357.

MacEachern, S. 1999. Dependent Nonparametric Processes. In: *Proceedings of the Section on Bayesian Statistical Science.* American Statistical Association.

MacEachern, S., Kottas, A., and Gelfand, A. 2001. *Spatial Nonparametric Bayesian Models.* Tech. rept. Tech. Rep. 01-10. Institute of Statistics and Decision Sciences, Duke University.

MacEachern, S. N. 1994. Estimating normal means with a conjugate style Dirichlet process prior. *Communication in Statistics: Simulation and Computation*, **23**, 727–741.

MacEachern, S. N., and Müller, P. 1998. Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, **7**, 223–238.

McCloskey, J. W. 1965. *A model for the distribution of species in an environment.* Ph.D. thesis, Michigan State University.

Muliere, P., and Secchi, P. 2003. Weak convergence of a Dirichlet-multinomial process. *Georgian Mathematical Journal*, **10**, 319–324.

Müller, P., Erkanli, A., and West, M. 1996. Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, **83**, 67–79.

Neal, R. M. 2000. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, **9**, 249–265.

Nobile, A., and Fearnside, A. T. 2007. Bayesian finite mixtures with an unknown number of components: the allocation sampler. *Statistics and Computing*, **17**, 147–162.

Patil, C. P., and Taillie, C. 1977. Diversity as a concept and its implications for random communities. *Bull. Int. Statist. Inst.*, **47**, 497–515.

Pitman, J. 1995. Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, **102**, 145–158.

Pitman, J. 1996. Some developments of the Blackwell-MacQueen urn scheme. Pages 245–267 of: T. S. Ferguson, L. S. Shapley, and MacQueen, J. B. (eds), *Statistics, Probability and Game Theory; Papers in Honor of David Blackwell.* Hayward: Institute of Mathematical Statistics.

Pitman, J., and Yor, M. 1997. The two-parameter Poisson–Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, **25**, 855–900.

Richardson, S., and Green, P. J. 1997. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society, B*, **59**, 731–792.

Sethuraman, J. 1994. A constructive definition of Dirichlet priors. *Statistica Sinica*, **4**, 639–650.

Sethuraman, J., and Tiwari, R. C. 1982. *Convergence of Dirichlet Measures and the Interpretation of Their Parameters.* Statistical Decision Theory and Related Topics III, 2, 305-315.

Teh, Yee Whye, Jordan, Michael I., Beal, Matthew J., and Blei, David M. 2006. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, **101**, 1566–1581.

Walker, S. G., Damien, P., Laud, P. W., and Smith, A. F. M. 1999. Bayesian nonparametric inference for random distributions and related functions (with discussion). *Journal of the Royal Statistical Society, B*, **61**, 485–527.

Wen, X., Fuhrman, S., Michaels, G. S., Carr, D. B., Smith, S., Barker, J. L., and Somogyi, R. 1998. Large-scale temporal gene expression mapping of central nervous system development. *Proceedings of the National Academy of Sciences of the United States of America*, 334–339.

West, M., Müller, P., and Escobar, M. D. 1994. Hierarchical priors and mixture models, with application in regression and density estimation. In: Smith, A. F. M, and Freeman, P. (eds), *Aspects of Uncertainty: A tribute to Lindley.* Wiley, New York.

Yeung, K. Y., Haynor, D. R., and Ruzzo, W. L. 2001. Validating clustering for gene expression data. *Bioinformatics*, 309–318.