Bayesian Nonparametrics

Peter Green

University of Bristol
and UTS, Sydney

March 2012 / UTS, Sydney

Outline

Themes

- Bayesians want to be nonparametric as much as frequentists do
- Random effects modelled in a flexible way
- Random distributions
- Exchangeability and de Finetti
- Urn models
- Infinite mixtures

Random distributions

Our statistical model includes a quantity $\theta \in \Omega$, which has distribution G, that is not fully known. To a Bayesian, G has to have a (prior) distribution, so G is a random distribution.

Example: $\Omega = \{0, 1\}$ (or any binary set). The only possible G is Bernoulli, say $G(\{0\}) = 1 - g$, $G(\{1\}) = g$. If g is known so is G.

If G is unknown so is $g \in [0, 1]$; a natural prior for g is Beta(α_0, α_1). A single $\theta \in \Omega$ still has a Bernoulli distribution (with $P(\theta = 1) = E(g) = \alpha_0/(\alpha_0 + \alpha_1)$).

The example begins to have a point (and is a model important in practice) if we have n conditionally independent θ_i from G; then $\sum_i \theta_i$ follows a Beta–Binomial model. The θ_i are unconditionally dependent.

Random distributions

Our statistical model includes a quantity $\theta \in \Omega$, which has distribution G, that is not fully known. To a Bayesian, G has to have a (prior) distribution, so G is a random distribution.

Example: $\Omega = \{0, 1\}$ (or any binary set). The only possible G is Bernoulli, say $G(\{0\}) = 1 - g$, $G(\{1\}) = g$. If g is known so is G.

If G is unknown so is $g \in [0, 1]$; a natural prior for g is Beta(α_0, α_1). A single $\theta \in \Omega$ still has a Bernoulli distribution (with $P(\theta = 1) = E(g) = \alpha_0/(\alpha_0 + \alpha_1)$).

The example begins to have a point (and is a model important in practice) if we have n conditionally independent θ_i from G; then $\sum_i \theta_i$ follows a Beta–Binomial model. The θ_i are unconditionally dependent.

Random distributions

Our statistical model includes a quantity $\theta \in \Omega$, which has distribution G, that is not fully known. To a Bayesian, G has to have a (prior) distribution, so G is a random distribution.

Example: $\Omega = \{0, 1\}$ (or any binary set). The only possible G is Bernoulli, say $G(\{0\}) = 1 - g$, $G(\{1\}) = g$. If g is known so is G.

If G is unknown so is $g \in [0, 1]$; a natural prior for g is Beta(α_0, α_1). A single $\theta \in \Omega$ still has a Bernoulli distribution (with $P(\theta = 1) = E(g) = \alpha_0/(\alpha_0 + \alpha_1)$).

The example begins to have a point (and is a model important in practice) if we have n conditionally independent θ_i from G; then $\sum_i \theta_i$ follows a Beta–Binomial model. The θ_i are unconditionally dependent.

Random distributions

Our statistical model includes a quantity $\theta \in \Omega$, which has distribution G, that is not fully known. To a Bayesian, G has to have a (prior) distribution, so G is a random distribution.

Example: $\Omega = \{0, 1\}$ (or any binary set). The only possible G is Bernoulli, say $G(\{0\}) = 1 - g$, $G(\{1\}) = g$. If g is known so is G.

If G is unknown so is $g \in [0, 1]$; a natural prior for g is Beta(α_0, α_1). A single $\theta \in \Omega$ still has a Bernoulli distribution (with $P(\theta = 1) = E(g) = \alpha_0/(\alpha_0 + \alpha_1)$).

The example begins to have a point (and is a model important in practice) if we have n conditionally independent θ_i from G; then $\sum_i \theta_i$ follows a Beta–Binomial model. The θ_i are unconditionally dependent.

Random distributions

Our statistical model includes a quantity $\theta \in \Omega$, which has distribution G, that is not fully known. To a Bayesian, G has to have a (prior) distribution, so G is a random distribution.

Example: $\Omega = \{1, 2, \ldots, K\}$ (or any finite set). The only possible G is Multinomial, say $G(\{k\}) = g_k$, $k = 1, 2, \ldots, K$. If g is known so is G.

If G is unknown so is g; a natural prior for g is Dirichlet$(\alpha_1, \alpha_2, \ldots \alpha_K)$. A single $\theta \in \Omega$ still has a Multinomial distribution (with $P(\theta = k) = E(g_k) = \alpha_k / (\sum_k \alpha_k)$).

The example begins to have a point (and is a model important in practice) if we have n conditionally independent θ_i from G; then we have a Dirichlet–Multinomial model.

Random distributions

Our statistical model includes a quantity $\theta \in \Omega$, which has distribution G, that is not fully known. To a Bayesian, G has to have a (prior) distribution, so G is a random distribution.

Example: $\Omega = \{1, 2, \ldots, K\}$ (or any finite set). The only possible G is Multinomial, say $G(\{k\}) = g_k$, $k = 1, 2, \ldots, K$. If g is known so is G.

If G is unknown so is g; a natural prior for g is Dirichlet$(\alpha_1, \alpha_2, \ldots \alpha_K)$. A single $\theta \in \Omega$ still has a Multinomial distribution (with $P(\theta = k) = E(g_k) = \alpha_k / (\sum_k \alpha_k)$).

The example begins to have a point (and is a model important in practice) if we have n conditionally independent θ_i from G; then we have a Dirichlet–Multinomial model.

Random distributions

Our statistical model includes a quantity $\theta \in \Omega$, which has distribution G, that is not fully known. To a Bayesian, G has to have a (prior) distribution, so G is a random distribution.

Example: $\Omega = \{1, 2, \ldots, K\}$ (or any finite set). The only possible G is Multinomial, say $G(\{k\}) = g_k$, $k = 1, 2, \ldots, K$. If g is known so is G.

If G is unknown so is g; a natural prior for g is Dirichlet$(\alpha_1, \alpha_2, \ldots \alpha_K)$. A single $\theta \in \Omega$ still has a Multinomial distribution (with $P(\theta = k) = E(g_k) = \alpha_k / (\sum_k \alpha_k)$).

The example begins to have a point (and is a model important in practice) if we have n conditionally independent θ_i from G; then we have a Dirichlet–Multinomial model.

Random distributions

Our statistical model includes a quantity $\theta \in \Omega$, which has distribution G, that is not fully known. To a Bayesian, G has to have a (prior) distribution, so G is a random distribution.

Example: $\Omega = \{1, 2, \ldots, K\}$ (or any finite set). The only possible G is Multinomial, say $G(\{k\}) = g_k$, $k = 1, 2, \ldots, K$. If g is known so is G.

If G is unknown so is g; a natural prior for g is Dirichlet$(\alpha_1, \alpha_2, \ldots \alpha_K)$. A single $\theta \in \Omega$ still has a Multinomial distribution (with $P(\theta = k) = E(g_k) = \alpha_k / (\sum_k \alpha_k)$).

The example begins to have a point (and is a model important in practice) if we have n conditionally independent θ_i from G; then we have a Dirichlet–Multinomial model.

The Dirichlet distribution

The Dirichlet$(\alpha_1, \alpha_2, \ldots \alpha_K)$ distribution has support the unit $(K-1)$-simplex $\{(x_1, x_2, \ldots, x_K) : x_i \geq 0, \sum_i x_i = 1\} \subset \mathcal{R}^K$, and density

$$\frac{\Gamma(\sum_i \alpha_i)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\ldots\Gamma(\alpha_K)} x_1^{\alpha_1 - 1} x_2^{\alpha_2 - 1} \ldots x_K^{\alpha_K - 1}$$

(strictly the density of any $K-1$ dimensional sub-vector).

Conjugacy under i.i.d. sampling

The Dirichlet (beta) prior distribution in these examples is not only 'natural', of course, it is also conjugate to the multinomial (binomial) likelihood.

If $g \sim \text{Dirichlet}(\alpha_1, \alpha_2, \ldots \alpha_K)$ and $\theta_i | g \sim g$, independently, $i = 1, 2, \ldots, n$, then

$$g|\theta_1, \theta_2, \ldots, \theta_n \sim \text{Dirichlet}(\alpha_1 + n_1, \alpha_2 + n_2, \ldots \alpha_K + n_K)$$

where $n_k = \#\{i : \theta_i = k\}$.

The posterior has the same form as the prior, and this assists with interpretation (the prior is like 'initial data') and computation.

Exchangeability

In the Dirichlet–Multinomial (or Beta–Binomial) model, $\{\theta_i\}_{i=1}^n$ are conditionally i.i.d. $\sim G$ given G, but not marginally independent. Rather they are (infinitely) exchangeable.

$\{\theta_i\}_{i=1}^n$ are exchangeable if the joint distribution is invariant to permutations of the labels $i = 1, 2, \ldots, n$.

$\{\theta_i\}_{i=1}^\infty$ are infinitely exchangeable if for every n the joint distribution of any subset $\{\theta_{j_i}\}_{i=1}^n$ of size n is the same.

Exchangeability

In the Dirichlet–Multinomial (or Beta–Binomial) model, $\{\theta_i\}_{i=1}^n$ are conditionally i.i.d. $\sim G$ given G, but not marginally independent. Rather they are (infinitely) exchangeable.

$\{\theta_i\}_{i=1}^n$ are exchangeable if the joint distribution is invariant to permutations of the labels $i = 1, 2, \ldots, n$.

$\{\theta_i\}_{i=1}^\infty$ are infinitely exchangeable if for every n the joint distribution of any subset $\{\theta_{j_i}\}_{i=1}^n$ of size n is the same.

Exchangeability

In the Dirichlet–Multinomial (or Beta–Binomial) model, $\{\theta_i\}_{i=1}^n$ are conditionally i.i.d. $\sim G$ given G, but not marginally independent. Rather they are (infinitely) exchangeable.

$\{\theta_i\}_{i=1}^n$ are exchangeable if the joint distribution is invariant to permutations of the labels $i = 1, 2, \ldots, n$.

$\{\theta_i\}_{i=1}^\infty$ are infinitely exchangeable if for every n the joint distribution of any subset $\{\theta_{j_i}\}_{i=1}^n$ of size n is the same.

Exchangeability and de Finetti

de Finetti's theorem (1931) says that when $\Omega = \{0, 1\}$, the joint distribution of $\{\theta_i\}_{i=1}^{\infty}$ is infinitely exchangeable if and only if for all n

$$p(\theta_1, \theta_2, \ldots, \theta_n) = \int_0^1 \alpha^{t_n}(1-\alpha)^{n-t_n} dF(\alpha)$$

for some distribution F on $[0, 1]$, where $t_n = \sum_{i=1}^{n} \theta_i$.

Further, the limiting frequency $\lim_{n\to\infty} t_n/n$ has distribution F.

Exchangeable binary outcomes are always like Bernoulli trials with a fixed but random success probability!

Exchangeability and de Finetti

de Finetti's theorem (1931) says that when $\Omega = \{0, 1\}$, the joint distribution of $\{\theta_i\}_{i=1}^{\infty}$ is infinitely exchangeable if and only if for all n

$$p(\theta_1, \theta_2, \ldots, \theta_n) = \int_0^1 \alpha^{t_n}(1-\alpha)^{n-t_n} dF(\alpha)$$

for some distribution F on $[0, 1]$, where $t_n = \sum_{i=1}^{n} \theta_i$.

Further, the limiting frequency $\lim_{n\to\infty} t_n/n$ has distribution F.

Exchangeable binary outcomes are always like Bernoulli trials with a fixed but random success probability!

Exchangeability and de Finetti

de Finetti's theorem (1931) says that when $\Omega = \{0, 1\}$, the joint distribution of $\{\theta_i\}_{i=1}^{\infty}$ is infinitely exchangeable if and only if for all n

$$p(\theta_1, \theta_2, \ldots, \theta_n) = \int_0^1 \alpha^{t_n}(1-\alpha)^{n-t_n} dF(\alpha)$$

for some distribution F on $[0, 1]$, where $t_n = \sum_{i=1}^{n} \theta_i$.

Further, the limiting frequency $\lim_{n\to\infty} t_n/n$ has distribution F.

Exchangeable binary outcomes are always like Bernoulli trials with a fixed but random success probability!

Pólya's urn

You start with B blue balls and R red balls in an urn. Repeatedly and indefinitely you draw a ball at random and replace it, adding an additional ball of the same colour. Let $\theta_i = 1$ if the ith ball drawn is blue, otherwise 0.

The joint distribution of $\{\theta_i\}_{i=1}^{\infty}$ is infinitely exchangeable! (Exercise: find an algebra-free proof!)

The limiting frequency $\lim_{n\to\infty} t_n/n$ has distribution Beta(B, R).

Pólya's urn

You start with B blue balls and R red balls in an urn. Repeatedly and indefinitely you draw a ball at random and replace it, adding an additional ball of the same colour. Let $\theta_i = 1$ if the ith ball drawn is blue, otherwise 0.

The joint distribution of $\{\theta_i\}_{i=1}^{\infty}$ is infinitely exchangeable! (Exercise: find an algebra-free proof!)

The limiting frequency $\lim_{n\to\infty} t_n/n$ has distribution Beta(B, R).

Pólya's urn

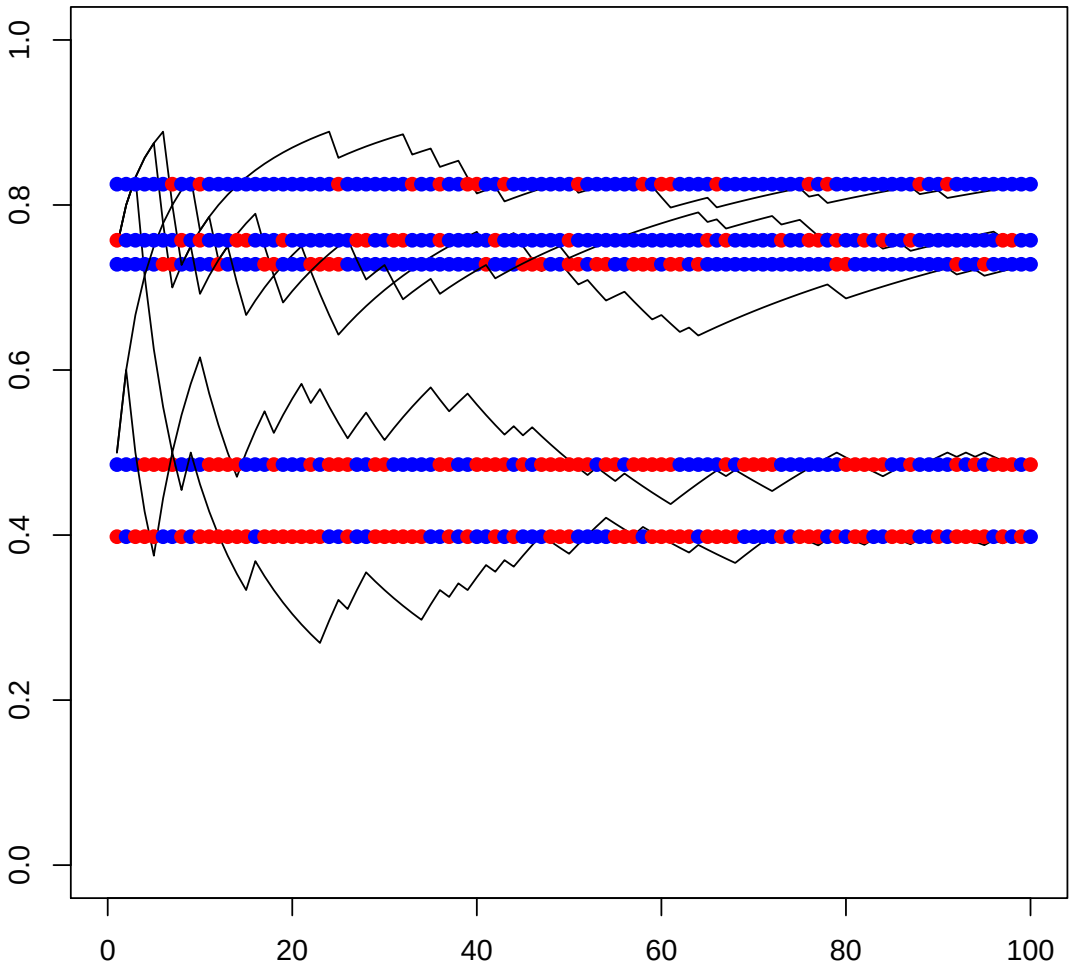
You start with B blue balls and R red balls in an urn. Repeatedly and indefinitely you draw a ball at random and replace it, adding an additional ball of the same colour. Let $\theta_i = 1$ if the ith ball drawn is blue, otherwise 0.

The joint distribution of $\{\theta_i\}_{i=1}^{\infty}$ is infinitely exchangeable! (Exercise: find an algebra-free proof!)

The limiting frequency $\lim_{n\to\infty} t_n/n$ has distribution Beta(B, R).

Pólya's urn

5 replicates of 100 draws from an urn model with $B = 2$, $R = 1$.



Exchangeability and de Finetti

When $\Omega = \{1, 2, \ldots, K\}$, the joint distribution of $\{\theta_i\}_{i=1}^{\infty}$ is infinitely exchangeable if and only if for all n

$$p(\theta_1, \theta_2, \ldots, \theta_n) = \int_0^1 \prod_{k=1}^{K} \alpha_k^{t_{n,k}} dF(\alpha)$$

for some distribution F on the unit $(K-1)$-simplex in $\mathcal{R}^K$, and $t_{n,k} = \#\{i \leq n : \theta_i = k\}$.

Further, the limiting frequencies $(\lim_{n\to\infty} t_{n,k}/n)_{k=1}^{K}$ have distribution F.

Exchangeability and de Finetti

When $\Omega = \{1, 2, \ldots, K\}$, the joint distribution of $\{\theta_i\}_{i=1}^{\infty}$ is infinitely exchangeable if and only if for all n

$$p(\theta_1, \theta_2, \ldots, \theta_n) = \int_0^1 \prod_{k=1}^{K} \alpha_k^{t_{n,k}} dF(\alpha)$$

for some distribution F on the unit $(K-1)$-simplex in $\mathcal{R}^K$, and $t_{n,k} = \#\{i \leq n : \theta_i = k\}$.

Further, the limiting frequencies $(\lim_{n\to\infty} t_{n,k}/n)_{k=1}^{K}$ have distribution F.

Random distributions

Our statistical model includes a quantity $\theta \in \Omega$, which has distribution G, that is not fully known. To a Bayesian, G has to have a (prior) distribution, so G is a random distribution.

What about general Ω? – most statistical models have at least one real parameter! The usual approach would select a parametric model for G – G is a distribution with prescribed functional form and one or more unknown hyperparameters, which may be taken as known, or inferred.

Could we instead leave G unspecified in form, as we could when Ω was finite?

And could we still expect conjugacy to help us?

Random distributions

Our statistical model includes a quantity $\theta \in \Omega$, which has distribution G, that is not fully known. To a Bayesian, G has to have a (prior) distribution, so G is a random distribution.

What about general Ω? – most statistical models have at least one real parameter! The usual approach would select a parametric model for G – G is a distribution with prescribed functional form and one or more unknown hyperparameters, which may be taken as known, or inferred.

Could we instead leave G unspecified in form, as we could when Ω was finite?

And could we still expect conjugacy to help us?

Exchangeability and de Finetti

For general Ω, Hewitt & Savage (1955) showed that the joint distribution of $\{\theta_i\}_{i=1}^{\infty}$ is infinitely exchangeable if and only if for all n

$$P(\theta_1 \in B_1, \theta_2 \in B_2, \ldots, \theta_n \in B_n) = \int_0^1 \prod_{i=1}^{n} Q(B_i)\mu(dQ)$$

for some probability measure μ on the set of probability measures on Ω.

Further, μ is the distribution of the limiting empirical measure $M(B) = n^{-1}\#\{i \leq n : \theta_i \in B\}$.

Exchangeability and de Finetti

For general Ω, Hewitt & Savage (1955) showed that the joint distribution of $\{\theta_i\}_{i=1}^{\infty}$ is infinitely exchangeable if and only if for all n

$$P(\theta_1 \in B_1, \theta_2 \in B_2, \ldots, \theta_n \in B_n) = \int_0^1 \prod_{i=1}^{n} Q(B_i)\mu(dQ)$$

for some probability measure μ on the set of probability measures on Ω.

Further, μ is the distribution of the limiting empirical measure $M(B) = n^{-1}\#\{i \leq n : \theta_i \in B\}$.

The Dirichlet process

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

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

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,

$$(G(B_1), \ldots, G(B_m)) \sim \text{Dirichlet}(\alpha G_0(B_1), \ldots, \alpha G_0(B_m))$$

The Dirichlet process, prior to posterior

If

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

and, given G, θ_i are i.i.d. from G, then

$$G|\theta_1, \theta_2, \ldots, \theta_n \sim DP(\alpha + n, \frac{\alpha}{\alpha + n} G_0 + \frac{n}{\alpha + n} F_n)$$

where F_n is the empirical distribution $n^{-1} \sum_i \delta_{\theta_i}$. ($\delta_\theta$ is the point mass distribution putting probability 1 on θ.)

So, yes, we still have conjugacy!

The Dirichlet process - view 0

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

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

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,

$$(G(B_1), \ldots, G(B_m)) \sim \text{Dirichlet}(\alpha G_0(B_1), \ldots, \alpha G_0(B_m))$$

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

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

The Dirichlet process - view 0

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

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

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,

$$(G(B_1), \ldots, G(B_m)) \sim \text{Dirichlet}(\alpha G_0(B_1), \ldots, \alpha G_0(B_m))$$

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

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

Ties – and why?

Even if G_0 is continuous, G is a.s. discrete, so i.i.d. draws $\{\theta_i, i = 1, 2, \ldots, n\}$ from G exhibit ties.

Why? Consider two draws θ_1, θ_2 from $G \sim DP(\alpha, G_0)$.

$$P(\theta_1, \theta_2 \in B|G) = G(B)^2.$$

But $G(B) \sim \text{Beta}(\alpha G_0(B), \alpha(1 - G_0(B))$, so so

$$P(\theta_1, \theta_2 \in B) = E[G(B)^2] = G_0(B)^2 + \frac{G_0(B)(1 - G_0(B))}{1 + \alpha}$$

$$= G_0(B) \times \frac{1 + \alpha G_0(B)}{1 + \alpha}$$

i.e. $P(\theta_2 \in B|\theta_1 \in B) = (1 + \alpha G_0(B))/(1 + \alpha) \to 1/(1 + \alpha)$ as $G_0(B) \to 0$. So if G_0 is continuous, $P(\theta_2 = \theta_1|\theta_1) = 1/(1 + \alpha)\ \forall \theta_1$.

Ties – and why?

Even if G_0 is continuous, G is a.s. discrete, so i.i.d. draws $\{\theta_i, i = 1, 2, \ldots, n\}$ from G exhibit ties.

Why? Consider two draws θ_1, θ_2 from $G \sim DP(\alpha, G_0)$.

$$P(\theta_1, \theta_2 \in B|G) = G(B)^2.$$

But $G(B) \sim \text{Beta}(\alpha G_0(B), \alpha(1 - G_0(B))$, so so

$$P(\theta_1, \theta_2 \in B) = E[G(B)^2] = G_0(B)^2 + \frac{G_0(B)(1 - G_0(B))}{1 + \alpha}$$

$$= G_0(B) \times \frac{1 + \alpha G_0(B)}{1 + \alpha}$$

i.e. $P(\theta_2 \in B|\theta_1 \in B) = (1 + \alpha G_0(B))/(1 + \alpha) \to 1/(1 + \alpha)$ as $G_0(B) \to 0$. So if G_0 is continuous, $P(\theta_2 = \theta_1|\theta_1) = 1/(1 + \alpha)\ \forall \theta_1$.

Ties – and why?

Even if G_0 is continuous, G is a.s. discrete, so i.i.d. draws $\{\theta_i, i = 1, 2, \ldots, n\}$ from G exhibit ties.

Why? Consider two draws θ_1, θ_2 from $G \sim DP(\alpha, G_0)$.

$$P(\theta_1, \theta_2 \in B|G) = G(B)^2.$$

But $G(B) \sim \text{Beta}(\alpha G_0(B), \alpha(1 - G_0(B))$, so so

$$P(\theta_1, \theta_2 \in B) = E[G(B)^2] = G_0(B)^2 + \frac{G_0(B)(1 - G_0(B))}{1 + \alpha}$$
$$= G_0(B) \times \frac{1 + \alpha G_0(B)}{1 + \alpha}$$

i.e. $P(\theta_2 \in B|\theta_1 \in B) = (1 + \alpha G_0(B))/(1 + \alpha) \to 1/(1 + \alpha)$ as $G_0(B) \to 0$. So if G_0 is continuous, $P(\theta_2 = \theta_1|\theta_1) = 1/(1 + \alpha)\ \forall \theta_1$.

The Dirichlet process - view 0 (ctd.)

The parameter α measures (inverse) *concentration*: given i.i.d. draws $\{\theta_i, i = 1, 2, \ldots, n\}$ from G,

- As $\alpha \to 0$, all θ_i are equal, drawn from G_0.
- As $\alpha \to \infty$, the θ_i are drawn i.i.d. from G_0.

Pólya urn representation

You can generate the $\{\theta_i\}$ i.i.d. from $G \sim \mathrm{DP}(\alpha, G_0)$ without explicitly creating G by

- Draw $\theta_1 \sim G_0$
- For $i = 1, 2, \ldots, n-1$, draw

$$\theta_{i+1} \sim \frac{\alpha}{\alpha + i} G_0 + \frac{1}{\alpha + i} \sum_{j=1}^{i} \delta_{\theta_j}$$

It is not so obvious now that the θ_i are exchangeable!

Example: ordinary urn: $\Omega = \{\texttt{Red}, \texttt{Blue}\}$, $\alpha =$ initial number of balls, $G_0 =$ initial proportions of each colour.

Pólya urn representation

You can generate the $\{\theta_i\}$ i.i.d. from $G \sim \mathrm{DP}(\alpha, G_0)$ without explicitly creating G by

- Draw $\theta_1 \sim G_0$
- For $i = 1, 2, \ldots, n-1$, draw

$$\theta_{i+1} \sim \frac{\alpha}{\alpha + i} G_0 + \frac{1}{\alpha + i} \sum_{j=1}^{i} \delta_{\theta_j}$$

It is not so obvious now that the θ_i are exchangeable!

Example: ordinary urn: $\Omega = \{\texttt{Red}, \texttt{Blue}\}$, $\alpha =$ initial number of balls, $G_0 =$ initial proportions of each colour.

Chinese restaurant process

A fun metaphor:

Customers enter a restaurant one-by-one.

- The first customer sits at table 1
- Subsequently, the $(i+1)$th customer:
  - joins an existing table with c previous customers, with probability $c/(\alpha+i)$
  - sits alone at the next free table, with probability $\alpha/(\alpha+i)$

This gives a partition of customers into tables that is exchangeable in the labels of both customers and tables – and has the same law as that describing ties in draws from a Dirichlet process.

The distribution of the clustering of customers is independent of their order of entering the restaurant!

Chinese restaurant process

A fun metaphor:

Customers enter a restaurant one-by-one.

- The first customer sits at table 1
- Subsequently, the $(i+1)$th customer:
  - joins an existing table with c previous customers, with probability $c/(\alpha+i)$
  - sits alone at the next free table, with probability $\alpha/(\alpha+i)$

This gives a partition of customers into tables that is exchangeable in the labels of both customers and tables – and has the same law as that describing ties in draws from a Dirichlet process.

The distribution of the clustering of customers is independent of their order of entering the restaurant!

Chinese restaurant process

A fun metaphor:

Customers enter a restaurant one-by-one.

- The first customer sits at table 1
- Subsequently, the $(i+1)$th customer:
  - joins an existing table with c previous customers, with probability $c/(\alpha+i)$
  - sits alone at the next free table, with probability $\alpha/(\alpha+i)$

This gives a partition of customers into tables that is exchangeable in the labels of both customers and tables – and has the same law as that describing ties in draws from a Dirichlet process.

The distribution of the clustering of customers is independent of their order of entering the restaurant!

Gibbs sampling application

(Unnecessary here, but useful motivation for later)

Since the $\{\theta_i\}$ are exchangeable, it is also true that the full conditionals are

$$\theta_i|\theta_{-i} \sim \frac{\alpha}{\alpha+n-1}G_0 + \frac{1}{\alpha+n-1}\sum_{j\neq i}\delta_{\theta_j}$$

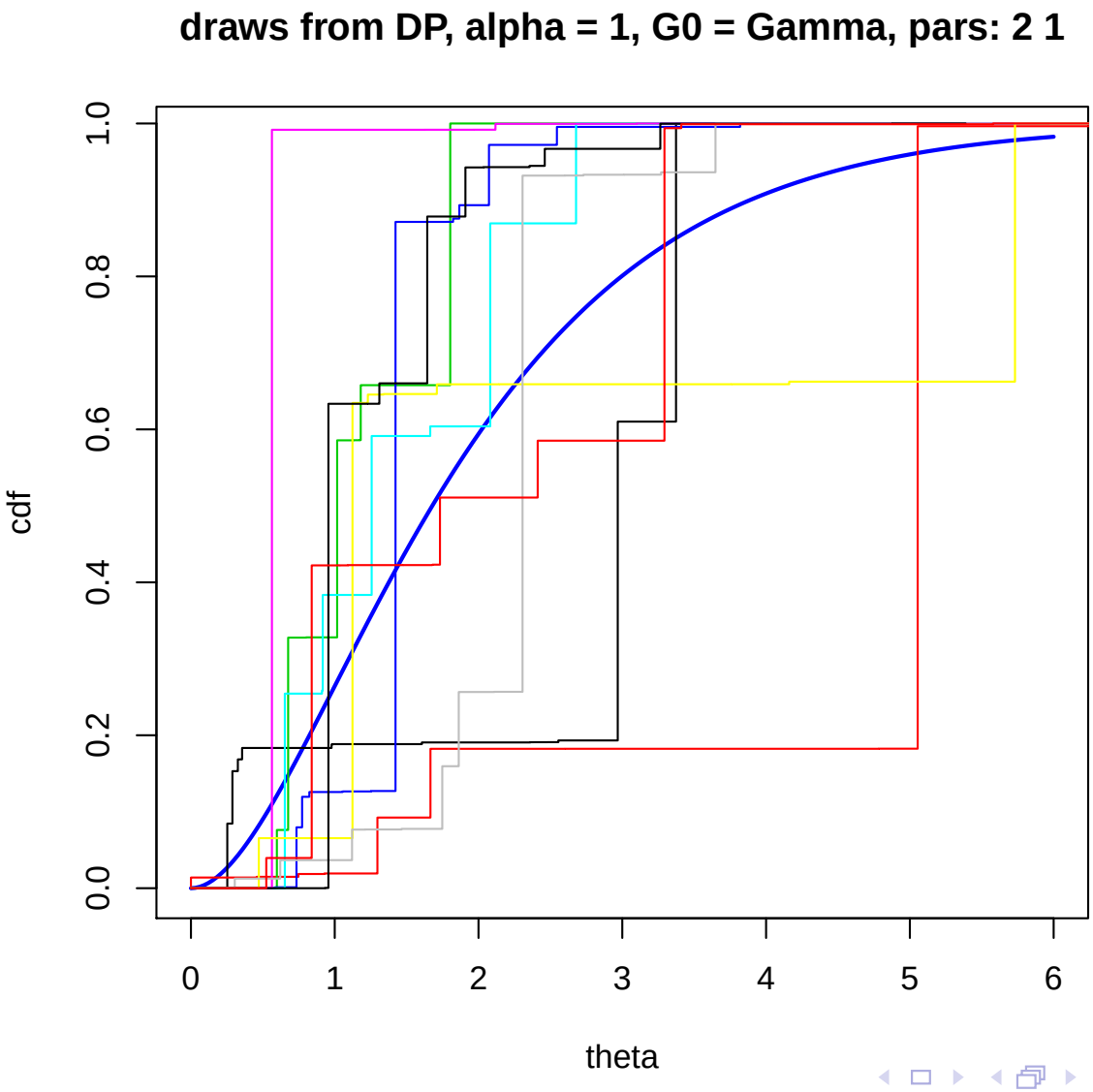
for all $i = 1, 2, \ldots, n$, where $\theta_{-i} = \{\theta_j, j \neq i\}$, which could be used to generate the $\{\theta_i\}$ by Gibbs sampling.

The Dirichlet process - view 1

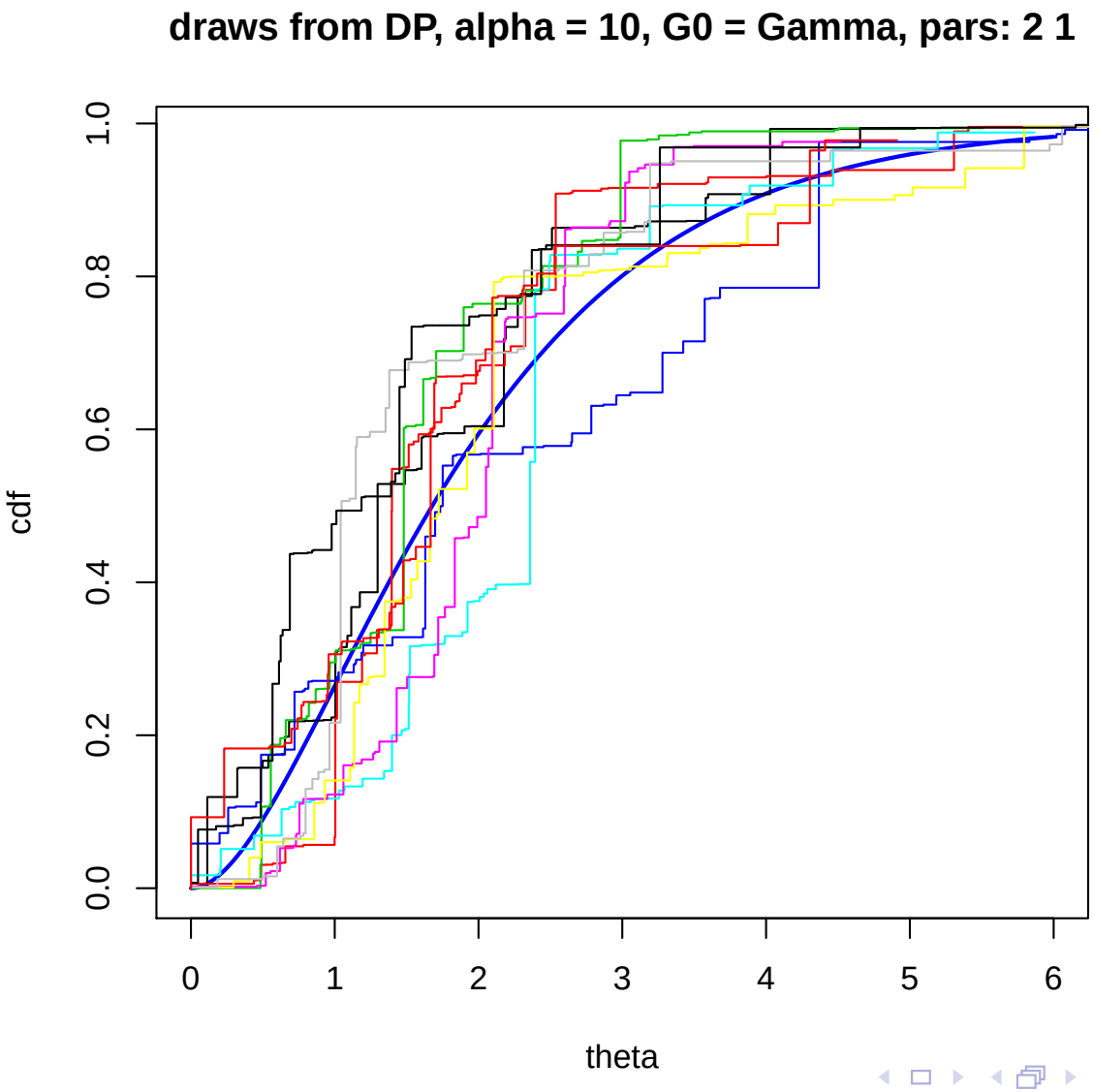
Sethuraman's 'stick-breaking' construction of G:

- draw $\theta_j^\star \sim G_0$, i.i.d., $j = 1, 2, \ldots$
- draw $V_j \sim \text{Beta}(1, \alpha)$, i.i.d., $j = 1, 2, \ldots$
- define G to be the discrete distribution putting probability $(1 - V_1)(1 - V_2) \ldots (1 - V_{j-1}) V_j$ on $\theta_j^\star$

The Dirichlet process



The Dirichlet process



Some more properties of the Dirichlet process

- lack of smoothness, e.g.
  $E(G(B)|(\theta_i)_{i=1}^n) = \alpha/(\alpha+n)G_0(B) < E(G(B))$ if no $\theta_i \in B$,
  however close they come
- $G(B_1)$ and $G(B_2)$ negatively correlated for any disjoint B_1, B_2
- $(G(B), G(B^c))$, $G_{|B}$ and $G_{|B^c}$ are independent, where
  $G_{|B}(A) = G(A|B)$
- $G_{|B} \sim \mathrm{DP}(\alpha G_0(B), G_{0|B})$
- expected number of distinct θ_i is $\sim \alpha \log(n/\alpha)$

Density estimation/infinite mixtures

The Dirichlet process model as it stands, generating randomly tied random variables $\{\theta_i\}$, is not a very useful model for data, but it is a useful ingredient in a hierarchical model setup. The simplest example is in mixture modelling.

The Dirichlet process mixture model

- $G \sim \mathrm{DP}(\alpha, G_0)$
- $\theta_i | G \sim G$, $i = 1, 2, \ldots, n$, i.i.d.
- $y_i | G, \theta \sim f(\cdot; \theta_i)$, independently

A kind of infinite mixture model, the number of components (distinct θ_i) is not bounded.

Density estimation/infinite mixtures

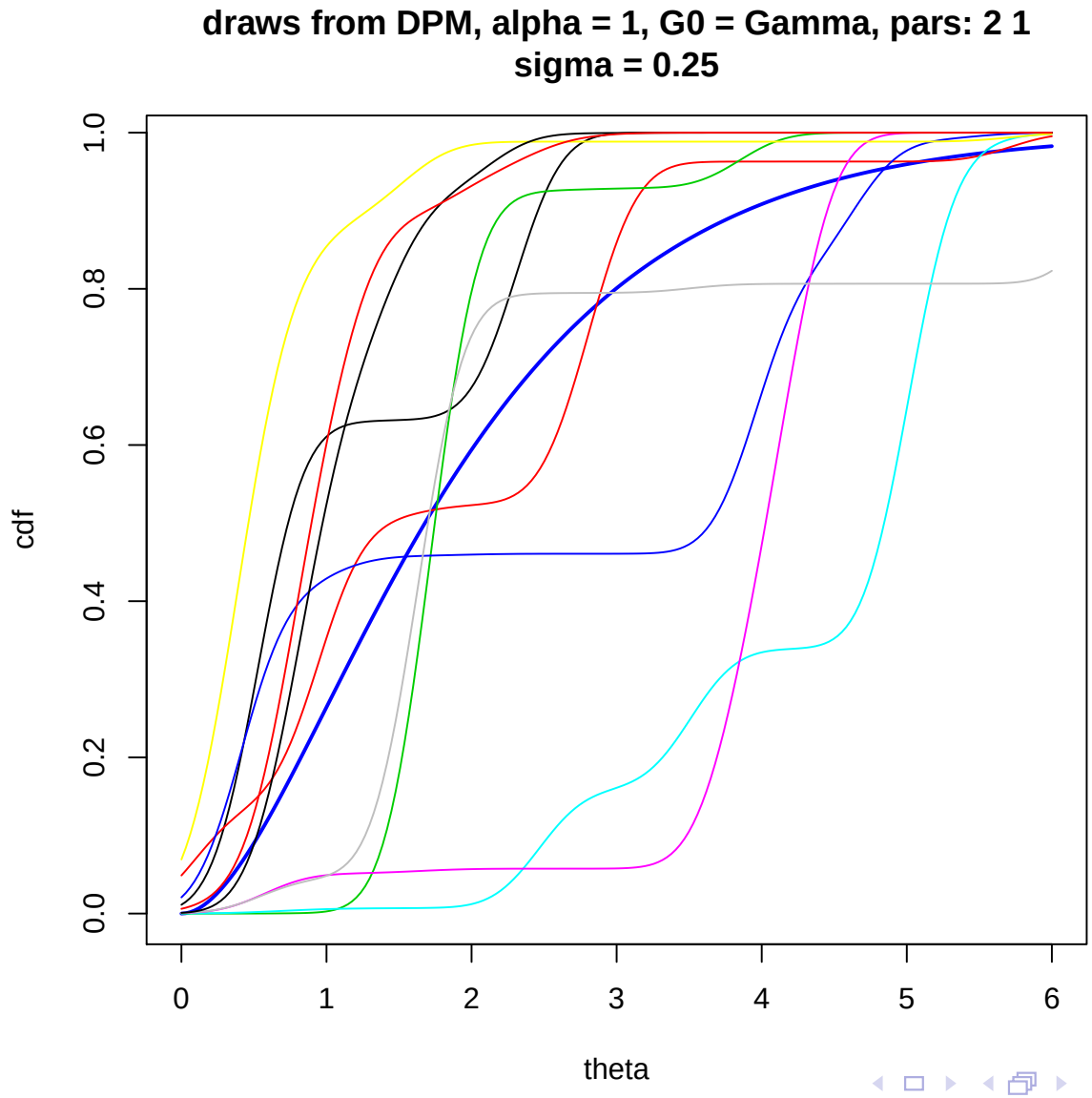
The Dirichlet process model as it stands, generating randomly tied random variables $\{\theta_i\}$, is not a very useful model for data, but it is a useful ingredient in a hierarchical model setup. The simplest example is in mixture modelling.

The Dirichlet process mixture model

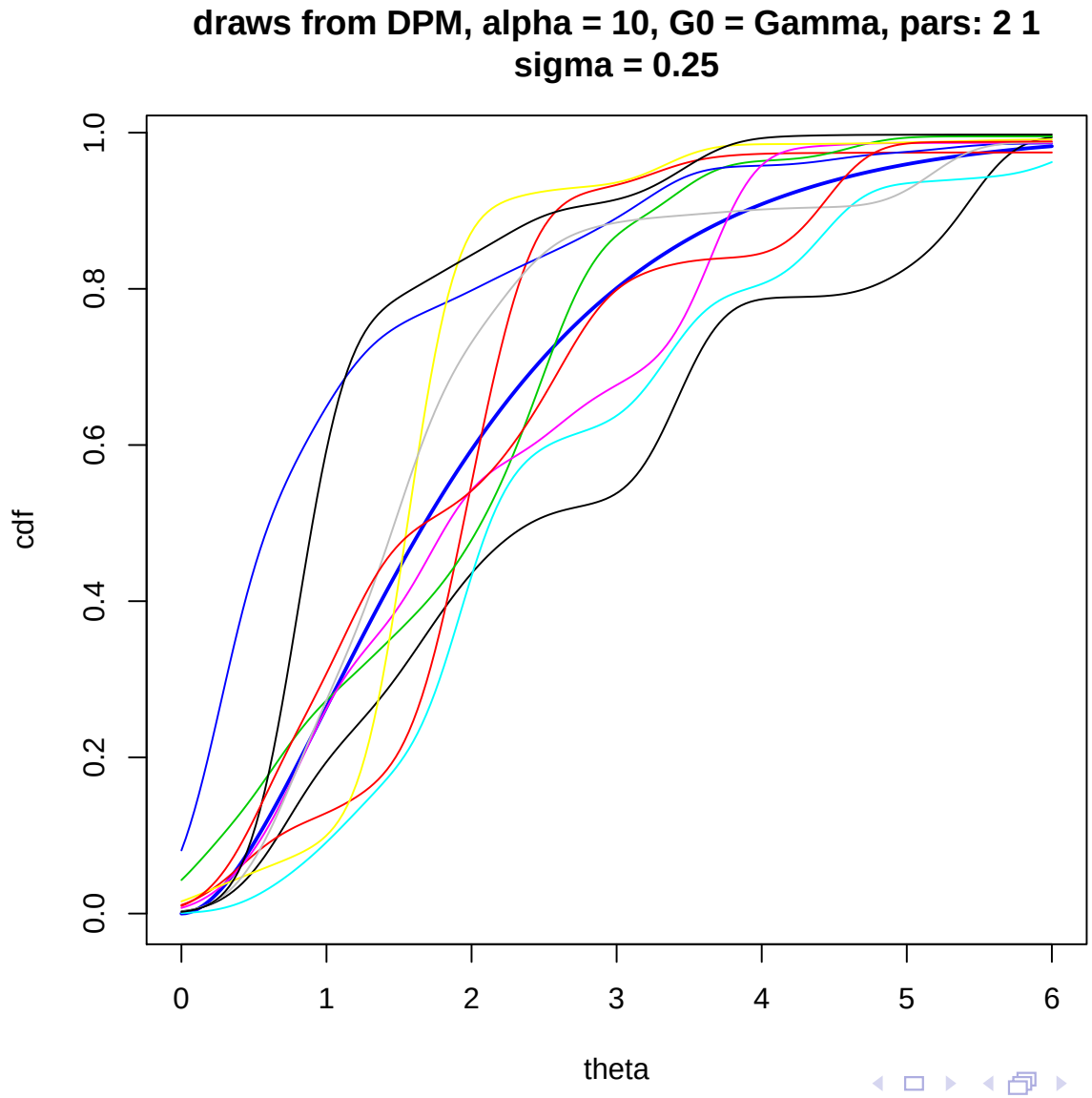
- $G \sim \mathrm{DP}(\alpha, G_0)$
- $\theta_i | G \sim G$, $i = 1, 2, \ldots, n$, i.i.d.
- $y_i | G, \theta \sim f(\cdot; \theta_i)$, independently

A kind of infinite mixture model, the number of components (distinct θ_i) is not bounded.

A Dirichlet process mixture



A Dirichlet process mixture



Pólya urn for a Dirichlet process mixture

The $(i+1)$th observation is allocated to a new cluster with probability

$$\propto \frac{\alpha}{\alpha+i}\int f(y_{i+1};\theta)G_0(d\theta)$$

and to existing cluster C_r^i with probability

$$\propto \frac{|C_r^i|}{\alpha+i}\frac{\int f(y_{i+1};\theta)\prod_{j\in C_r^i} f(y_j;\theta)G_0(d\theta)}{\int \prod_{j\in C_r^i} f(y_j;\theta)G_0(d\theta)}$$

– the so-called Weighted Chinese restaurant process. This is particularly useful when G_0 is conjugate for $\prod_i f(y_i;\theta)$, then all these integrals have explicit closed form.

More general models based on the DP

It is a small step to replace the 3rd stage here:

- $G \sim \mathrm{DP}(\alpha, G_0)$
- $\theta_i | G \sim G$, $i = 1, 2, \ldots, n$, i.i.d.
- $y_i | G, \theta \sim f(\cdot; \theta_i)$, independently

with

- $y_i | G, \theta \sim f_i(\cdot; x_i, \theta_i)$, independently

In fact, all we need is that the likelihood of the data as a function of θ, given G, is a product of functions of the individual θ_i. This allows huge generality.

Only the joint distribution of the $\{\theta_i\}$ contributes to the likelihood, not G itself, so assuming G is not specifically an object of inference either, we can formulate the model for $\{\theta_i\}$ more directly, without losing anything.

More general models based on the DP

It is a small step to replace the 3rd stage here:

- $G \sim \mathrm{DP}(\alpha, G_0)$
- $\theta_i | G \sim G$, $i = 1, 2, \ldots, n$, i.i.d.
- $y_i | G, \theta \sim f(\cdot; \theta_i)$, independently

with

- $y_i | G, \theta \sim f_i(\cdot; x_i, \theta_i)$, independently

In fact, all we need is that the likelihood of the data as a function of θ, given G, is a product of functions of the individual θ_i. This allows huge generality.

Only the joint distribution of the $\{\theta_i\}$ contributes to the likelihood, not G itself, so assuming G is not specifically an object of inference either, we can formulate the model for $\{\theta_i\}$ more directly, without losing anything.

More general models based on the DP

It is a small step to replace the 3rd stage here:

- $G \sim \mathrm{DP}(\alpha, G_0)$
- $\theta_i | G \sim G$, $i = 1, 2, \ldots, n$, i.i.d.
- $y_i | G, \theta \sim f(\cdot; \theta_i)$, independently

with

- $y_i | G, \theta \sim f_i(\cdot; x_i, \theta_i)$, independently

In fact, all we need is that the likelihood of the data as a function of θ, given G, is a product of functions of the individual θ_i. This allows huge generality.

Only the joint distribution of the $\{\theta_i\}$ contributes to the likelihood, not G itself, so assuming G is not specifically an object of inference either, we can formulate the model for $\{\theta_i\}$ more directly, without losing anything.

The Dirichlet process - view 2

Partition model: partition $\{1, 2, \ldots, n\} = \bigcup_{j=1}^{d} C_j$ at random, so that

$$p(C_1, C_2, \ldots, C_d) = \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)} \alpha^d \prod_{j=1}^{d} (n_j - 1)!$$

where $n_j = \#C_j$. (NB preference for unequal cluster sizes!) Draw $\theta_j^\star \sim G_0$, i.i.d., $j = 1, \ldots, d$, and set $\theta_i = \theta_j^\star$ if $i \in C_j$.

G is invisible in view 2.

The Dirichlet process - view 3

Finite mixture model $\sum_j w_j f(\cdot|\theta_j^\star)$ with a Dirichlet prior on the weights:

- Draw $(w_1, w_2, \ldots, w_k) \sim \text{Dirichlet}(\delta, \ldots, \delta)$
- Draw $c_i \in \{1, 2, \ldots, k\}$ with $P\{c_i = j\} = w_j$, i.i.d., $i = 1, \ldots, n$
- Draw $\theta_j^\star \sim G_0$, i.i.d., $j = 1, \ldots, k$
- Set $\theta_i = \theta_{c_i}^\star$

Let $k \to \infty$, $\delta \to 0$ such that $k\delta \to \alpha$.

The result is a Dirichlet process mixture model, if you ignore the 'empty components'. G is also invisible in view 3.

Dirichlet process mixtures – reprise

Items are clustered, according to a specific tractable distribution parameterised by $\alpha > 0$, and within each cluster the parameter θ is drawn i.i.d. from G_0.

How nonparametric is that?

Dirichlet process – summary

- The only direct generalisation of the approach that is most natural in the case of discrete θ
- Retains interpretability and computational advantages
- Some unexpected and/or undesirable properties
- Posterior consistency
- Many different perspectives possible, suggesting generalisations in different directions
  - random measures
  - stick-breaking
  - Pólya urn representation/algorithm
  - partitions

Children of the Dirichlet process

- Lessons learned from the DP, and different directions for generalisation
- Conjugacy, neutrality and survival analysis
- Pólya trees
- Species sampling models
- (Doubly) nonparametric regression: dependent DP and kernel stick-breaking priors
- Hybrid DP for high-dimensional and functional parameters
- Hierarchical DP
- Exchangeable partition functions
- Bernstein polynomials

Multiple notations for partitions

- $\boldsymbol{c}$ 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 $i \in C_j$

!Take care with multiplicities, and distinction between allocations and partitions: labelling of C_j is arbitrary, likewise values of $\{c_i\}$.

Pólya urn for a Dirichlet process mixture

Suppose G_0 is conjugate for $\prod_i f(y_i;\theta)$. Write $m(y_A)$ for marginal likelihood.

The $(i+1)$th observation is allocated to a new cluster with probability

$$\propto \frac{\alpha}{\alpha+i}\int f(y_{i+1};\theta)G_0(d\theta) = \frac{\alpha}{\alpha+i}m(y_{i+1})$$

and to existing cluster C_r^i with probability

$$\propto \frac{|C_r^i|}{\alpha+i}\frac{\int f(y_{i+1};\theta)\prod_{j\in C_r^i} f(y_j;\theta)G_0(d\theta)}{\int \prod_{j\in C_r^i} f(y_j;\theta)G_0(d\theta)} = \frac{|C_r^i|}{\alpha+i}\frac{m(y_{C_r^i\cup\{i+1\}})}{m(y_{C_r^i})}$$

– the so-called Weighted Chinese restaurant (WCR) process.

Gibbs sampling of allocation indices

Because of exchangeability, the WCR probabilities can be used for Gibbs sampling of the allocation indices c_i.

This was key to the early adoption of DP models in applied Bayesian nonparametric methods. But this Gibbs sampler is easily generalised to many other models.

The first factor ($\alpha/(\alpha+i)$ or $|C_r^i|/(\alpha+i)$) is simply the ratio of the prior probabilities of the appropriate partition with and without allocation of the new item.

Examples of models for which the partition distribution is explictly available, and factorises so that the ratios needed are simple include the Dirichlet-multinomial finite mixture model, and the Pitman–Yor two-parameter Poisson–Dirichlet process, as well as the DP mixture model – and the idea is not limited to (re-)allocating one item at a time.

Gibbs sampling of allocation indices

Because of exchangeability, the WCR probabilities can be used for Gibbs sampling of the allocation indices c_i.

This was key to the early adoption of DP models in applied Bayesian nonparametric methods. But this Gibbs sampler is easily generalised to many other models.

The first factor ($\alpha/(\alpha+i)$ or $|C_r^i|/(\alpha+i)$) is simply the ratio of the prior probabilities of the appropriate partition with and without allocation of the new item.

Examples of models for which the partition distribution is explictly available, and factorises so that the ratios needed are simple include the Dirichlet-multinomial finite mixture model, and the Pitman–Yor two-parameter Poisson–Dirichlet process, as well as the DP mixture model – and the idea is not limited to (re-)allocating one item at a time.

Gibbs sampling of allocation indices

Because of exchangeability, the WCR probabilities can be used for Gibbs sampling of the allocation indices c_i.

This was key to the early adoption of DP models in applied Bayesian nonparametric methods. But this Gibbs sampler is easily generalised to many other models.

The first factor ($\alpha/(\alpha+i)$ or $|C_r^i|/(\alpha+i)$) is simply the ratio of the prior probabilities of the appropriate partition with and without allocation of the new item.

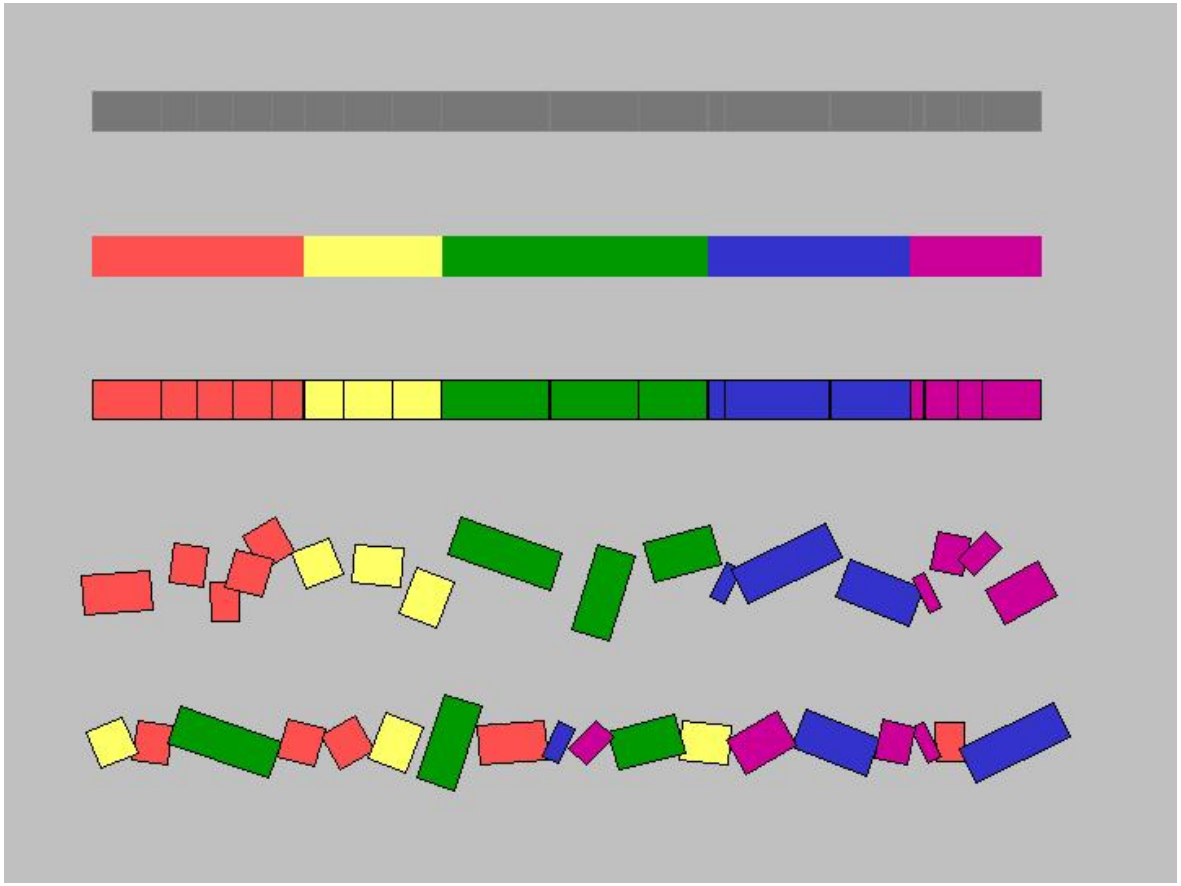
Examples of models for which the partition distribution is explictly available, and factorises so that the ratios needed are simple include the Dirichlet-multinomial finite mixture model, and the Pitman–Yor two-parameter Poisson–Dirichlet process, as well as the DP mixture model – and the idea is not limited to (re-)allocating one item at a time.

A coloured Dirichlet process

To define a variant on the DP in which not all clusters are exchangeable:

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

Colouring and breaking sticks



Coloured partition distribution

The CDP generates the following partition model: partition $\{1, 2, \ldots, n\} = \bigcup_k \bigcup_{j=1}^{d_k} C_{kj}$ at random, so that

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

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

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

Note that the clustering remains exchangeable over items. For $i \in C_{kj}$, set $k_i = k$ and $\theta_i = \theta_j^\star$, where $\theta_j^\star$ are drawn i.i.d. from G_{0k}.

Coloured partition distribution

The CDP generates the following partition model: partition $\{1, 2, \ldots, n\} = \bigcup_k \bigcup_{j=1}^{d_k} C_{kj}$ at random, so that

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

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

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

Note that the clustering remains exchangeable over items. For $i \in C_{kj}$, set $k_i = k$ and $\theta_i = \theta_j^\star$, where $\theta_j^\star$ are drawn i.i.d. from G_{0k}.

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
  $\propto \alpha_k \times (\gamma_k + n_k^{-i})/(\alpha_k + n_k^{-i}) \times m(Y_i)$
- the existing cluster C_{kj}, with probability
  $\propto n_{kj}^{-i} \times (\gamma_k + n_k^{-i})/(\alpha_k + n_k^{-i}) \times m(Y_i | Y_{C_{kj}^{-i}})$

Distribution estimation allowing continuity

The Pólya tree is a construction that generalises the Dirichlet process in a way that allows the modeller to impose continuity.

We start an arbitrary infinite binary tree partition of Ω: $\Omega = B_0 \cup B_1$, with $B_0 \cap B_1 = \emptyset$, and continue: $B_\epsilon = B_{\epsilon 0} \cup B_{\epsilon 1}$ with $B_{\epsilon 0} \cap B_{\epsilon 1} = \emptyset$, for any finite sequence ϵ of 0's and 1's. To 'centre' at G_0, we typically choose B_ϵ so that $G_0(B_\epsilon) = 2^{-m}$ if ϵ has length m.

We also have a tree of non-negative parameters $\mathcal{A} = \{\alpha_\epsilon\}$ and a tree of independent random variables $\mathcal{C} = \{C_\epsilon\}$ with $C_{\epsilon 0} \sim \text{Beta}(\alpha_{\epsilon 0}, \alpha_{\epsilon 1})$ and $C_{\epsilon 0} = 1 - C_{\epsilon 1}$.

The we can define a random distribution G by requiring $G(B_{\epsilon 0}) = G(B_\epsilon)C_{\epsilon 0}$ and $G(B_{\epsilon 1}) = G(B_\epsilon)C_{\epsilon 1}$.

Distribution estimation allowing continuity

The Pólya tree is a construction that generalises the Dirichlet process in a way that allows the modeller to impose continuity.

We start an arbitrary infinite binary tree partition of Ω: $\Omega = B_0 \cup B_1$, with $B_0 \cap B_1 = \emptyset$, and continue: $B_\epsilon = B_{\epsilon 0} \cup B_{\epsilon 1}$ with $B_{\epsilon 0} \cap B_{\epsilon 1} = \emptyset$, for any finite sequence ϵ of 0's and 1's. To 'centre' at G_0, we typically choose B_ϵ so that $G_0(B_\epsilon) = 2^{-m}$ if ϵ has length m.

We also have a tree of non-negative parameters $\mathcal{A} = \{\alpha_\epsilon\}$ and a tree of independent random variables $\mathcal{C} = \{C_\epsilon\}$ with $C_{\epsilon 0} \sim \text{Beta}(\alpha_{\epsilon 0}, \alpha_{\epsilon 1})$ and $C_{\epsilon 0} = 1 - C_{\epsilon 1}$.

The we can define a random distribution G by requiring $G(B_{\epsilon 0}) = G(B_\epsilon)C_{\epsilon 0}$ and $G(B_{\epsilon 1}) = G(B_\epsilon)C_{\epsilon 1}$.

Distribution estimation allowing continuity

The Pólya tree is a construction that generalises the Dirichlet process in a way that allows the modeller to impose continuity.

We start an arbitrary infinite binary tree partition of Ω: $\Omega = B_0 \cup B_1$, with $B_0 \cap B_1 = \emptyset$, and continue: $B_\epsilon = B_{\epsilon 0} \cup B_{\epsilon 1}$ with $B_{\epsilon 0} \cap B_{\epsilon 1} = \emptyset$, for any finite sequence ϵ of 0's and 1's. To 'centre' at G_0, we typically choose B_ϵ so that $G_0(B_\epsilon) = 2^{-m}$ if ϵ has length m.

We also have a tree of non-negative parameters $\mathcal{A} = \{\alpha_\epsilon\}$ and a tree of independent random variables $\mathcal{C} = \{C_\epsilon\}$ with $C_{\epsilon 0} \sim \mathrm{Beta}(\alpha_{\epsilon 0}, \alpha_{\epsilon 1})$ and $C_{\epsilon 0} = 1 - C_{\epsilon 1}$.

The we can define a random distribution G by requiring $G(B_{\epsilon 0}) = G(B_\epsilon) C_{\epsilon 0}$ and $G(B_{\epsilon 1}) = G(B_\epsilon) C_{\epsilon 1}$.

Distribution estimation allowing continuity

The Pólya tree is a construction that generalises the Dirichlet process in a way that allows the modeller to impose continuity.

We start an arbitrary infinite binary tree partition of Ω: $\Omega = B_0 \cup B_1$, with $B_0 \cap B_1 = \emptyset$, and continue: $B_\epsilon = B_{\epsilon 0} \cup B_{\epsilon 1}$ with $B_{\epsilon 0} \cap B_{\epsilon 1} = \emptyset$, for any finite sequence ϵ of 0's and 1's. To 'centre' at G_0, we typically choose B_ϵ so that $G_0(B_\epsilon) = 2^{-m}$ if ϵ has length m.

We also have a tree of non-negative parameters $\mathcal{A} = \{\alpha_\epsilon\}$ and a tree of independent random variables $\mathcal{C} = \{C_\epsilon\}$ with $C_{\epsilon 0} \sim \text{Beta}(\alpha_{\epsilon 0}, \alpha_{\epsilon 1})$ and $C_{\epsilon 0} = 1 - C_{\epsilon 1}$.

The we can define a random distribution G by requiring $G(B_{\epsilon 0}) = G(B_\epsilon)C_{\epsilon 0}$ and $G(B_{\epsilon 1}) = G(B_\epsilon)C_{\epsilon 1}$.

The Pólya tree

If we set $\alpha_\epsilon = c/2^m$ where m is the length of the sequence ϵ, we recover the Dirichlet process.

If we set $\alpha_\epsilon = cm^2$, G is absolutely continuous w.p. 1.

Prior to posterior

If G follows a Pólya process a priori, and $\theta \sim G$ then $G|\theta$ also follows a Pólya process, but with α_ϵ replaced by $\alpha'_\epsilon = \alpha_\epsilon + I[\theta \in B_\epsilon]$.

The Pólya tree

If we set $\alpha_\epsilon = c/2^m$ where m is the length of the sequence ϵ, we recover the Dirichlet process.

If we set $\alpha_\epsilon = cm^2$, G is absolutely continuous w.p. 1.

Prior to posterior

If G follows a Pólya process a priori, and $\theta \sim G$ then $G|\theta$ also follows a Pólya process, but with α_ϵ replaced by $\alpha'_\epsilon = \alpha_\epsilon + I[\theta \in B_\epsilon]$.

The Pólya tree

If we set $\alpha_\epsilon = c/2^m$ where m is the length of the sequence ϵ, we recover the Dirichlet process.

If we set $\alpha_\epsilon = cm^2$, G is absolutely continuous w.p. 1.

Prior to posterior

If G follows a Pólya process a priori, and $\theta \sim G$ then $G|\theta$ also follows a Pólya process, but with α_ϵ replaced by $\alpha'_\epsilon = \alpha_\epsilon + I[\theta \in B_\epsilon]$.

EPPFs

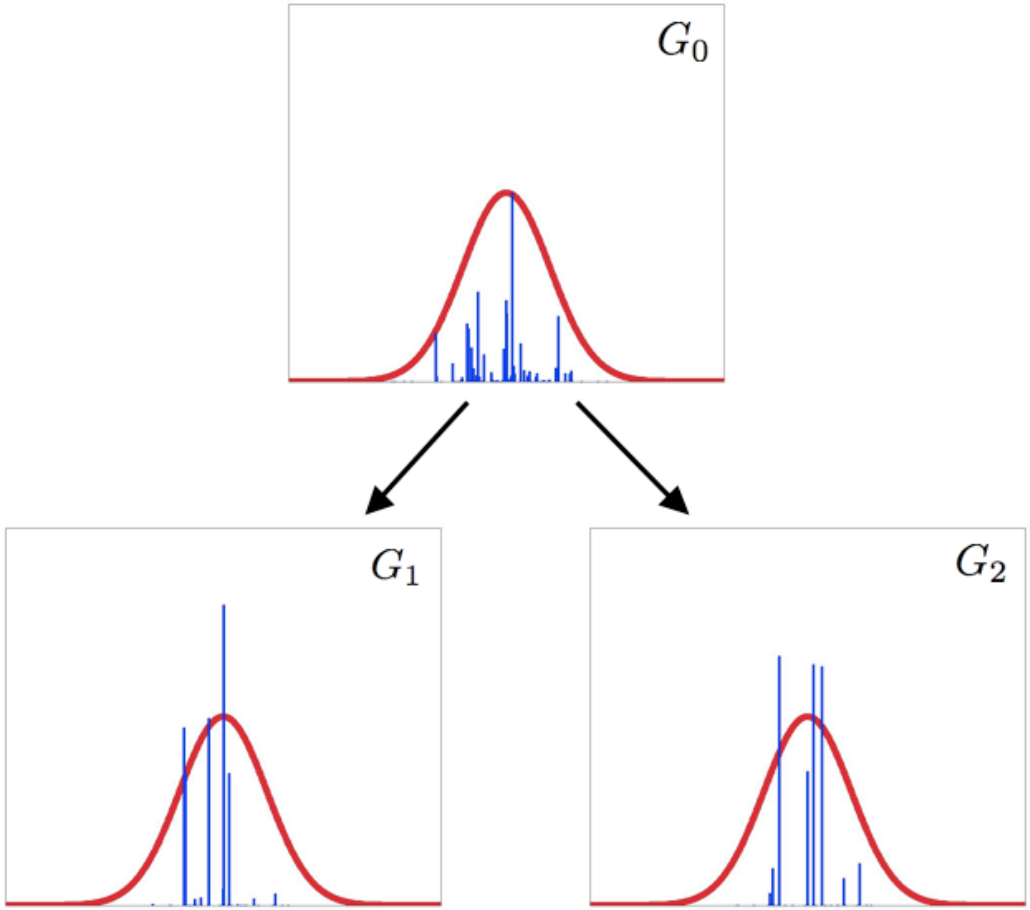
Pitman (1995) characterised random partition models that are

- exchangeable over items
- exchangeable over clusters
- 'heritable' – the marginal clustering of any subset of items follows a model of the same form

by means of exchangeable partition probability functions. These connect with normalised random measures with independent increments, and generate explicit predictive distributions (Pólya urn algorithms).

Hierarchical DP and overlapping clustering

Teh, Jordan, Beal and Blei (2006); use DP as base distribution G_0 to share clusters across multiple clustering problems.



Regression: DDP

'Bayesian nonparametric regression' ought to mean 'nonparametric' both in the distributional and in the regression sense: can we model an unknown distributional relationship $Y|x \sim f(\cdot|x)$ (ideally for moderately high dimensional x) nonparametrically, and fully probabilistically, in a way that is computationally practical?

The Dependent Dirichlet process (DDP) takes the stick-breaking representation $G = \sum_h w_h \delta_{\theta_h}$ and introduces dependence on covariates x into (potentially) both the weights w_h and the point locations θ_h. Dependence of w_h on x is computationally awkward, so typically

$$Y|x \sim G_x = \sum_h w_h \delta_{\theta_h(x)}$$

is assumed, where $\theta_h(x)$ are i.i.d. stochastic processes indexed by x (often in practice gaussian processes).

Spatial DDP

Gelfand/Kottas/MacEachern example based on French rainfall data

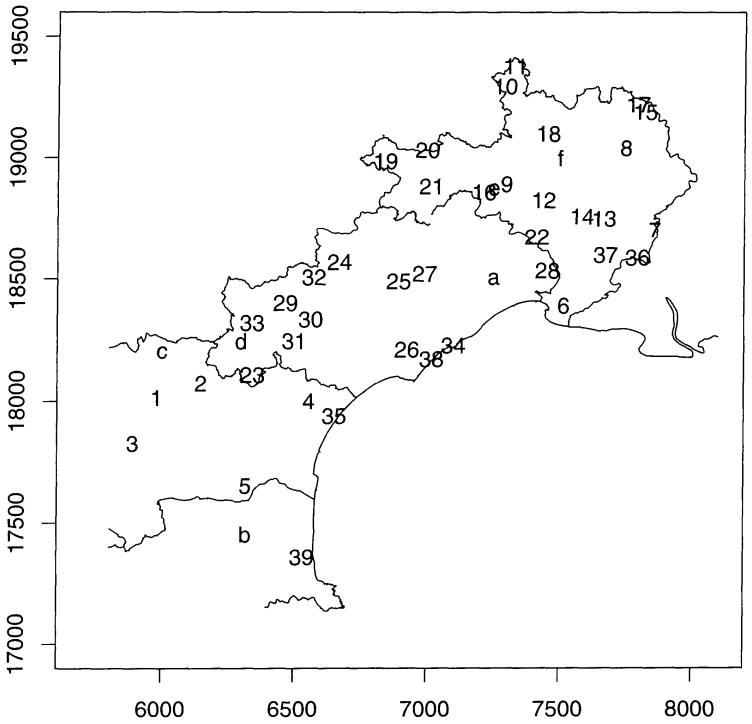


Figure 1. Geographic Map of the Languedoc-Roussillon Region in Southern France Showing the 39 Sites Where the Precipitation Data Have Been Observed. The six new sites considered for spatial prediction in the simulation experiment are denoted by a, b, c, d, e, and f. The boundaries of three French departments are also drawn.

Spatial DDP

Contrasting regular and spatial DP predictive inferences

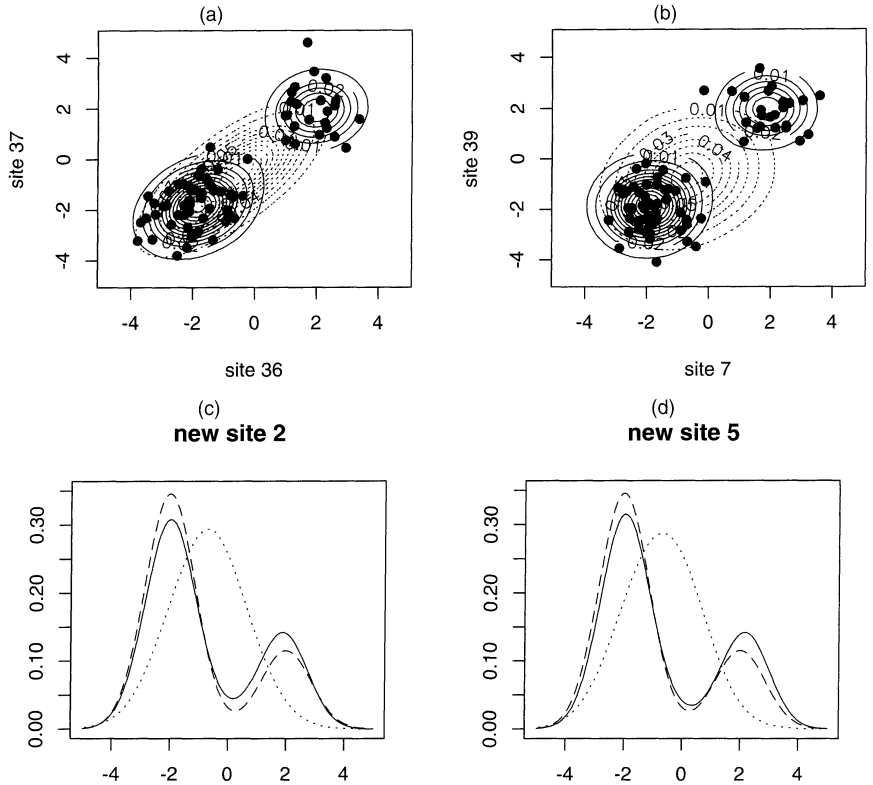


Figure 4. Simulated Data, Case II. (a) and (b) Bivariate posterior predictive densities for pairs of sites ($\boldsymbol{s}_{36}, \boldsymbol{s}_{37}$) and ($\boldsymbol{s}_7, \boldsymbol{s}_{39}$) overlaid on corresponding plots of data. (c) and (d) Posterior predictive densities at new sites $\tilde{\boldsymbol{s}}_2$ = (6,316, 17,452) and $\tilde{\boldsymbol{s}}_5$ = (7,250, 18,870) and the associated true densities (dashed lines). The solid lines correspond to the spatial DP mixture model; the dotted lines, to the parametric GP mixture model.

Regression: weighted mixture of DPs

Dunson, Pillai and Park (2007) take

$$G_x = \sum_{i=1}^{n} \left(\frac{\gamma_i K(x, x_i)}{\sum_{l=1}^{n} \gamma_l K(x, x_l)} \right) G_i^\star \quad \text{where} \quad G_i^\star \sim \mathrm{DP}(\alpha, G_0) \quad \text{i.i.d.}$$

Regression: weighted mixture of DPs

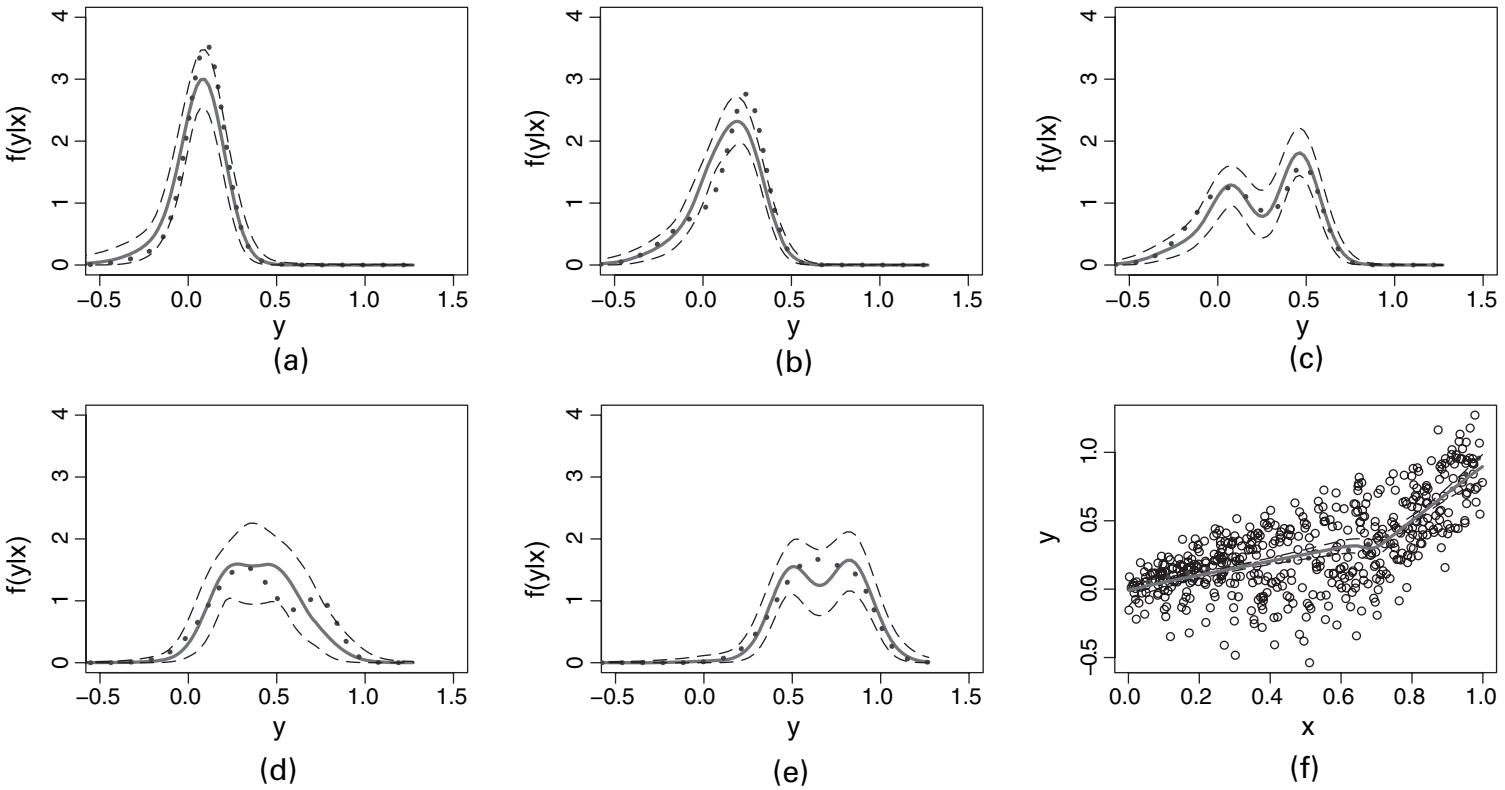


Fig. 2. True conditional densities of $y|x$ (·····), posterior mean estimates (———) and 99% pointwise credible intervals (-------) in simulation case 2: (a) $x = 0.1$ (10%); (b) $x = 0.25$ (25%); (c) $x = 0.49$ (50%); (d) $x = 0.75$ (75%); (e) $x = 0.88$ (90%); (f) data, along with the true and estimated mean regression curves

Regression: weighted mixture of DPs

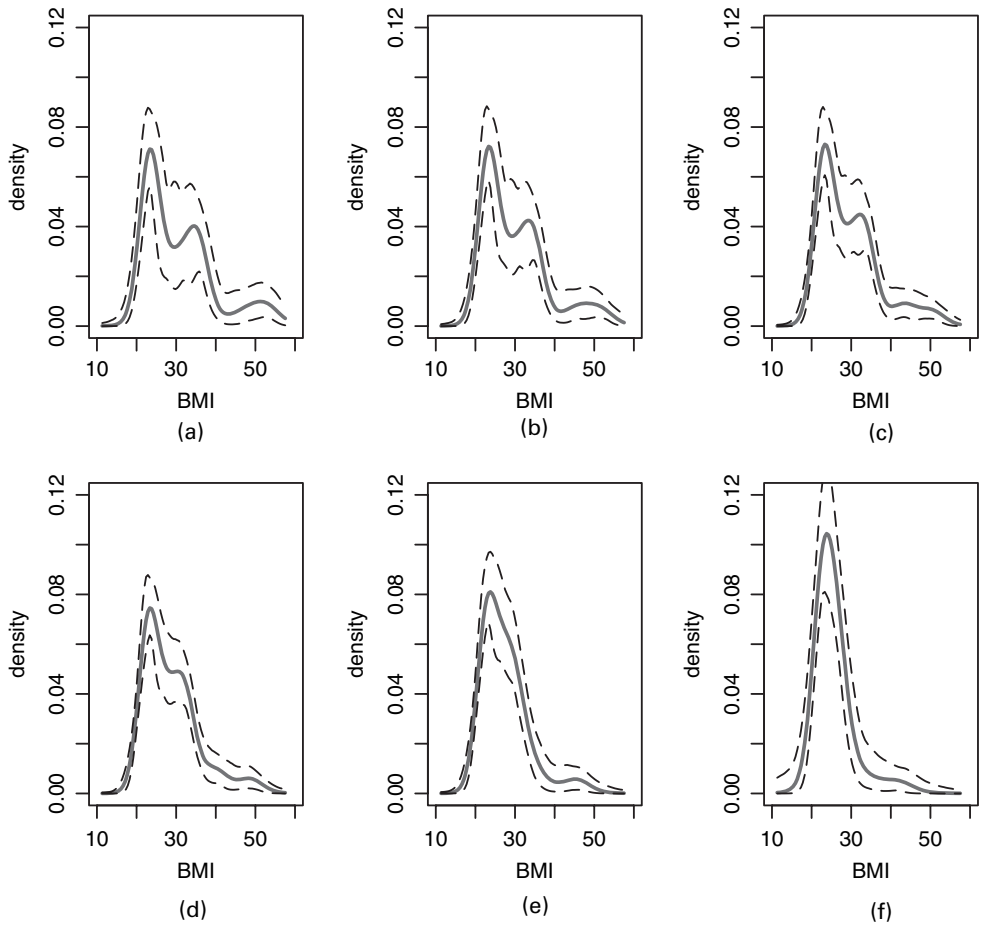


Fig. 3. Predictive densities for BMI conditional on a range of values for the level of LH, with age fixed at the sample mean (———, posterior predictive means; – – –, 99% pointwise credible intervals): (a) LH = 0.18 (1%); (b) LH = 1.34 (10%); (c) LH = 2.48 (25%); (d) LH = 3.98 (50%); (e) LH = 6.72 (75%); (f) LH = 10.14 (90%)

Functional data: the hybrid DP

Petrone, Guindani & Gelfand, *JRSSB*, 2009.

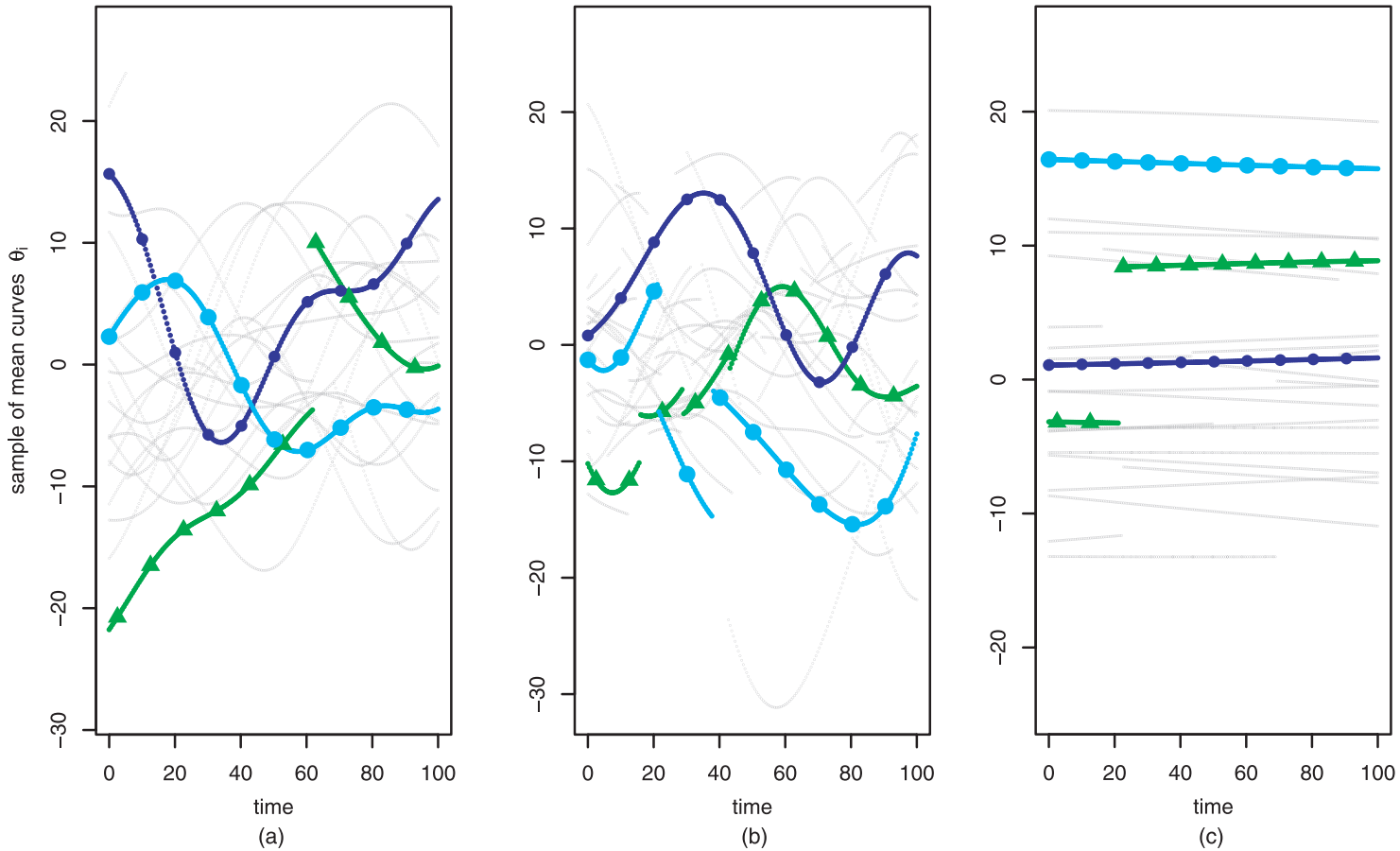


Fig. 1. Samples from the hDP$_k$ prior, for various values of the hyperparameters (some pure and hybrid curves are highlighted; the pure species have continuous trajectories whereas hybrid curves are clearly characterized by their discontinuities; details are given in Section 4.2): (a) $\phi_0 = 3$ and $\phi_q = 0.1$; (b) $\phi_0 = 3$ and $\phi_q = 3$; (c) $\phi_0 = 0.01$ and $\phi_q = 0.1$

Children of the Dirichlet process

- Lessons learned from the DP, and different directions for generalisation
- Conjugacy, neutrality and survival analysis
- Pólya trees
- Species sampling models
- (Doubly) nonparametric regression: dependent DP and kernel stick-breaking priors
- Hybrid DP for high-dimensional and functional parameters
- Hierarchical DP
- Exchangeable partition functions
- Bernstein polynomials

- *Bayesian nonparametrics*, Nils Hjort, Chris Holmes, Peter Müller, and Stephen Walker (2009), Cambridge University Press
- Nonparametric Bayesian Data Analysis, Peter Müller and Fernando A. Quintana (2004), *Statistical Science*, **19**, 95–110. DOI 10.1214/088342304000000017
- Bayesian nonparametric inference for random distributions and related functions, Stephen Walker, Paul Damien, Purushottam Laud and Adrian Smith (1999), *JRSS*, B, **61**, 485-527.

- Email: peter.green@uts.edu.au