Introduction to Bayesian nonparametrics

Tom Thorne

18th June 2018

Overview

- Bayesian nonparametrics
- Finite mixture models
- Dirichlet processes
- Markov Chain Monte Carlo
- Hierarchical Dirichlet processes
- Some other Bayesian nonparametric models

Bayesian nonparametrics

Bayesian parametric models

An example of a Bayesian parametric model:
$X|\theta \sim p_\theta$, $\theta \in \Theta$ with e.g. $\Theta \subset \mathbb{R}^d$
with a prior $p(\theta)$. Then p_θ could be for example a normal distribution.

Bayesian nonparametric models

$X|G \sim G$ with $G \sim Q$

- Does not mean that there are no parameters – instead there are an infinite number.
- Placing a prior on G makes it a *Bayesian* nonparametric model.

Why use Bayesian nonparametric models?

Motivations

Practical:

- Scales automatically with complexity of the data
- Relatively easy to integrate into existing MCMC samplers
- Not computationally demanding

Philosophical:

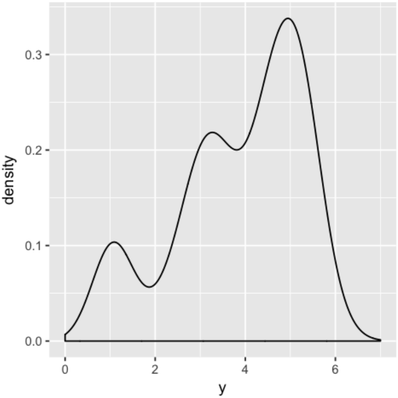
- Prior has a large support
- (*Hopefully*) no need to adjust the prior after observing the data

Bayesian parametric example – Finite mixtures

Observations are generated from a mixture with a fixed number of components:

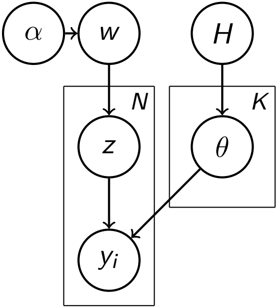
$$p(y|\theta, w) = \sum_{k=1}^{K} w_k p(y|\theta_k)$$

A simple example is a mixture of normal distributions:



Dirichlet priors for finite mixtures

- Prior on parameters H
- Prior on cluster membership probabilities w
- Allocation to clusters z_i
- Observations y_i



$$
\begin{aligned}
\theta_1, \ldots, \theta_K &\sim H \\
w &\sim \mathrm{Dirichlet}\left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right) \\
z_i &\sim \mathrm{Multinomial}(w) \\
y_i &\sim F(\theta_{z_i})
\end{aligned}
$$

Dirichlet priors

We can define the posterior of the model as:

$$p(\theta, z, w|y) \propto p(y|\theta, z)p(z|w)p(w)p(\theta)$$

The Dirichlet prior on w allows us to integrate out to give:

$$p(z|\theta, y) \propto p(y|\theta, z)p(z)$$

Then we can construct a Gibbs sampler that samples from $p(\theta|y, z)$ and $p(z|\theta, y)$

Dirichlet priors

In the Gibbs sampler we can sweep over each z_i and update it conditional on the other parameters:

$$p(z_i = j | y, z_{-i}, \theta) \propto p(y_i | \theta_j) p(z_i = j | z_{-i})$$

Where

$$p(z_i = j | z_{-i}) = \frac{n_j + \frac{\alpha}{K}}{N - 1 + \alpha}$$

with n_j the number of $z_{-i} = j$.

Markov Chain Monte Carlo

```
Initialisation;
for s ∈ 1, ..., Steps do
    for i ∈ 1, ..., N do
        for k ∈ 1, ..., K do
            p_k ← p(y_i|θ_k) p(z_i = k|z_{-i});
        end
        Normalise p;
        Sample z_i ~ Mult(p);
    end
    for k ∈ 1, ..., K do
        Update θ_k given y_j where z_j = k;
    end
    Store z, K, θ
end
```

Markov Chain Monte Carlo

```
Initialisation;
for s ∈ 1, ..., Steps do
    for i ∈ 1, ..., N do
        for k ∈ 1, ..., K do
            p_k ← p(y_i|θ_k)p(z_i = k|z_{-i});
        end
        Normalise p;
        Sample z_i ~ Mult(p);
    end
    for k ∈ 1, ..., K do
        Update θ_k given y_j where z_j = k;
    end
    Store z, K, θ
end
```

Update z for each data point

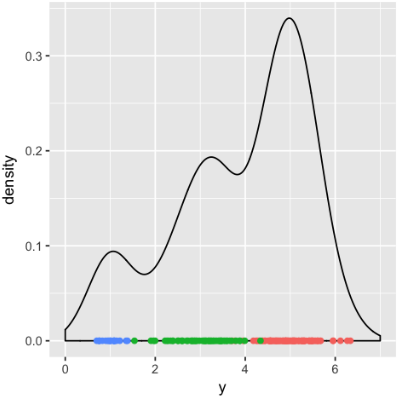
Markov Chain Monte Carlo

Initialisation;
for $s \in 1, \ldots, Steps$ **do**
  for $i \in 1, \ldots, N$ **do**
    for $k \in 1, \ldots, K$ **do**
      $p_k \leftarrow p(y_i|\theta_k)p(z_i = k|z_{-i})$;
    end
    Normalise p;
    Sample $z_i \sim \mathrm{Mult}(p)$;
  end
  for $k \in 1, \ldots, K$ **do**
    Update θ_k given y_j where $z_j = k$;
  end
  Store z, K, θ
end

Update θ for each cluster

Finite mixture - Example

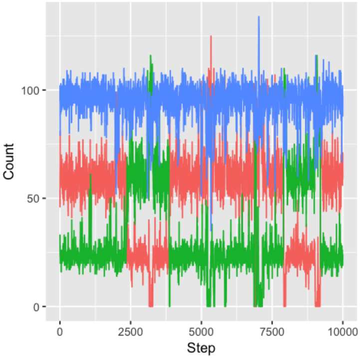
```
y<-c(rnorm(100,5,0.5),rnorm(60,3,0.6),rnorm(20,1,0.2))
```



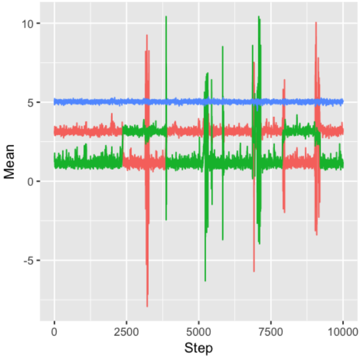
Finite mixture - Example

Markov Chain Monte Carlo traces:

Counts of membership for each cluster

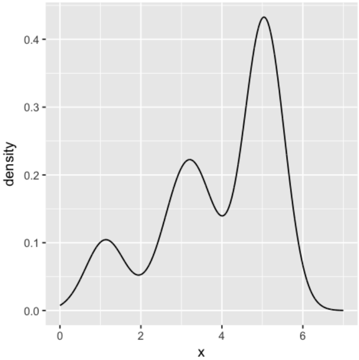


Normal distribution means for each cluster

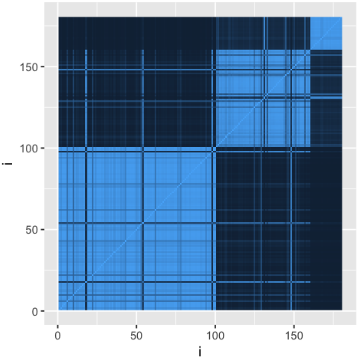


Finite mixture - Example

Density estimation:



Looking at posterior samples, for each z_i, z_j, how often does $z_i = z_j$?



Challenges

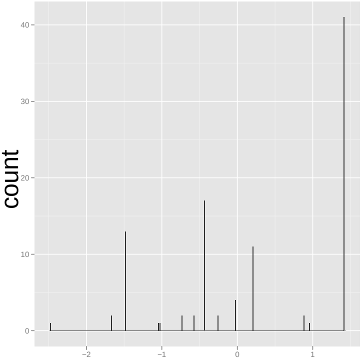
- How do we set the number of clusters?
  - Look at the data?
  - Model selection?
- What happens if we collect a new, much larger data set?

Dirichlet processes

The Dirichlet process is a nonparametric extension of the Dirichlet distribution

$G \sim \mathrm{DP}(\alpha, H)$

- A distribution over distributions
- Samples exhibit clustering behaviour
- Concentration parameter α
- Centering measure H



Ferguson, T. S. A Bayesian Analysis of Some Nonparametric Problems. The Annals of Statistics 1, 209–230 (1973).

Stick breaking construction

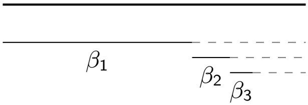
Sethuraman J, A constructive definition of Dirichlet prior. Stat Sin 2:639–650 (1994)

$$u_k|\gamma \sim \mathrm{Beta}(1,\gamma),$$
$$\beta_k = u_k \prod_{i=1}^{k-1}(1-u_i),$$

denoted as $\beta \sim \mathrm{GEM}(\gamma)$. Then if $G \sim \mathrm{DP}(\alpha, H)$:

$$G = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k}$$

where $\theta_k \sim H$.



$\beta_1 = u_1$, $\beta_2 = u_2(1-u_1)$, $\beta_3 = u_3(1-u_2)(1-u_1)$

Polya urn representation

The stick breaking construction is an infinite mixture, and so hard to work with. Fortunately we can marginalise out G.

If $G \sim \mathrm{DP}(\gamma, H)$, for a set of observations $\theta_i \sim G$:

$$\theta_N | \theta_1 \dots \theta_{N-1}, \alpha, H \sim \frac{\alpha}{\alpha + N - 1} H + \sum_{k=1}^{K} \frac{n_k}{\alpha + N - 1} \delta_{\theta_k^*}$$

where $\theta_1^*, \dots, \theta_K^*$ are the unique values in $\theta_1, \dots, \theta_{N-1}$, and n_k is the number of θ_i having value θ_k^*

Blackwell, D. & MacQueen, J. B. Ferguson Distributions Via Polya Urn Schemes. The Annals of Statistics 1, 353–355 (1973).

Polya urn representation

Going back to the Dirichlet distribution with w integrated out:

$$p(z_i = j | z_{-i}) = \frac{n_j + \frac{\alpha}{K}}{N - 1 + \alpha}$$

If $K \to \infty$, then:

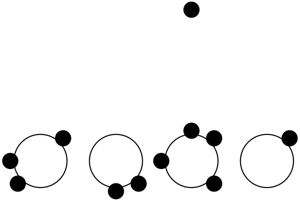
$$p(z_i = j | z_{-i}) = \frac{n_j}{N - 1 + \alpha}$$

and so

$$p(z_i = \text{new} | z_{-i}) = \frac{\alpha}{N - 1 + \alpha}$$

Restaurant process

If $G \sim \mathrm{DP}(\alpha, H)$, for a set of observations $\theta \sim G$:



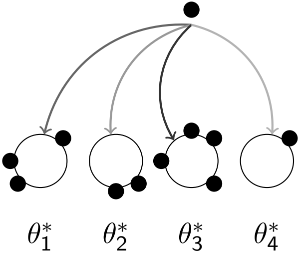
θ_1^* θ_2^* θ_3^* θ_4^*

Analogy for the Dirichlet process due to Pitman and Dubins

D. Aldous, Exchangeability and Related Topics. In l'École d'été de probabilités de Saint-Flour, XIII, pages 1-198. 1983

Restaurant process

If $G \sim \mathrm{DP}(\alpha, H)$, for a set of observations $\theta \sim G$:

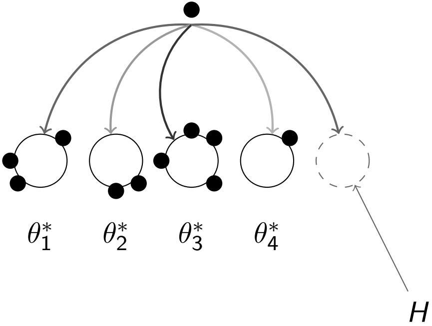


Analogy for the Dirichlet process due to Pitman and Dubins

D. Aldous, Exchangeability and Related Topics. In l'École d'été de probabilités de Saint-Flour, XIII, pages 1-198. 1983

Restaurant process

If $G \sim \mathrm{DP}(\alpha, H)$, for a set of observations $\theta \sim G$:

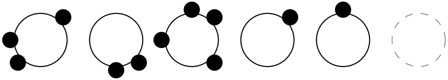


Analogy for the Dirichlet process due to Pitman and Dubins

D. Aldous, Exchangeability and Related Topics. In l'École d'été de probabilités de Saint-Flour, XIII, pages 1-198. 1983

Restaurant process

If $G \sim \mathrm{DP}(\alpha, H)$, for a set of observations $\theta \sim G$:



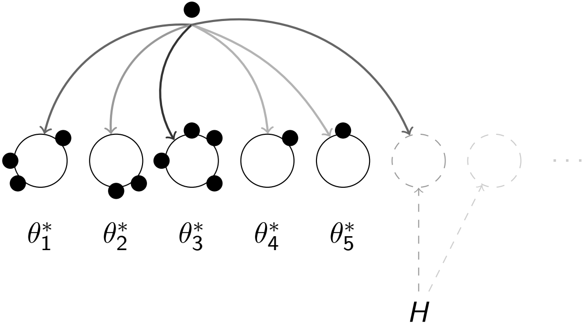
θ_1^* θ_2^* θ_3^* θ_4^* θ_5^*

Analogy for the Dirichlet process due to Pitman and Dubins

D. Aldous, Exchangeability and Related Topics. In l'École d'été de probabilités de Saint-Flour, XIII, pages 1-198. 1983

Restaurant process

If $G \sim \mathrm{DP}(\alpha, H)$, for a set of observations $\theta \sim G$:



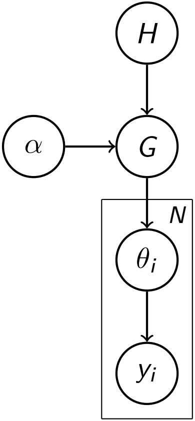
Analogy for the Dirichlet process due to Pitman and Dubins

D. Aldous, Exchangeability and Related Topics. In l'École d'été de probabilités de Saint-Flour, XIII, pages 1-198. 1983

Dirichlet process mixture models

We can use the DP prior in a mixture:

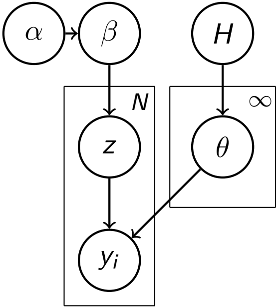
$$G \sim DP(\alpha, H)$$
$$\theta_i \sim G$$
$$y_i \sim F(\theta_i)$$



Dirichlet process mixture models

We can use the DP prior in a mixture:

$$\begin{aligned}\theta_k^* &\sim H\\ \beta &\sim \mathrm{GEM}(\alpha)\\ z_i &\sim \beta\\ y_i &\sim F(\theta_{z_i}^*)\end{aligned}$$



Markov Chain Monte Carlo

- Data points $y_1, \ldots, y_n$
- Latent variables $z_1, \ldots, z_n$
- K parameter sets $\theta_1, \ldots, \theta_K$
- Likelihood $f(y|\theta_k)$

Markov Chain Monte Carlo

Indicator updates

Existing cluster:

$$p(z_i = k | z_{-i}, y, \theta) \propto p(z_i = k | z_{-i}) f(y_i | \theta_k)$$

Where z_{-i} is the set of cluster allocations excluding i.

New cluster:

$$p(z_i = K+1 | z_{-i}, y, \theta) \propto p(z_i = K+1 | z_{-i}) \int_\theta f(y_i | \theta) p(\theta) d\theta$$

When $f(y|\theta)$ and $p(\theta)$ are non-conjugate we can't directly evaluate the integral.

Markov Chain Monte Carlo

Indicator updates

Existing cluster:

$$p(z_i = k | z_{-i}) = \frac{n_k}{N - 1 + \alpha}$$

Where n_k is the size of cluster k in z_{-i}.

New cluster:

$$p(z_i = K + 1 | z_{-i}) = \frac{\alpha}{N - 1 + \alpha}$$

Markov Chain Monte Carlo

for $s \in 1, \ldots, Steps$ **do**
  for $i \in 1, \ldots, N$ **do**
    for $k \in 1, \ldots, K$ **do**
      $p_k \leftarrow f(y_i|\theta_k)p(z_i = k|z_{-i})$;
    end
    for $l \in 1, \ldots, L$ **do**
      Sample $\theta_{K+l} \sim p(\theta)$;
      $p_{K+l} \leftarrow \frac{1}{L} f(y_i|\theta_{K+l})p(z_i = \mathrm{new}|z_{-i})$;
    end
    Sample $z_i \sim \mathrm{Mult}(p)$;
    Tidy z,θ
  end
  for $k \in 1, \ldots, K$ **do**
    Update θ_k given y_j where $z_j = k$;
  end
  Store z, K, θ
end

Algorithm 8 in Neal, R. M. Markov Chain Sampling Methods for Dirichlet Process Mixture Models. Journal of Computational and Graphical Statistics 9, 249 (2000).

Markov Chain Monte Carlo

for $s \in 1, \ldots, Steps$ **do**
    for $i \in 1, \ldots, N$ **do**
        for $k \in 1, \ldots, K$ **do**
            $p_k \leftarrow f(y_i|\theta_k)p(z_i = k|z_{-i})$;
        end
        for $l \in 1, \ldots, L$ **do**
            Sample $\theta_{K+l} \sim p(\theta)$;
            $p_{K+l} \leftarrow \frac{1}{L} f(y_i|\theta_{K+l})p(z_i = \mathrm{new}|z_{-i})$;
        end
        Sample $z_i \sim \mathrm{Mult}(p)$;
        Tidy z,θ
    end
    for $k \in 1, \ldots, K$ **do**
        Update θ_k given y_j where $z_j = k$;
    end
    Store z, K, θ
end

Update z for each data point.

Algorithm 8 in Neal, R. M. Markov Chain Sampling Methods for Dirichlet Process Mixture Models. Journal of Computational and Graphical Statistics 9, 249 (2000).

Markov Chain Monte Carlo

for $s \in 1, \ldots, Steps$ **do**
  for $i \in 1, \ldots, N$ **do**
    for $k \in 1, \ldots, K$ **do**
      $p_k \leftarrow f(y_i|\theta_k)p(z_i = k|z_{-i})$;
    end
    for $l \in 1, \ldots, L$ **do**
      Sample $\theta_{K+l} \sim p(\theta)$;
      $p_{K+l} \leftarrow \frac{1}{L} f(y_i|\theta_{K+l})p(z_i = \text{new}|z_{-i})$;
    end
    Sample $z_i \sim \text{Mult}(p)$;
    Tidy z,θ
  end
  for $k \in 1, \ldots, K$ **do**
    Update θ_k given y_j where $z_j = k$;
  end
  Store z, K, θ
end

If any cluster has no members, delete it. If we created a new cluster, increase K and set θ_K appropriately.

Algorithm 8 in Neal, R. M. Markov Chain Sampling Methods for Dirichlet Process Mixture Models. Journal of Computational and Graphical Statistics 9, 249 (2000).

Markov Chain Monte Carlo

for $s \in 1, \ldots, Steps$ **do**
  for $i \in 1, \ldots, N$ **do**
    for $k \in 1, \ldots, K$ **do**
      $p_k \leftarrow f(y_i|\theta_k)p(z_i = k|z_{-i})$;
    end
    for $l \in 1, \ldots, L$ **do**
      Sample $\theta_{K+l} \sim p(\theta)$;
      $p_{K+l} \leftarrow \frac{1}{L} f(y_i|\theta_{K+l})p(z_i = \mathrm{new}|z_{-i})$;
    end
    Sample $z_i \sim \mathrm{Mult}(p)$;
    Tidy z,θ
  end
  for $k \in 1, \ldots, K$ **do**
    Update θ_k given y_j where $z_j = k$;
  end
  Store z, K, θ
end

Update θ for each cluster.

Algorithm 8 in Neal, R. M. Markov Chain Sampling Methods for Dirichlet Process Mixture Models. Journal of Computational and Graphical Statistics 9, 249 (2000).

Example application

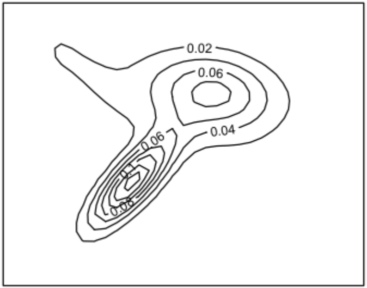
Bivariate normal mixture

```
library(DPpackage)
library(mvtnorm)
# assume cov1,2,3 are covariance matrices
y<-rbind(rmvnorm(n1,c(5,5),cov1),
    rmvnorm(n2,c(3,3),cov2),
    rmvnorm(n3,c(1.5,6),cov3))
mcmc <- list(nburn=1000,nsave=10000,nskip=10)
prior <- list(alpha=1,m1=rep(4,2),
    psiinv1=diag(0.2,2),nu1=4,tau1=1,tau2=100)
results <- DPdensity(y,prior=prior,
    mcmc=mcmc,state=NULL,status=TRUE)
```

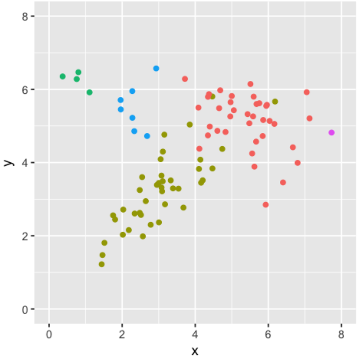
https://cran.r-project.org/web/packages/DPpackage/index.html

Results

Posterior density estimate

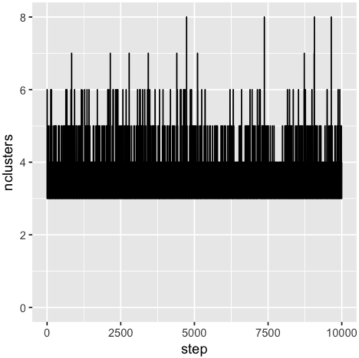


Example of cluster allocation from a single sample

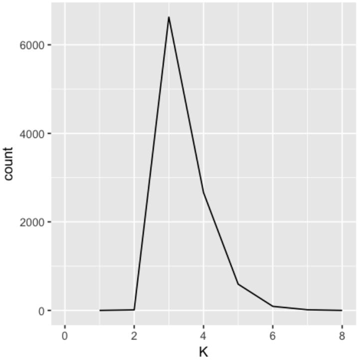


The Dirichlet process has a tendency to infer small extra clusters.

Results

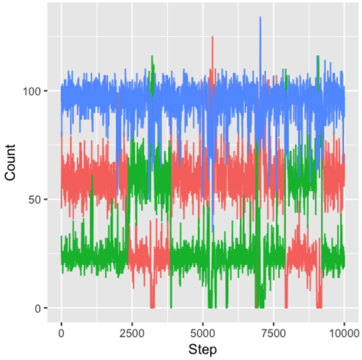


8
6
4
2
0
nclusters
0
2500
5000
7500
10000
step



6000
4000
2000
0
count
0
2
4
6
8
K

Practicalities



Likelihood is invariant under swapping of the labels:

- If we are only interested in density estimation this is not a problem
- Can use MAP estimate of assignments
- Place a constraint on the parameters e.g. $\mu_1 < \mu_2 < \mu_3, \ldots$
- Relabelling strategies

Jasra, A., Holmes, C. & Stephens, D. Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling. Stat Sci 20, 50–67 (2005).

Rodríguez, C. & Walker, S. G. Label Switching in Bayesian Mixture Models: Deterministic Relabeling Strategies. Journal of Computational and Graphical Statistics 23, 25–45 (2014).

Properties

Number of clusters:

- As $n \to \infty$, $\frac{K}{\log n} \to \alpha$
- With fixed concentration parameter, the DP posterior *does not* converge to the true number of components in the mixture.
  Miller, J. W. & Harrison, M. T. Inconsistency of Pitman-Yor process mixtures for the number of components. The Journal of Machine Learning Research 15, 3333–3370 (2014).

Cluster sizes

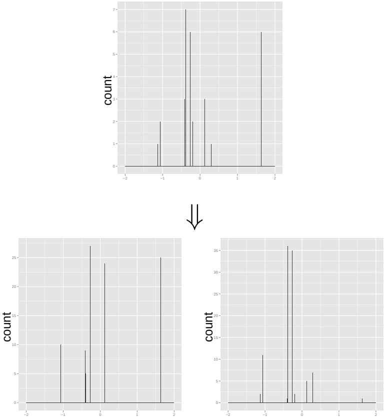
- Prior favours few large clusters and many small ones

An alternative approach:

Miller, J. W. & Harrison, M. T. Mixture models with a prior on the number of components. (2015).

Hierarchical Dirichlet processes

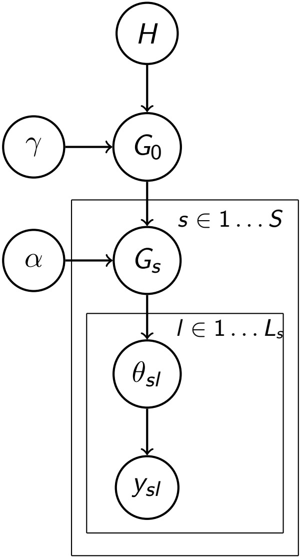
Dirichlet process with another Dirichlet process as base measure



- Divide observations into groups
- Within a group observations are distributed as $G_j \sim \mathrm{DP}(\alpha, G_0)$
- Common measure $G_0 \sim \mathrm{DP}(\gamma, H)$
- Observations all drawn from a shared set of points from the discrete distribution G_0

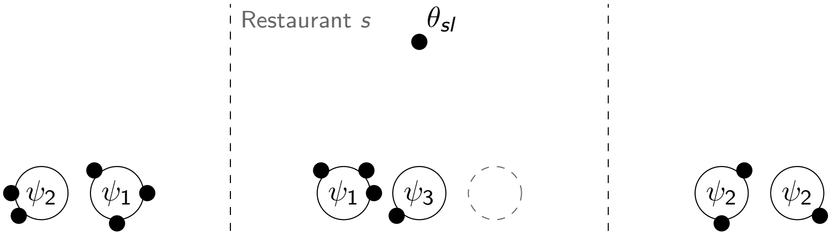
Teh, Y. W., Jordan, M. I., Beal, M. J. & Blei, D. M. Hierarchical Dirichlet Processes. Journal of the American Statistical Association 101, 1566–1581 (2012).

Hierarchical Dirichlet processes



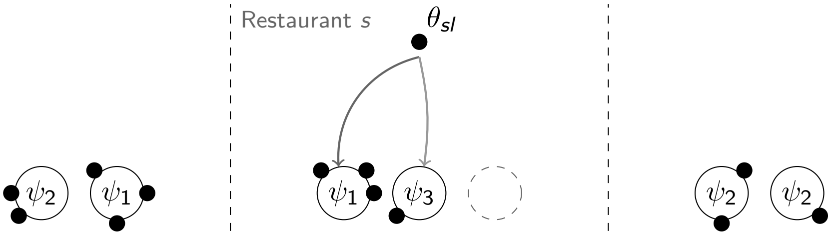
Restaurant franchise

$$\theta_{sl} \sim G_s,\ G_s \sim DP(\alpha, G_0),\ G_0 \sim DP(\gamma, H)$$



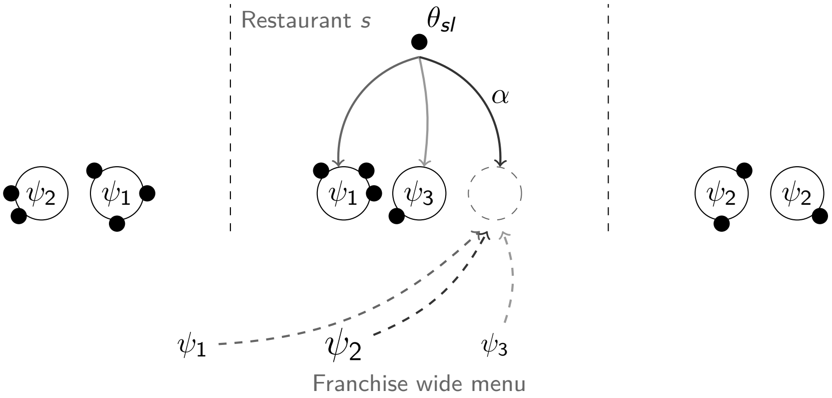
Restaurant franchise

$$\theta_{sl} \sim G_s,\ G_s \sim DP(\alpha, G_0),\ G_0 \sim DP(\gamma, H)$$



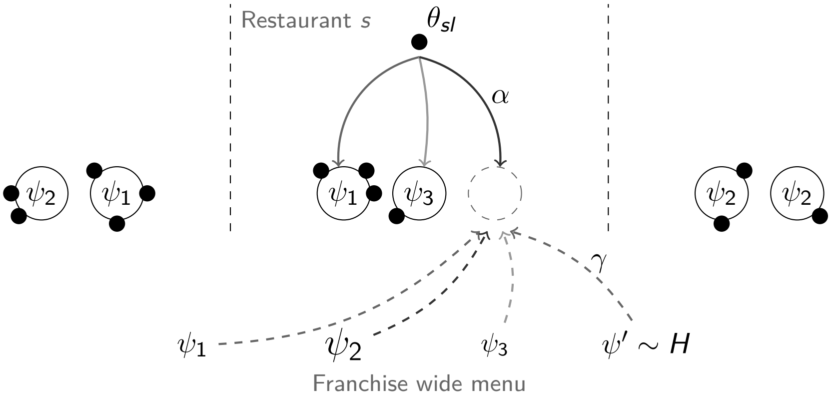
Restaurant franchise

$$\theta_{sl} \sim G_s,\ G_s \sim DP(\alpha, G_0),\ G_0 \sim DP(\gamma, H)$$



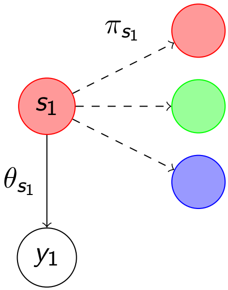
Restaurant franchise

$$\theta_{sl} \sim G_s,\ G_s \sim DP(\alpha, G_0),\ G_0 \sim DP(\gamma, H)$$



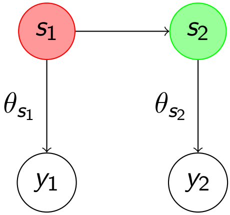
Example - HDP-HMM

Hidden Markov Models – Generally requires prior specification of the number of hidden states



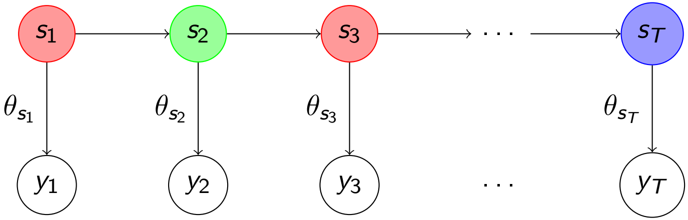
Example - HDP-HMM

Hidden Markov Models – Generally requires prior specification of the number of hidden states



Example - HDP-HMM

Hidden Markov Models – Generally requires prior specification of the number of hidden states

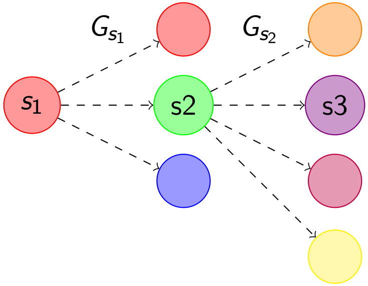


Example - HDP-HMM

- Make transition distribution out of each state G_s with DP prior on G_s?
- Problem – no coupling of the underlying set of states (atoms of $G_1, G_2, \dots$ are independent sets of draws from H)

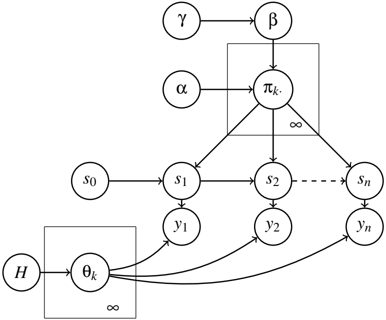
$$s_2 | s_1 = k \sim G_k,\ G_k \sim \mathrm{DP}(\alpha, H)$$

$$s_3 | s_2 = j \sim G_j,\ G_j \sim \mathrm{DP}(\alpha, H)$$



Example - HDP-HMM

- Use a Hierarchical Dirichlet Process to have a common, shared set of states
- Corresponds to the dishes served across all restaurants in the franchise



- Centering measure H
- Shared state distribution β
- Transition distributions $\pi_{i,\cdot}$
- State sequence $s_0, \ldots, s_n$
- Observations $y_1, \ldots, y_n$

Beal, M. J., Ghahramani, Z. & Rasmussen, C. E. The Infinite Hidden Markov Model. (2002).

Some other Bayesian nonparametric priors

- Dependent Dirichlet processes (MacEachern 2000)
- Pitman-Yor processes (Pitman & Yor Annals of Probability, 25, 855–900, 1997)
- Polya trees (Ferguson, Annals of Statistics, 2, 615–629, 1974, Lavine, Annals of Statistics, 20, 1222-1235, 1992)
- Indian Buffet Process (Griffiths & Ghahramani, 2006)
- Gaussian processes

Some further reading

- Müller, P., Quintana, F. A., Jara, A. & Hanson, T. Bayesian Nonparametric Data Analysis. (Springer, 2015).
- Phadia, E. G. Prior Processes and Their Applications. (Springer, 2016).
- Hjort, N. L., Holmes, C., Müller, P. & Walker, S. G. Bayesian Nonparametrics. (Cambridge University Press, 2010).