Lecture 15: Bayesian non-parametrics: the Chinese Restaurant Process and normal mixtures.

Notes and Problem sheets are available at

http://www.stats.ox.ac.uk/~nicholls/BayesMethods/
Notation... repeated values

Sequential simulation of \((\theta_1, ..., \theta_n)\) (L14) may give repeated \(\theta\)-values (if we choose one of the “old” \(\theta\)'s). Write

\[
\sum_{i=1}^{n} \delta_{\theta_i} = \sum_{k=1}^{K} n_k \delta_{\theta^*_k}.
\]

\(\theta^* = (\theta^*_1, ..., \theta^*_K)\) are the unique \(\theta\)-values in \((\theta_1, ..., \theta_n)\).

\(K\) is random, the number of distinct values in \((\theta_1, ..., \theta_n)\).

\(n_k\) is the number of times \(\theta^*_k\) appears in \((\theta_1, ..., \theta_n)\).

For \(k = 1, ..., K\) let \(S_k = \{i; \theta_i = \theta^*_k\}\) so we have a partition \(S = \{S_1, ..., S_K\}\) with \(n_k = |S_k|\).

The mapping \(\theta \rightarrow (\theta^*, S)\) is invertible: for \(i = 1, ..., n\) let \(k_i = \{k : i \in S_k\}\); since \(S\) is a partition \(k_i\) is unique; set \(\theta_i = \theta^*_{k_i}\).
Sequential simulation: generative model for $\theta^*, S$

1. $\theta_1^* \sim H$; set $K = 1$, $S_1 = 1$ and $S = \{S_1\}$.

2. for $j = 1, \ldots, n - 1$
   
   (a) With probability $\alpha/(\alpha + j)$: simulate $\theta_{K+1}^* \sim H$; set $S_{K+1} = \{j + 1\}$ and $S \leftarrow S \cup S_K$; set $K \leftarrow K + 1$.

   (b) Otherwise: for $k = 1, \ldots, K$ set $n_k = |S_k|$; simulate $k \sim (n_1, \ldots, n_K)/j$; set $S_k \leftarrow S_k \cup \{j + 1\}$.

This simulates the marginal distribution $(\theta^*, S)|\alpha, H$.

Finally, set $\theta = \theta(\theta^*, S)$ using the mapping on the previous slide.

If $\tilde{\theta} \sim G$ with $G \sim DP(\alpha, H)$ then $\theta \sim \tilde{\theta}$. We henceforth work with the parameterisation $(\theta^*, S)$.
The Chinese Restaurant Process, AKA CRP

The sequential simulation of parameters according to a $DP(\alpha, H)$ is analogous to restaurant seating!

1) There is $j = 1$ one customer in the restaurant. They are seated at table $k = 1$.

2) Suppose that after the customer $j = 2, 3, ..., n$ arrives, there are $n^j_k$ people seated at table $k$, and $K^j$ tables in all are occupied.

3) The $j + 1$’st arrival chooses a new table with probability $\alpha / (\alpha + j)$ and table $k$ with probability $n^j_k / (\alpha + j)$.

This divides up $n$ customers over $K$ tables. The set $S_k$ lists customers at table $k$.

Put an independent parameter $\theta^*_k \sim H$ on table $k = 1, 2, ..., K$. Together, $(S, \theta^*) \sim G$ with $G \sim DP(\alpha, H)$. 
The CRP probability distribution. Let \( S = (S_1, \ldots, S_K) \) be a partition of \( \{1, 2, \ldots, n\} \), and let \( n_k = \text{card}(S_k) \). The CRP for \( n \) customers realises partition \( S \) with probability

\[
P_\alpha(S) = \frac{\Gamma(\alpha) \alpha^K \prod_{k=1}^{K} \Gamma(n_k)}{\Gamma(\alpha + n)}.
\]

Intuition: Suppose the sequence of table assignments is

\[
T = (1, 1, 2, 1, 2, 3, 3, 2, 2, 4)
\]

for \( n = 10 \) customers so \( S = \{\{1, 2, 4\}, \{3, 5, 8, 9\}, \{6, 7\}, \{10\}\} \).

Table assignment \( T \) and partition \( S \) are 1 to 1, so

\[
P_\alpha(S) = 1 \times \frac{1}{\alpha + 1} \times \frac{\alpha}{\alpha + 2} \times \frac{2}{\alpha + 3} \times \frac{1}{\alpha + 4} \times \frac{\alpha}{\alpha + 5} \times \frac{1}{\alpha + 6} \times \frac{2}{\alpha + 7} \times \frac{3}{\alpha + 8} \times \frac{\alpha}{\alpha + 9}
\]

\[
= \alpha^{K-1} 2! 3! 1! 0! \prod_{i=2}^{10} (\alpha + i - 1)^{-1}.
\]
Proof: The $i$’th arrival brings a denominator factor $(\alpha + i - 1)^{-1}$, so the denominator is $\prod_{i=2}^{n}(\alpha + i - 1)$.

Now look at the numerator. Suppose the customers seated at table $k$ are $S_k = \{i_1, i_2, ..., i_{n_k}\}$. When $i_1$ arrived there was no-one sitting at table $k$, and $k - 1$ tables were occupied, so $i_1$ choses table $k$ wp $\alpha/(\alpha + i_1 - 1)$.

After that, for $j = 2, ..., n_k$, there were $j - 1$ seated at table $k$ when $i_j$ arrived, so $i_j$ chose table $k$ wp $(j - 1)/(\alpha + i_j - 1)$ so the numerator factor is $\alpha(n_k - 1)!$. 
If we end up with \( K \) tables then there are \( K - 1 \) events in which a new table is chosen so

\[
P_\alpha(S) = \frac{\alpha^{K-1} \prod_{k=1}^{K} (n_k - 1)!}{\prod_{i=2}^{n} (\alpha + i - 1)}
\]

\[
\quad = \frac{\alpha^K \prod_{k=1}^{K} (n_k - 1)!}{\prod_{i=1}^{n} (\alpha + i - 1)}
\]

For \( n \) integer \( \Gamma(n) = (n - 1)! \) and for \( x > 0 \), \( x\Gamma(x) = \Gamma(x + 1) \) so we can write this in terms of \( \Gamma \) functions,

\[
P_\alpha(S) = \frac{\Gamma(\alpha) \alpha^K \prod_{k=1}^{K} \Gamma(n_k)}{\Gamma(\alpha + n)}
\]

[\text{EOP}]
The joint distribution of $\theta^*$, $S$

The $\theta^*$-values are simulated independently from $H$. Because

$$\pi(d\theta^*, S) = \pi(S)\pi(d\theta^*|S),$$

we have

$$\pi(d\theta^*, S) = P_\alpha(S) \prod_{k=1}^{K(S)} H(d\theta^*_k)$$

with $K(S) = \text{card}(S)$. Here $S$ acts like a model index.

For example, if $H$ is a distribution on $\Omega = \mathbb{R}^p$ and

$$H(d\theta^*_k) = h(\theta^*_k)d\theta^*_k,$$

so $H$ has density $h$ wrt volume measure in $\mathbb{R}^p$, then

$$\pi(\theta^*, S) = P_\alpha(S) \prod_{k=1}^{K(S)} h(\theta^*_k).$$

The dimension of $\theta^*$ is $K$. 
Let $\Xi_n$ be the set of all partitions of $\{1, \ldots, n\}$.

The state space of $\pi(\theta^*, S')$ is

$$\Omega^* = \bigcup_{S \in \Xi_n} \Omega^K(S) \times \{S\}.$$  

If $A \subset \Omega^*$ and $(\theta^*, S') \sim \pi(\theta^*, S')$ above then

$$\Pr((\theta^*, S') \in A) = \pi(A)$$

$$= \sum_{S \in \Xi_n} \int_{\Omega^K(S)} \mathbb{1}(\theta^*, S') \in A \left[ P_\alpha(S) \prod_{k=1}^{K(S)} h(\theta^*_k) \right] d\theta^*_1, \ldots, d\theta^*_K.$$
Bayesian inference for a Dirichlet process mixture

Suppose our data are $n$ samples $y_i \sim f(y_i|\theta_i), i = 1, \ldots, n$ from a normal mixture with an unknown number of components.

In a DP model for a mixture $y_i \sim f(y_i|\theta_i)$ our prior for $\theta$ is

$$\theta \sim G \quad \text{with} \quad G \sim DP(\alpha, H).$$

In $S, \theta^*$ notation, Bayes rule for conditional probabilities is*

$$\pi(\theta^*, S|y) \propto f(y|\theta^*, S)\pi(S)\pi(\theta^*|S)$$

for $(\theta^*, S) \in \Omega^*$ so

$$\pi(\theta^*, S|y) \propto f(y|\theta^*, S)P_\alpha(S) \prod_{k=1}^{K} h(\theta^*_k).$$

The number of components $K$ in $\theta^*$ is a random variable.

*This is a Very Useful Relation. It takes us straight to an expression for the posterior in a general DP-process mixture! We can write this straight down.
Normal mixture for the Galaxy data

Each component of the mixture has an unknown mean and variance, $\theta^*_k = (\mu^*_k, \sigma^*_k)$. Our base distribution $H$ has density

$$h(\theta^*_k) = h_\mu(\mu^*_k)h_\sigma(\sigma^*_k).$$

If $S = (S_1, \ldots, S_K)$ is a partition of $\{1, 2, \ldots, n\}$ with $n = 82$, and $i \in S_{k_i}$ then

$$y_i|S, \mu^*, \sigma^* \sim N(\mu^*_{k_i}, \sigma^*_k).$$

This determines the likelihood. Our priors are

$$h_\mu(\mu^*_k) = N(\mu^*_k; \mu_0, \sigma_0^2)$$

and

$$h_\sigma(\sigma^*_k) = \Gamma(\sigma^*_k; \alpha_0, \beta_0).$$
with $\mu_0 = 20, \sigma_0 = 10, \alpha_0 = 2$ and $\beta_0 = 1/9$ fixed hyperparameters and I take $\alpha = 1$ in the DP-prior, so the posterior is

\[
\pi(S, \mu^*, \sigma^*|y) \propto f(y; \mu^*, \sigma^*, S) \pi(\mu^*, \sigma^*|S) P_\alpha(S)
\]

\[
\propto \prod_{k=1}^{K} \prod_{i \in S_k} N(y_i; \mu^*_k, \sigma^*_k^2)
\]

\[
\times \prod_{k=1}^{K} N(\mu^*_k; \mu_0, \sigma_0^2) I\Gamma(\sigma^*_k^2; \alpha_0, \beta_0)
\]

\[
\times \alpha^K \prod_{k=1}^{K} \Gamma(n_k).
\]

We dropped the denominator in the expression for $P(S)$ as it does not depend on $S$. 
Remarks on conjugate priors and “Collapsed Gibbs”

I chose Normal/inv-Gamma for the $\mu^*/\sigma^*$-prior to keep things simple and conjugate, so I could Gibbs-sample $\mu^*, \sigma^*$.

If we just did straightforward MH-MCMC on $\mu^*$ and $\sigma^*$ that wouldn’t be necessary.

Conjugate priors are popular in this field as it allows us to integrate out $\theta^* = (\mu^*, \sigma^*)$ completely and just sample the discrete dbn $\pi(S|y)$. This is the “collapsed Gibbs sampler”. It is efficient. If our purpose is clustering, $S$ is all we need anyway.

The downside is we can’t model $\mu^*$ and $\sigma^*$ with freedom.
Remarks on the choice of $\alpha$

I chose $\alpha = 1$. This controls the prior distribution on the number of clusters. I used simulation (of the CRP) to check this distribution was sensible. The prior mean (PS4) is

$$E(K) = \sum_{i=1}^{n} \frac{\alpha}{\alpha + i - 1}.$$  

which is about $E(K) = 5$ here.

This is a bit lower (was 10) than we used in our prior in L13 - we would take $\alpha$ a bit larger if we wanted to match that.

It is often straightforward to impose a hyper-prior on $\alpha$ and infer it along with everything else.
Gibbs sampler for $\mu^*, \sigma^*$ Iterate through the parameters sampling them conditionally. The conditional distribution for $\mu^*_k$ is

$$
\pi(\mu^*_k | \mu^*_{-k}, \sigma^*, y) \propto N(\mu^*_k; \mu_0, \sigma^2_0) \prod_{i \in S_k} N(y_i; \mu^*_k, \sigma^*_k) \cdot N(y_i; \mu^*_k, \sigma^*_k).
$$

We can complete the square and find $\mu^*_k | \sigma^*_k, y \sim N(a, b)$ with

$$
a = b \left( \frac{n_k \bar{y}_k}{\sigma^*_k} + \mu_0 \right), \quad b = \left( \frac{n_k}{\sigma^*_k} + \frac{1}{\sigma^2_0} \right)^{-1},
$$

where $n_k = |S_k|$ and $\bar{y}_k = n_k^{-1} \sum_{i \in S_k} y_i$.

A similar calculation gives $\sigma^*^2 | \mu^*_k, y \sim \text{IG}(c, d)$ with

$$
c = \alpha_0 + n_k/2, \quad d = \beta_0 + \frac{1}{2} \sum_{i \in S_k} (y_i - \mu^*_k)^2.
$$