Bayesian Nonparametric Latent Feature Models

Zoubin Ghahramani
University of Cambridge

Thomas L. Griffiths
University of California at Berkeley

zoubin@eng.cam.ac.uk
http://learning.eng.cam.ac.uk/zoubin/

Newton Institute 2008
Clustering

Basic idea: each data point belongs to a cluster

Many clustering methods exist:

- mixture models
- hierarchical clustering
- spectral clustering

Goal: to partition data into groups in an unsupervised manner
A binary matrix representation for clustering

- Rows are data points
- Columns are clusters
- Since each data point is assigned to one and only one cluster, the rows sum to one.
- Finite mixture models: number of columns is finite
- Infinite mixture models: number of columns is countably infinite
Infinite mixture models  
(e.g. Dirichlet Process Mixtures)

Why?

• You might not believe a priori that your data comes from a finite number of mixture components (e.g. strangely shaped clusters; heavy tails; structure at many resolutions)

• Inflexible models (e.g. a mixture of 6 Gaussians) can yield unreasonable inferences and predictions.

• For many kinds of data, the number of clusters might grow over time: clusters of news stories or emails, classes of objects, etc.

• You might want your method to automatically infer the number of clusters in the data.
Notice that more structure (clusters) appear as you draw more points.

(figure inspired by Neal)
Infinite mixture models

\[ p(x) = \sum_{k=1}^{K} \pi_k p_k(x) \]

How?

- Start from a finite mixture model with \( K \) components and take the limit\(^1\) as number of components \( K \to \infty \)

- But you have infinitely many parameters!

- Rather than optimize the parameters (ML, MAP), you integrate them out (Bayes) using, e.g:
  - sampling (MCMC)
  - variational methods
  - expectation propagation (EP)

\(^1\)Dirichlet Process Mixtures; Chinese Restaurant Processes
A binary matrix representation for clustering

- Rows are data points
- Columns are clusters
- Since each data point is assigned to one and only one cluster...
- ...the rows sum to one.
More General Priors on Binary Matrices

- Rows are data points
- Columns are latent features
- We can think of infinite binary matrices...
  ...where each data point can now have multiple features, so...
  ...the rows can sum to more than one.

(Griffiths and Ghahramani, 2005)
Another way of thinking about this:

- there are multiple overlapping clusters
- each data point can belong to several clusters simultaneously.

If there are $K$ features, then there are $2^K$ possible binary latent representations for each data point.
Why?

- Many statistical models can be thought of as modelling data in terms of hidden or latent variables.

- Clustering algorithms (e.g. using mixture models) represent data in terms of which cluster each data point belongs to.

- But clustering models are restrictive, they do not have distributed representations.

- Consider modelling people’s movie preferences (the “Netflix” problem). A movie might be described using features such as “is science fiction”, “has Charlton Heston”, “was made in the US”, “was made in 1970s”, “has apes in it”... these features may be unobserved (latent).

- The number of potential latent features for describing a movie (or person, news story, image, gene, speech waveform, etc) is unlimited.
From finite to infinite binary matrices

\( z_{nk} = 1 \) means object \( n \) has feature \( k \):

\[
z_{nk} \sim \text{Bernoulli}(\theta_k)
\]

\[
\theta_k \sim \text{Beta}(\alpha/K, 1)
\]

- Note that \( P(z_{nk} = 1|\alpha) = E(\theta_k) = \frac{\alpha/K}{\alpha/K+1} \), so as \( K \) grows larger the matrix gets sparser.

- So if \( Z \) is \( N \times K \), the expected number of nonzero entries is \( N\alpha/(1 + \alpha/K) < N\alpha \).

- Even in the \( K \to \infty \) limit, the matrix is expected to have a finite number of non-zero entries.
From finite to infinite binary matrices

We can integrate out $\theta$, leaving:

\[
P(Z|\alpha) = \int P(Z|\theta)P(\theta|\alpha) d\theta
\]

\[
= \prod_k \frac{\Gamma(m_k + \frac{\alpha}{K})\Gamma(N - m_k + 1)}{\Gamma(\frac{\alpha}{K}) \Gamma(N + 1 + \frac{\alpha}{K})}
\]

The conditional feature assignments are:

\[
P(z_{nk} = 1|z_{-n,k}) = \int_0^1 P(z_{nk}|\theta_k)p(\theta_k|z_{-n,k}) d\theta_k
\]

\[
= \frac{m_{-n,k} + \frac{\alpha}{K}}{N + \frac{\alpha}{K}},
\]

where $z_{-n,k}$ is the set of assignments of all objects, not including $n$, for feature $k$, and $m_{-n,k}$ is the number of objects having feature $k$, not including $n$.

We can take limit as $K \rightarrow \infty$.

From finite to infinite binary matrices

A technical difficulty: the probability for any particular matrix goes to zero as $K \to \infty$:

$$\lim_{K \to \infty} P(Z|\alpha) = 0$$

However, if we consider equivalence classes of matrices in left-ordered form obtained by reordering the columns: $[Z] = lof(Z)$ we get:

$$\lim_{K \to \infty} P([Z]|\alpha) = \exp \left\{ -\alpha H_N \right\} \frac{\alpha^{K_+}}{\prod_{h>0} K_h!} \prod_{k \leq K_+} \frac{(N - m_k)!(m_k - 1)!}{N!}.$$

- $K_+$ is the number of features assigned (i.e. non-zero columns).
- $H_N = \sum_{n=1}^{N} \frac{1}{n}$ is the $N$th harmonic number.
- $K_h$ are the number of features with history $h$ (a technicality).
- This distribution is **infinitely exchangeable**, i.e. it is not affected by the ordering on objects. This is important for its use as a prior in settings where the objects have no natural ordering.
(a) The class matrix on the left is transformed into the class matrix on the right by the function $lof()$. The resulting left-ordered matrix was generated from a Chinese restaurant process (CRP) with $\alpha = 10$.

(b) A left-ordered feature matrix. This matrix was generated from the prior on infinite binary matrices with $\alpha = 10$. 

Binary matrices in left-ordered form
Indian buffet process

“Many Indian restaurants in London offer lunchtime buffets with an apparently infinite number of dishes”

- First customer starts at the left of the buffet, and takes a serving from each dish, stopping after a Poisson($\alpha$) number of dishes as her plate becomes overburdened.
- The $n$th customer moves along the buffet, sampling dishes in proportion to their popularity, serving himself with probability $m_k/n$, and trying a Poisson($\alpha/n$) number of new dishes.
- The customer-dish matrix is our feature matrix, $Z$. 
Two generalizations

- a two-parameter generalization of the Indian Buffet Process
- from binary to non-binary latent features
I. A two-parameter generalization of the IBP?

Limitation:

- The hyperparameter $\alpha$ controls the number of features per object $K_n \overset{\text{def}}{=} \sum_k z_{nk} \sim \text{Poisson}(\alpha)$
- But $\alpha$ also controls the total number of features possessed by a set of $N$ objects, i.e. the variability across rows of $Z$.
- This seems limited—we really want independent control over the mean number of features and the variability across rows.
I. A two-parameter generalization of the IBP

\[ z_{nk} = 1 \text{ means object } n \text{ has feature } k \]

One-parameter IBP

\[ z_{nk} \sim \text{Bernoulli}(\theta_k) \]
\[ \theta_k \sim \text{Beta}(\alpha/K, 1) \]

Two-parameter IBP

\[ z_{nk} \sim \text{Bernoulli}(\theta_k) \]
\[ \theta_k \sim \text{Beta}(\alpha\beta/K, \beta) \]

Properties of the two-parameter IBP

- Number of features per object is \( \text{Poisson}(\alpha) \)
- Setting \( \beta = 1 \) reduces to IBP.
- Parameter \( \beta \) is feature repulsion, \( 1/\beta \) is feature stickiness.
- Total expected number of features is \( \bar{K}_+ = \alpha \sum_{n=1}^{N} \frac{\beta}{\beta + n - 1} \rightarrow \alpha\beta \log N \)
- \( \lim_{\beta \to 0} \bar{K}_+ = \alpha \)
- \( \lim_{\beta \to \infty} \bar{K}_+ = N\alpha \)

Joint work with Peter Sollich
I. A two-parameter generalization of the IBP

First customer starts at the left of the buffet, and takes a serving from each dish, stopping after a Poisson(α) number of dishes as her plate becomes overburdened.

The $n$th customer moves along the buffet, sampling dishes in proportion to their popularity, serving himself with probability $m_k/(\beta - 1 + n)$, and trying a Poisson($\alpha \beta/(\beta - 1 + n)$) number of new dishes.
II. From binary to non-binary latent features

In many models we might want non-binary latent features.

A simple way to generate non-binary latent feature matrices from $Z$:

$$ F = Z \otimes V $$

where $\otimes$ is the elementwise (Hadamard) product of two matrices, and $V$ is a matrix of independent random variables (e.g. Gaussian, Poisson, Discrete, ...).
What do we do with $Z$?

Model data.
Modelling Data

Latent variable model: let $\mathbf{X}$ be the $N \times D$ matrix of observed data, and $\mathbf{Z}$ be the $N \times K$ matrix of binary latent features

$$P(\mathbf{X}, \mathbf{Z}|\alpha, \beta) = P(\mathbf{X}|\mathbf{Z})P(\mathbf{Z}|\alpha, \beta)$$

By combining the IBP with different likelihood functions we can get different kinds of models:

- Models for graph structures (w/ Wood, Griffiths, 2006)
- Models for protein complexes (w/ Chu, Wild, 2006)
- Models for choice behaviour (Görür & Rasmussen, 2006)
- Models for users in collaborative filtering (w/ Meeds, Roweis, Neal, 2006)
- Sparse latent trait, pPCA and ICA models (w/ Knowles, 2007)
- Models for overlapping clusters (w/ Heller, 2007)
Posterior Inference in IBPs

\[ P(Z, \alpha|X) \propto P(X|Z)P(Z|\alpha)P(\alpha) \]

Gibbs sampling:

\[ P(z_{nk} = 1|Z_{-(nk)}, X, \alpha) \propto P(z_{nk} = 1|Z_{-(nk)}, \alpha)P(X|Z) \]

- If \( m_{-n,k} > 0 \),
  \[ P(z_{nk} = 1|z_{-n,k}) = \frac{m_{-n,k}}{N} \]

- For infinitely many \( k \) such that \( m_{-n,k} = 0 \): Metropolis steps with truncation* to sample from the number of new features for each object.

- If \( \alpha \) has a Gamma prior then the posterior is also Gamma \( \rightarrow \) Gibbs sample.

**Conjugate sampler:** assumes that \( P(X|Z) \) can be computed.

**Non-conjugate sampler:** \( P(X|Z) = \int P(X|Z, \theta)P(\theta)d\theta \) cannot be computed, requires sampling latent \( \theta \) as well (e.g. approximate samplers based on (Neal 2000) non-conjugate DPM samplers).

*Slice sampler:* works in the non-conjugate case, is not approximate, and has an adaptive truncation level using a stick-breaking construction of the IBP (Teh, et al, 2007).
An application of IBPs

“A Non-Parametric Bayesian Method for Inferring Hidden Causes”
(Frank Wood, Tom Griffiths, & ZG, 2006)

\[ Y \] - binary latent factors (diseases)
\[ Z \] - graph structure (\( \sim \) IBP)
\[ X \] - observed binary features (symptoms)

“Noisy-or” observations: 
\[ P(x_{nt} = 1|Z, Y, \lambda, \epsilon) = 1 - (1 - \lambda) \sum_k z_{nk} y_{kt} (1 - \epsilon) \]
An application of IBPs

“A Non-Parametric Bayesian Method for Inferring Hidden Causes”
(with Frank Wood and Tom Griffiths)

Gibbs sampling traces

Comparison to RJMCMC

Figure 5: Trace plots and histograms for the Gibbs sampler applied to the signs exhibited by 50 stroke patients. The left column shows the current value of $\epsilon, \lambda, p, \alpha,$ and $K_+$ as the sampler progressed, where $K_+$ is obtained by examining the current $Z$ sample. The right column shows histograms of the same variables computed over the samples.

Figure 3: Learning the number of hidden causes using both RJMCMC and Gibbs sampling. Each line show the mean and standard deviation of the expected value of the dimensionality of the model ($K$ for RJMCMC, and $K_+$ for Gibbs) taken over 500 iterations of sampling for each of 10 datasets.
An application of IBPs

“A Non-Parametric Bayesian Method for Inferring Hidden Causes”
(with Frank Wood and Tom Griffiths)

Inferring stroke localization from patient symptoms:

Figure 6: Causal structure with highest posterior probability. Two grouping of signs are highlighted. In solid black, we find a grouping of poor optokinetic nystagmus, lack of facial control, weakness, decreased rapid alternating movements, abnormal deep tendon reflexes, Babinski sign, and double simultaneous stimulation neglect, all on the left side, consistent with a right frontal/parietal infarct. In dashed black, we find a grouping of comprehension deficit, non-fluency, repetition, anomia, visual field deficit, facial weakness, and general weakness, with the latter three on the right side, generally consistent with a left temporal infarct.

(50 stroke patients, 56 symptoms/signs)
Infinite Independent Components Analysis

Model: \( Y = G(Z \otimes X) + E \)

where \( Y \) is the data matrix, \( G \) is the mixing matrix \( Z \sim \text{IBP}(\alpha, \beta) \) is a mask matrix, \( X \) is heavy tailed sources and \( E \) is Gaussian noise.

Fig. 1. True and inferred \( Z \) and algorithm convergence.

(David Knowles & ZG, 2007)
Modelling Dyadic Data

genes $\times$ patients

Figure 5: Gene expression results. (A) The top-left is $X$ sorted according to contiguous features in the final $U$ and $V$ in the Markov chain. The bottom-left is $V^\top$ and the top-right is $U$. The bottom-right is $W$. (B) The same as (A), but the expected value of $X$, $\hat{X} = UWV^\top$. We have highlighted regions that have both $u_{ik}$ and $v_{jl}$ on. For clarity, we have only shown the (at most) two largest contiguous regions for each feature pair.

users $\times$ movies

(w/ Meeds, Roweis, Neal, 2006)
Summary

- A new class of non-parametric Bayesian model.

http://learning.eng.cam.ac.uk/zoubin
zoubin@eng.cam.ac.uk
Appendix
An application of IBPs - II

“Identifying protein complexes in high-throughput protein interaction screens using an infinite latent feature model” (with Wei Chu, David L Wild, and Roland Krause)

• Data: protein-protein interaction screens.

• Proteins belong to complexes.

• A protein can belong to multiple complexes.

• We don’t know how many complexes there are.

• We want to infer the complexes from the interaction screen data.
An application of IBPs, continued

We use data from affinity purification/mass spectrometry (APMS) experiments:

Individual proteins are tagged and used as baits to form physiological complexes with other proteins in the cell.

(i) Native Protein Complexes

(ii) Purification Matrix, B

(iii) Adjacency Matrix, A

(iv) Complex Membership, Z
An application of IBPs, continued

Adjacency matrix: $A = B^\top B$

Von Neumann Diffusion Kernel: $K := \sum_{\ell=1}^{\infty} \gamma^{\ell-1} A^\ell = A (1 - \gamma A)^{-1}$
where $\gamma$ is the diffusion factor.

The normalized kernel is an appropriate measure of similarity:

$$D_{ij} = \frac{K_{ij}}{\sqrt{K_{ii}K_{jj}}}$$

The prior:

$$Z \sim \text{IBP}(\alpha)$$
where $z_{ci} = 1$ means protein $n$ belongs to complex $c$.

The likelihood:

$$P(D|Z) = \prod_{\{ij: z_i^\top z_j > 0\}} (D_{ij})^{z_i^\top z_j} \prod_{\{ij: z_i^\top z_j = 0\}} (1 - D_{ij})$$

Inference via Gibbs sampling.
The RNA Polymerase complexes. (i) presents the purification results using 9 bait proteins. (ii) presents the corresponding normalized von Neumann kernel matrix, where the gray scale indicates the probability of pairwise membership defined by $D$. The proteins are sorted according to the inferred complex membership.

The four largest complexes identified by our algorithm. The bars indicate the probability of membership of the proteins. The top 3 complexes correspond to RNAP II, RNAP III and RNAP I respectively. The cross indices indicate the members of the three RNA polymerase complexes according to the MIPS protein complex database.
Dirichlet Process Mixtures (Infinite Mixtures)

Start with a finite mixture of $K$ components to model data $\{x^{(1)}, \ldots x^{(N)}\}$:

$$p(x^{(n)}|\theta) = \sum_{k=1}^{K} \pi_k p_k(x^{(n)}|\theta_k)$$

Represent using latent indicator variables:

$$p(x^{(n)}|\theta) = \sum_{k=1}^{K} P(s^{(n)} = k|\pi) p_k(x^{(n)}|\theta_k, s^{(n)} = k)$$

where $s^{(n)} = k$ means $x^{(n)}$ came from component $k$. Distribution of indicators $s = (s^{(1)}, \ldots, s^{(N)})$ given $\pi$ is multinomial

$$P(s^{(1)}, \ldots, s^{(N)}|\pi) = \prod_{k=1}^{K} \pi_k^{N_k}, \quad N_k \overset{\text{def}}{=} \sum_{n=1}^{N} \delta(s^{(n)}, k).$$

Give the mixing proportions $\pi$ a symmetric Dirichlet prior

$$p(\pi|\alpha) = \frac{\Gamma(\alpha)}{\Gamma(\alpha/K)^K} \prod_{k=1}^{K} \pi_k^{\alpha/K-1}$$
Dirichlet Process Mixtures (Infinite Mixtures) - II

Distribution of indicators $s = (s^{(1)}, \ldots, s^{(N)})$ given $\pi$ is multinomial

$$P(s^{(1)}, \ldots, s^{(N)}|\pi) = \prod_{k=1}^{K} \pi_k^{N_k}, \quad N_k \overset{\text{def}}{=} \sum_{n=1}^{N} \delta(s^{(n)}, k).$$

Give the mixing proportions $\pi$ a symmetric Dirichlet prior

$$p(\pi|\alpha) = \frac{\Gamma(\alpha)}{\Gamma(\alpha/K)^K} \prod_{k=1}^{K} \pi_k^{\alpha/K - 1}$$

Integrate out the mixing proportions, $\pi$, to obtain:

$$P(s^{(1)}, \ldots, s^{(N)}|\alpha) = \int d\pi \ P(s|\pi)P(\pi|\alpha) = \frac{\Gamma(\alpha)}{\Gamma(N + \alpha)} \prod_{k=1}^{K} \frac{\Gamma(N_k + \alpha/K)}{\Gamma(\alpha/K)}$$

Take limit as $K \to \infty$, but instead of looking at distribution over indicator variables, $s^{(n)}$, consider the corresponding distribution over partitions of $n$ data points.\(^2\)

E.g.: (1 2 3) (4) (5) or (1 3) (2 5) (4) or (1 2 3 4 5) or (1) (2 5) (3) (4)...

\(^2\)The number of such partitions is the $n$th Bell number
Chinese Restaurant Process

The CRP generates samples from the distribution over partitions induced by a DPM.

Generating from a CRP:

- customer 1 enters the restaurant and sits at table 1. 
  \( K = 1, \ N = 1, \ N_1 = 1 \)

- for \( N = 2, \ldots \),
  
  - customer \( N \) sits at table \( k \) with prob \( \frac{N_k}{N-1+\alpha} \) for \( k = 1 \ldots K \)
  
  - \( K + 1 \) with prob \( \frac{\alpha}{N-1+\alpha} \) (new table)

  if new table was chosen then \( K \leftarrow K + 1 \) endif

endfor

“Rich get richer” property.

(Aldous 1985; Pitman 2002)
Clustering with Gaussian Mixtures
(Density Estimation)

Data: $\mathcal{D} = \{x^{(n)}\}$ for $n = 1, \ldots, N$

$\mathbf{x}^{(n)} \in \mathbb{R}^D$

Parameters: $\theta = ((\mu^{(1)}, \Sigma^{(1)}) \ldots, (\mu^{(m)}, \Sigma^{(m)}), \pi)$

Model:

$x^{(n)} \sim \sum_{n=1}^{m} \pi_n p_i(x^{(n)})$

where

$p_i(x^{(n)}) = \mathcal{N}(\mu^{(i)}, \Sigma^{(i)})$

Goal: To infer $\theta$ from the data and predict the density $p(x|\mathcal{D}, m)$
Bayes Rule Applied to Machine Learning

\[ P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} \]

\( P(D|\theta) \) likelihood of \( \theta \)
\( P(\theta) \) prior probability of \( \theta \)
\( P(\theta|D) \) posterior of \( \theta \) given \( D \)

Model Comparison:

\[ P(m|D) = \frac{P(D|m)P(m)}{P(D)} \]

\[ P(D|m) = \int P(D|\theta, m)P(\theta|m) \, d\theta \]

Prediction:

\[ P(x|D, m) = \int P(x|\theta, D, m)P(\theta|D, m) \, d\theta \]

\[ P(x|D, m) = \int P(x|\theta, m)P(\theta|D, m) \, d\theta \quad \text{(if } x \text{ is iid given } \theta) \]
Model Comparison: two examples

- e.g. selecting $m$, the number of Gaussians in a mixture model
  \[ P(m|\mathcal{D}) = \frac{P(\mathcal{D}|m)P(m)}{P(\mathcal{D})}, \]

- e.g. selecting $m$ the order of a polynomial in a nonlinear regression model
  \[ P(D|m) = \int P(D|\theta, m)P(\theta|m)\,d\theta \]

A possible procedure:
1. place a prior on $m$, $P(m)$
2. given data, use Bayes rule to infer $P(m|\mathcal{D})$

What is the problem with this procedure?
Real data is complicated

Example 1:
You are trying to model people’s patterns of movie preferences. You believe there are “clusters” of people, so you use a mixture model...

- How should you pick $P(m)$, your prior over how many clusters there are? teenagers, people who like action movies, people who like romantic comedies, people who like horror movies, people who like movies with Marlon Brando, people who like action movies but not science fiction, etc etc...

- Even if there are a few well defined clusters, they are unlikely to be Gaussian in the variables you measure. To model complicated distributions you might need many Gaussians for each cluster.

- Conclusion: any small finite number seems unreasonable
Real data is complicated

Example 2:
You are trying to model crop yield as a function of rainfall, amount of sunshine, amount of fertilizer, etc. You believe this relationship is nonlinear, so you decide to model it with a polynomial.

• How should you pick $P(m)$, your prior over what is the order of the polynomial?

• Do you believe the relationship could be linear? quadratic? cubic? What about the interactions between input variables?

• Conclusion: any order polynomial seems unreasonable.

How do we adequately capture our beliefs?
Non-parametric Bayesian Models

- Bayesian methods are most powerful when your prior adequately captures your beliefs.

- Inflexible models (e.g. mixture of 5 Gaussians, 4th order polynomial) yield unreasonable inferences.

- Non-parametric models are a way of getting very flexible models.

- Non-parametric models can automatically infer an adequate model size/complexity from the data, without needing to explicitly do Bayesian model comparison.\(^3\)

\(^3\)Even if you believe there are infinitely many possible clusters, you can still infer how many clusters are represented in a finite set of \(n\) data points.
Dirichlet Process Mixtures (Infinite Mixtures) - III

Starting from

\[
P(s|\alpha) = \frac{\Gamma(\alpha)}{\Gamma(n + \alpha)} \prod_{j=1}^{K} \frac{\Gamma(n_j + \alpha/K)}{\Gamma(\alpha/K)}
\]

**Conditional Probabilities: Finite K**

\[
P(s^{(i)} = j|s_{-n}, \alpha) = \frac{n_{-n,j} + \alpha/K}{n - 1 + \alpha}
\]

where \(s_{-n}\) denotes all indices except \(n\), and \(n_{-n,j} \overset{\text{def}}{=} \sum_{\ell \neq i} \delta(s^{(\ell)}, j)\)

DP: more populous classes are more likely to be joined

**Conditional Probabilities: Infinite \(K\)**

Taking the limit as \(K \to \infty\) yields the conditionals

\[
P(s^{(i)} = j|s_{-n}, \alpha) = \begin{cases} 
\frac{n_{-n,j}}{n-1+\alpha} & \text{\(j\) represented} \\
\frac{\alpha}{n-1+\alpha} & \text{all \(j\) not represented}
\end{cases}
\]

Left over mass, \(\alpha\), ⇒ **countably infinite** number of indicator settings.