The Indian Buffet Process and Extensions

Zoubin Ghahramani

University of Cambridge

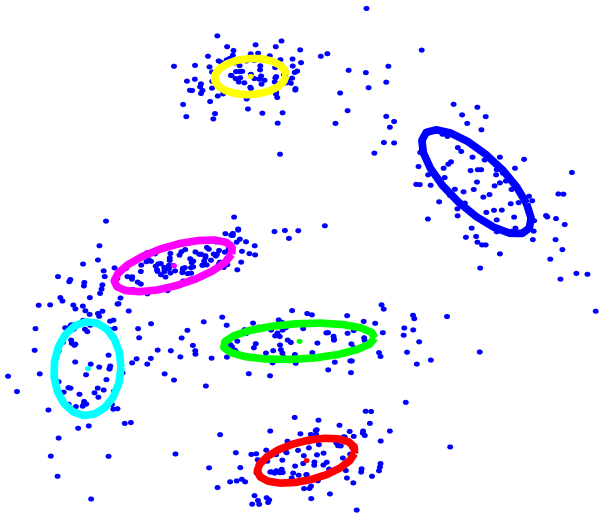
zoubin@eng.cam.ac.uk
http://learning.eng.cam.ac.uk/zoubin/

Bayesian Nonparametrics Workshop

Moncalieri, Italy 2009

Clustering

Basic idea: each data point belongs to a cluster



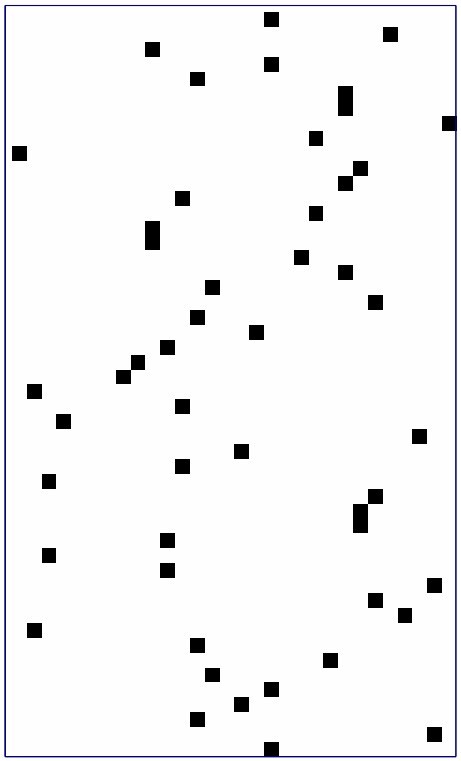
Goals:

- to model the distribution of data;
- to partition data into groups;
- to infer the number of groups

A Classical Approach: mixture modelling with finitely many components

A Bayesian Nonparametric Approach: Dirichlet process mixtures, with countably infinitely many components

A binary matrix representation for clustering

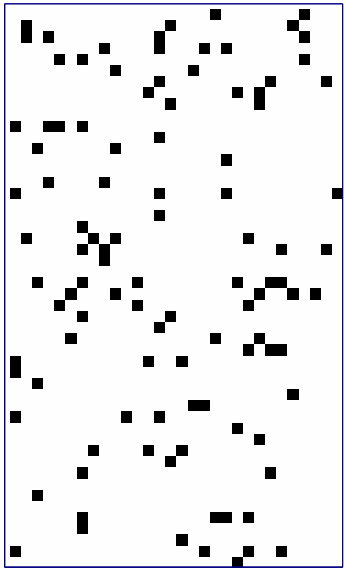


- Rows are data points
- Columns are clusters
- Since each data point is assigned to one and only one cluster, rows sum to one.
- Finite mixture models: number of columns is finite
- Dirichlet Process Mixtures (DPM): number of columns is countably infinite

The Chinese restaurant process (CRP; Aldous, 1985) is the distribution on partitions of the data induced by a DPM.

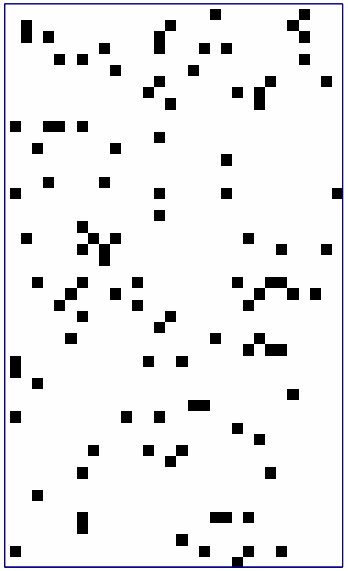
Thus, we can think of the CRP as a distribution on such binary matrices.

More general distributions 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.

More general distributions on binary matrices



Another way of thinking about this:

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

If there are K latent features, then there are 2^K possible settings of the binary latent features 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...
- 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.

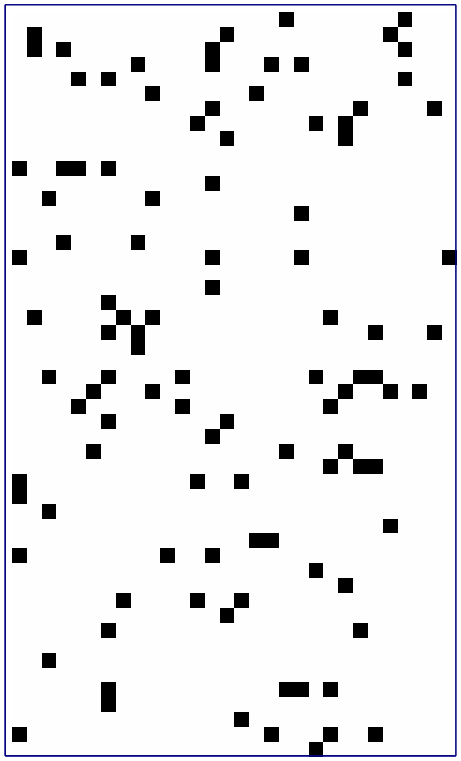
Other reasons: graph structures, stick breaking, Beta processes, time series!

From finite to infinite binary matrices

$z_{nk} = 1$ means object n has feature k:

$$z_{nk} \sim \text{Bernoulli}(\theta_k)$$

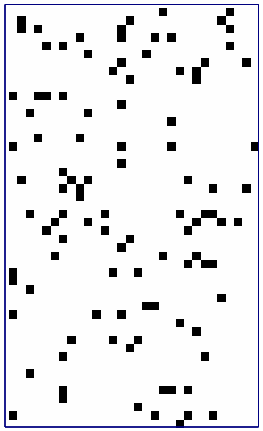
$$\theta_k \sim \text{Beta}(\alpha/K, 1)$$



- 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 $\mathbf{Z}$ 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.

(Griffiths and Ghahramani, 2005)

From finite to infinite binary matrices



We can integrate out θ, leaving:

$$
\begin{aligned}
P(\mathbf{Z}|\alpha) &= \int P(\mathbf{Z}|\theta)P(\theta|\alpha)d\theta \\
&= \prod_k \frac{\Gamma(m_k + \frac{\alpha}{K})\Gamma(N - m_k + 1)}{\Gamma(\frac{\alpha}{K})} \frac{\Gamma(1 + \frac{\alpha}{K})}{\Gamma(N + 1 + \frac{\alpha}{K})}
\end{aligned}
$$

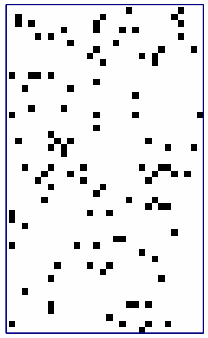
The conditional feature assignments are:

$$
\begin{aligned}
P(z_{nk} = 1|\mathbf{z}_{-n,k}) &= \int_0^1 P(z_{nk}|\theta_k)p(\theta_k|\mathbf{z}_{-n,k})\ d\theta_k \\
&= \frac{m_{-n,k} + \frac{\alpha}{K}}{N + \frac{\alpha}{K}}
\end{aligned}
$$

where $\mathbf{z}_{-n,k}$ 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 $K \to \infty$.

"Rich get richer", like in Chinese Restaurant Processes.

From finite to infinite binary matrices



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

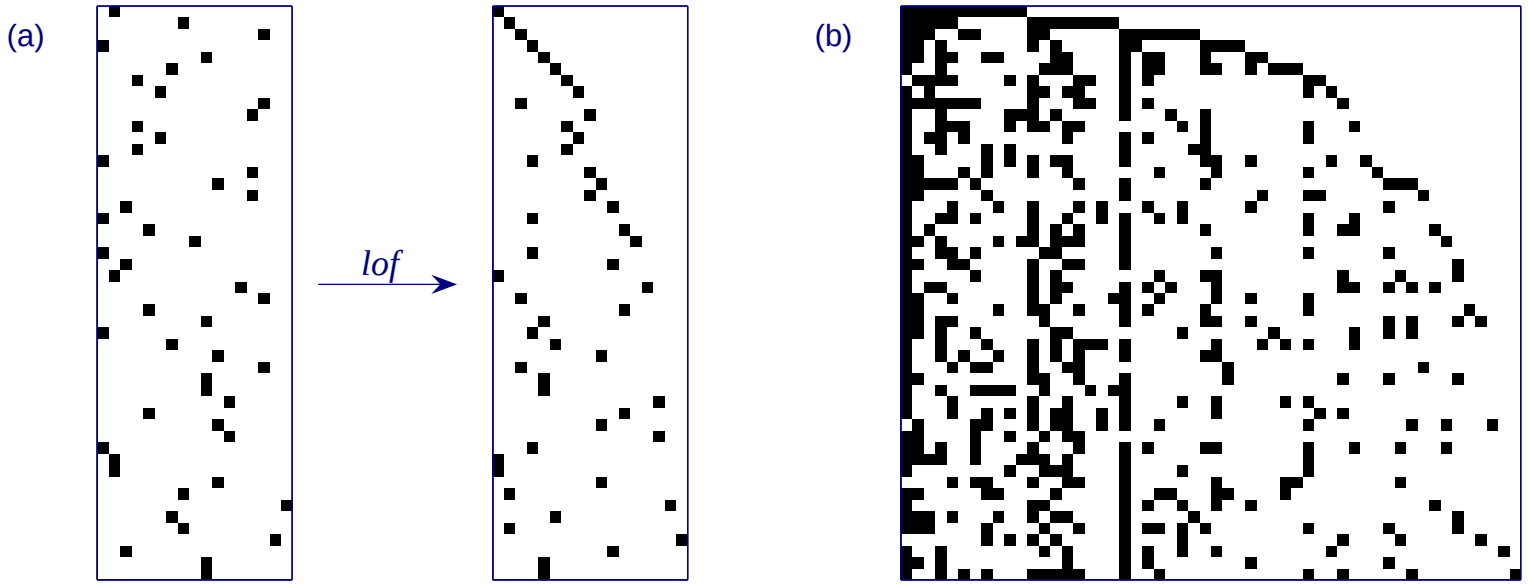
$$\lim_{K\to\infty} P(\mathbf{Z}|\alpha) \quad = \quad 0$$

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

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

- K_+ is the number of features assigned (i.e. non-zero columns).
- $H_N = \sum_{n=1}^{N} \frac{1}{n}$ is the Nth 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.

Binary matrices in left-ordered form

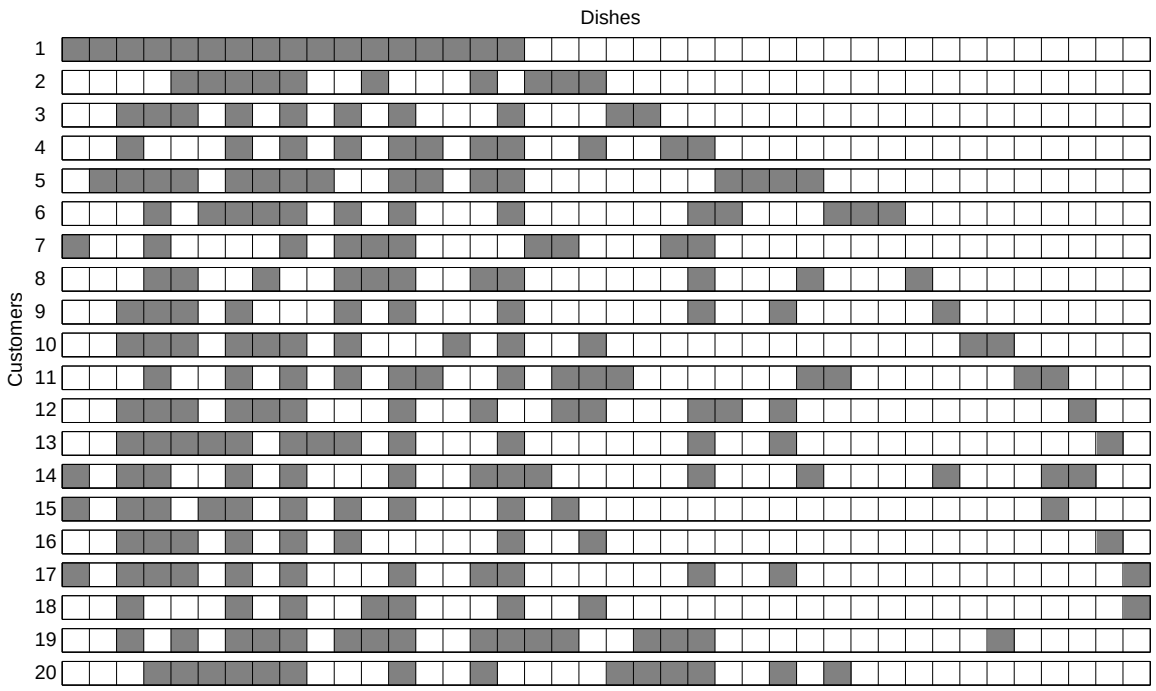


(a) The matrix on the left is transformed into the matrix on the right by the function $lof()$. The resulting left-ordered matrix was generated from a Chinese restaurant process (CRP) with $\alpha = 10$.

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

Indian buffet process

(Griffiths and Ghahramani, 2005)

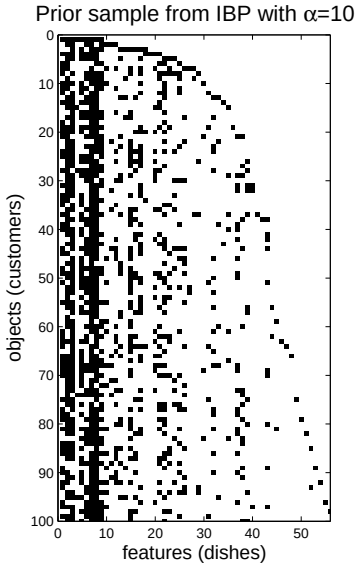


"Many Indian restaurants in London offer lunchtime buffets with an apparently infinite number of dishes"

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

Properties of the Indian buffet process

$$P([\mathbf{Z}]|\alpha) = \exp\{-\alpha H_N\}\frac{\alpha^{K_+}}{\prod_{h>0} K_h!}\prod_{k\leq K_+}\frac{(N-m_k)!(m_k-1)!}{N!}$$



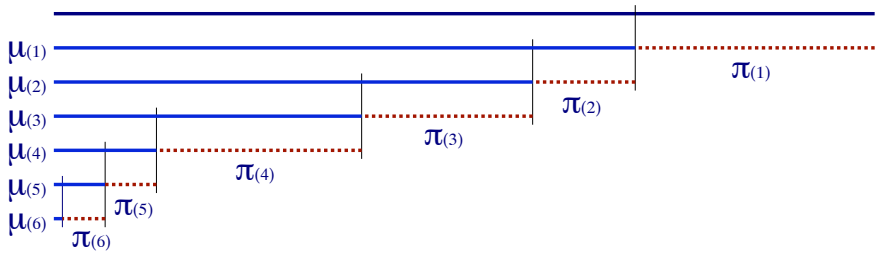


Figure 1: Stick-breaking construction for the DP and IBP. The black stick at top has length 1. At each iteration the vertical black line represents the break point. The brown dotted stick on the right is the weight obtained for the DP, while the blue stick on the left is the weight obtained for the IBP.

Shown in (Griffiths and Ghahramani, 2005):

- It is infinitely exchangeable.
- The number of ones in each row is Poisson(α)
- The expected total number of ones is αN.
- The number of nonzero columns grows as $O(\alpha \log N)$.

Additional properties:

- Has a stick-breaking representation (Teh, Görür, Ghahramani, 2007)
- Can be interpreted using a Beta-Bernoulli process (Thibaux and Jordan, 2007)

What do we do with $\mathbf{Z}$?

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

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

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 overlapping clusters (w/ Heller, 2007)
- Models for choice behaviour (Görür, Jäkel & Rasmussen, 2006)
- Models for users in collaborative filtering (w/ Meeds, Roweis, Neal, 2006)
- Sparse latent factor models (w/ Knowles, 2007)

Posterior Inference in IBPs

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

Gibbs sampling: $P(z_{nk} = 1|\mathbf{Z}_{-(nk)}, \mathbf{X}, \alpha) \propto P(z_{nk} = 1|\mathbf{Z}_{-(nk)}, \alpha)P(\mathbf{X}|\mathbf{Z})$

- If $m_{-n,k} > 0$, $P(z_{nk} = 1|\mathbf{z}_{-n,k}) = \frac{m_{-n,k}}{N}$
- 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 α has a Gamma prior then the posterior is also Gamma $\rightarrow$ Gibbs sample.

Conjugate sampler: assumes that $P(\mathbf{X}|\mathbf{Z})$ can be computed.

Non-conjugate sampler: $P(\mathbf{X}|\mathbf{Z}) = \int P(\mathbf{X}|\mathbf{Z}, \theta)P(\theta)d\theta$ cannot be computed, requires sampling latent θ as well (c.f. (Neal 2000) non-conjugate DPM samplers).

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

Particle Filter: (Wood & Griffiths, 2007).

Accelerated Gibbs Sampling: maintaining a probability distribution over some of the variables (Doshi-Velez & Ghahramani, 2009).

Variational inference: (Doshi-Velez, Miller, van Gael, & Teh, 2009).

An application of IBPs

"A Non-Parametric Bayesian Method for Inferring Hidden Causes"
(Wood, Griffiths, Ghahramani, 2006)

Inferring stroke localization from patient symptoms:

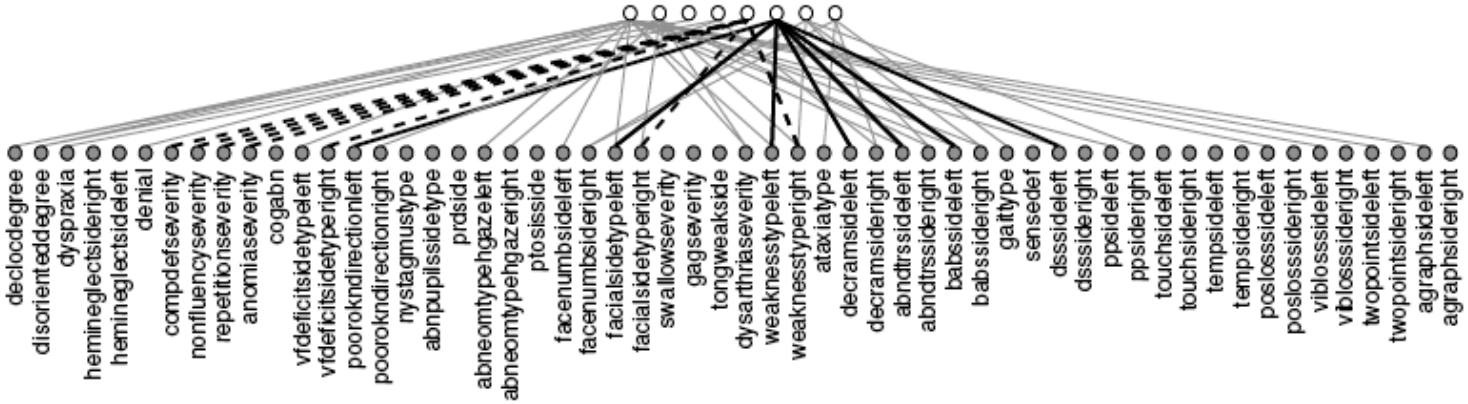
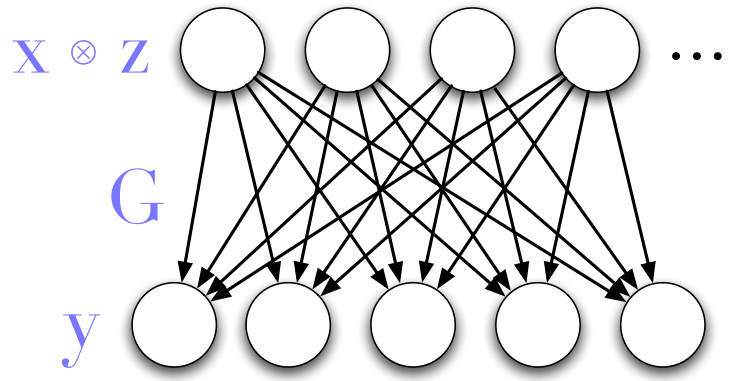


Figure 6: Causal structure with highest posterior probability. Two grouping of signs are highlighted. In solid black, we find a grouping of poor optokinetic nystagmus, lack of facial control, weakness, decreased rapid alternating movements, abnormal deep tendon reflexes, Babinski sign, and double simultaneous 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)

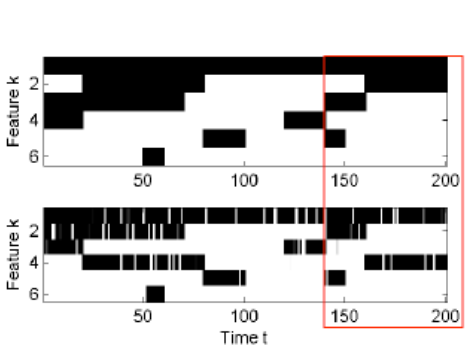
The IBP models the graph structure connecting hidden causes to symptoms

Infinite Sparse Latent Factor Models

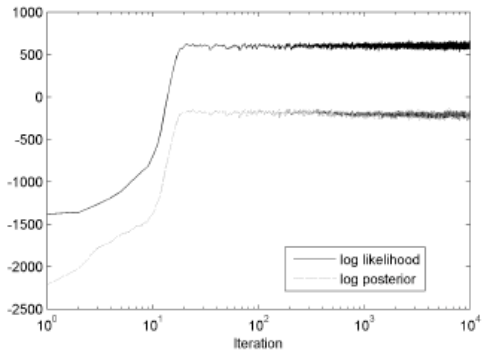


Model: $\mathbf{Y} = \mathbf{G}(\mathbf{Z} \otimes \mathbf{X}) + \mathbf{E}$

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



(a) *Top:* True $\mathbf{Z}$. *Bottom:* Inferred $\mathbf{Z}$. Red box denotes test data.



(b) Plot of the log likelihood and posterior for the duration of the iICA$_2$ run.

Fig. 1. True and inferred $\mathbf{Z}$ and algorithm convergence.

The IBP models the sparsity structure in the latent variables

(w/ Knowles, 2007)

Modelling Dyadic Data

genes $\times$ patients

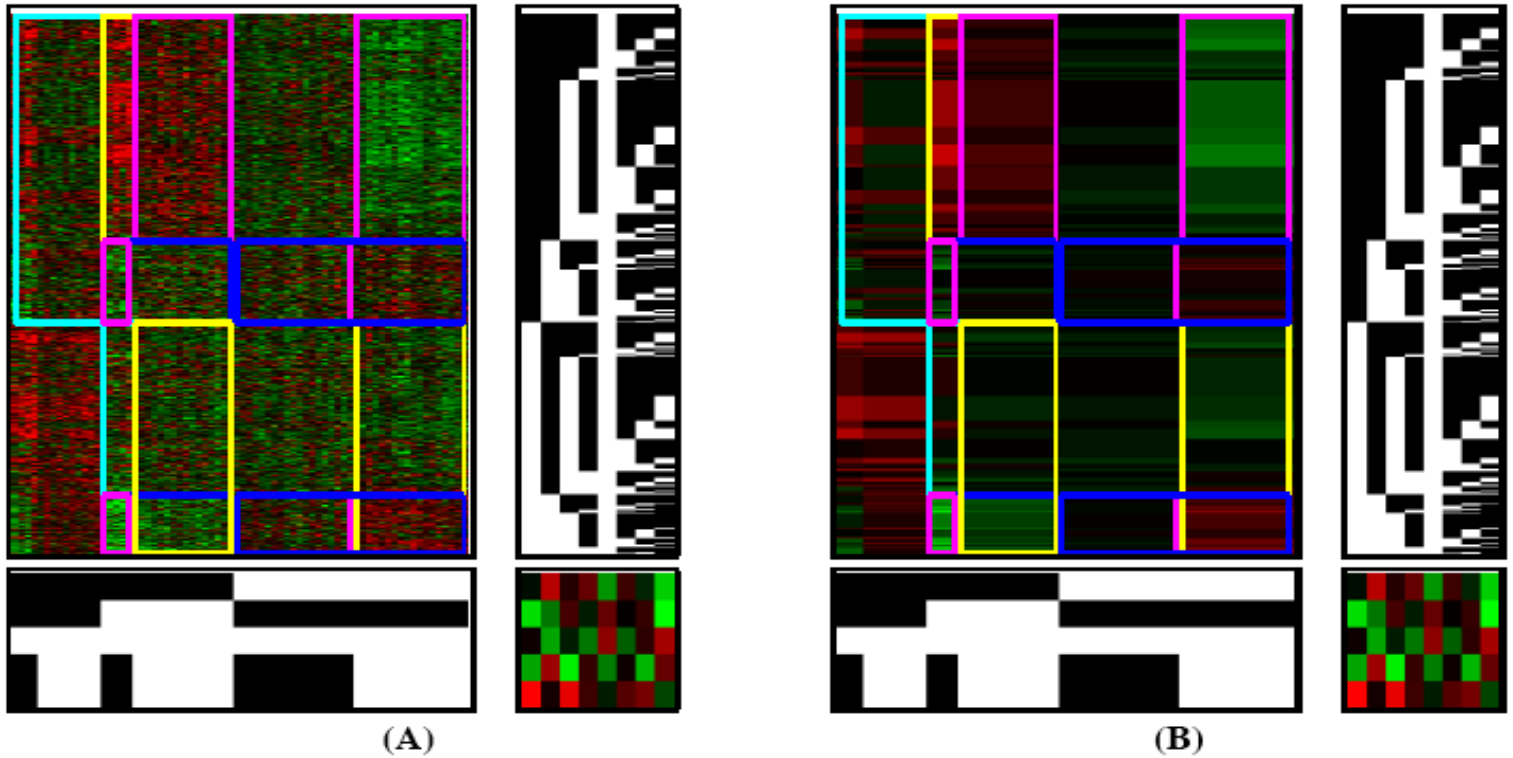


Figure 5: Gene expression results. **(A)** The top-left is $\mathbf{X}$ sorted according to contiguous features in the final $\mathbf{U}$ and $\mathbf{V}$ in the Markov chain. The bottom-left is $\mathbf{V}^\top$ and the top-right is $\mathbf{U}$. The bottom-right is $\mathbf{W}$. **(B)** The same as **(A)**, but the expected value of $\mathbf{X}$, $\hat{\mathbf{X}} = \mathbf{U}\mathbf{W}\mathbf{V}^\top$. We have hilighted 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

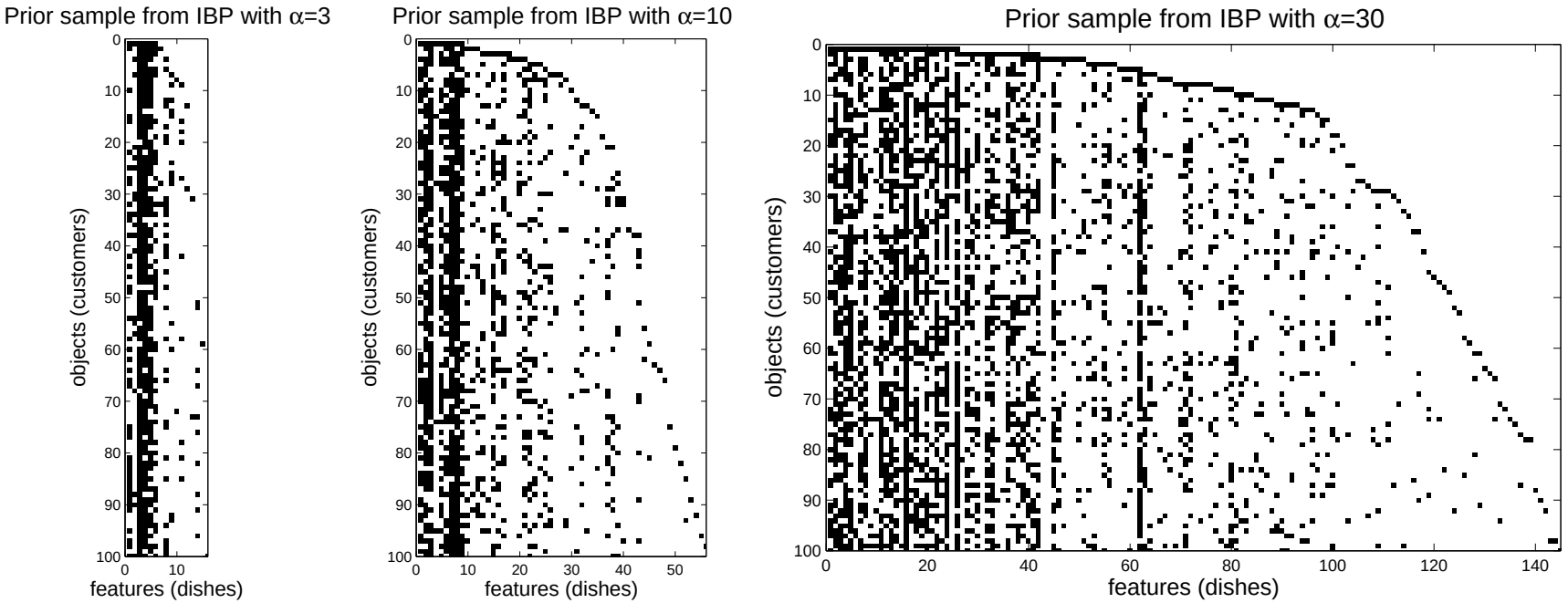
The IBP models latent features of genes, patients, users, movies.

(w/ Meeds, Roweis, Neal, 2006)

Three generalizations

- a two-parameter generalization of the Indian buffet process
- from binary to non-binary latent features
- time series models

I. A two-parameter generalization of the IBP?



Limitation:

- The hyperparameter α controls the number of features per object $K_n \stackrel{\text{def}}{=} \sum_k z_{nk} \sim \text{Poisson}(\alpha)$
- But α also controls the total number of features possessed by a set of N objects, i.e. the variability across rows of $\mathbf{Z}$.
- 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

$z_{nk} = 1$ means object n has feature k

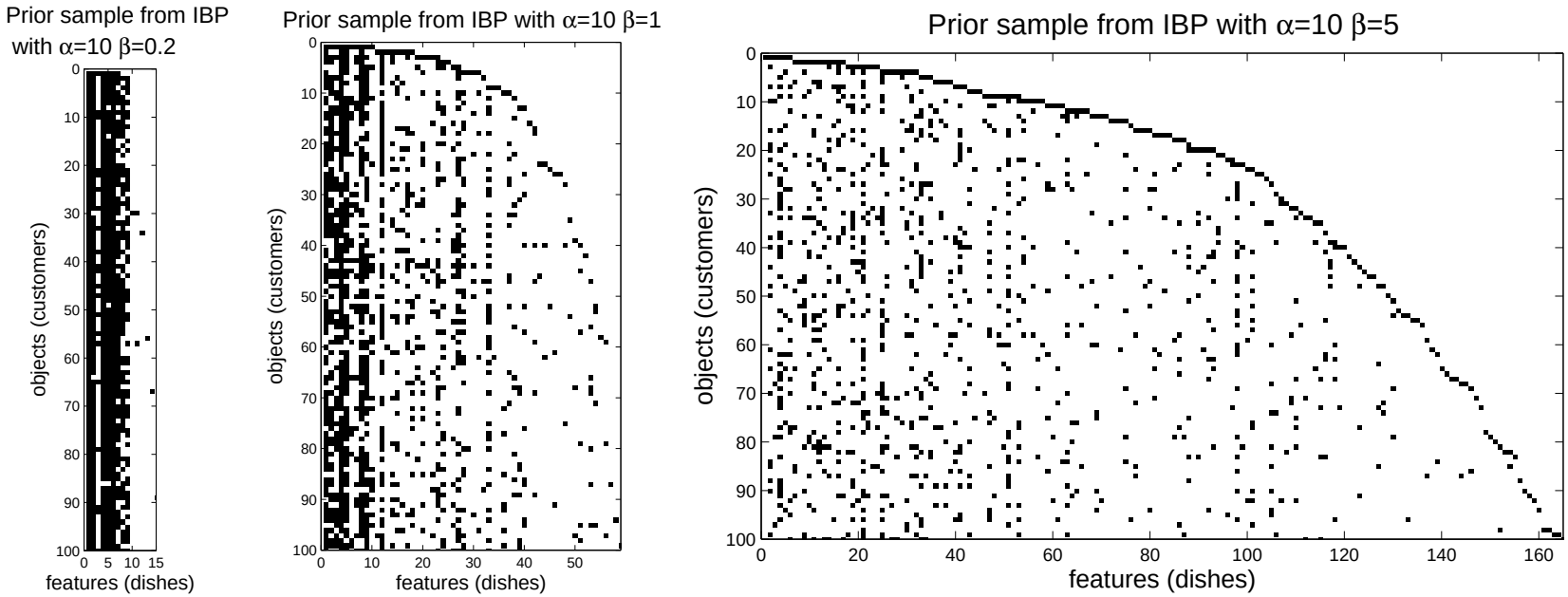
One-parameter IBP	Two-parameter IBP
$z_{nk} \sim \text{Bernoulli}(\theta_k)$	$z_{nk} \sim \text{Bernoulli}(\theta_k)$
$\theta_k \sim \text{Beta}(\alpha/K, 1)$	$\theta_k \sim \text{Beta}(\alpha\beta/K, \beta)$

Properties of the two-parameter IBP

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

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 nth 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.

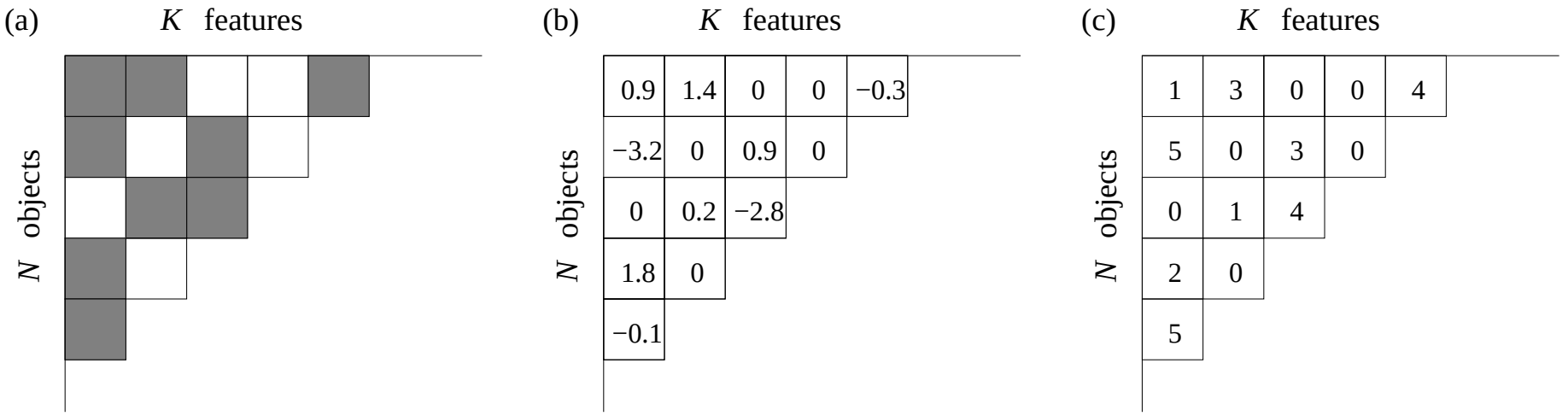
II. From binary to non-binary latent features

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

A simple way to generate non-binary latent feature matrices from $\mathbf{Z}$:

$$\mathbf{F} = \mathbf{Z} \otimes \mathbf{V}$$

where $\otimes$ is the elementwise (Hadamard) product of two matrices, and $\mathbf{V}$ is a matrix of independent random variables (e.g. Gaussian, Poisson, Discrete, ...).



III. Markov Indian buffet process and time series

Let the Z_{nk} have a *Markov structure*: e.g.

$$P(Z_{nk} = 1 | Z_{n-1,k} = 0) = \theta_{k,0,1}$$

Why? For time series data, we want latent factors to turn on and off in a manner that depends on time.

The Markov IBP (MIBP) defines such a process, which has IBP marginals.

(van Gael, Teh, Ghahramani, 2009)

More generally we can have the IBP be *dependent* on covariates (Williamson).

The MIBP can be used to generalise the hidden Markov model...

III. Markov Indian buffet process and time series

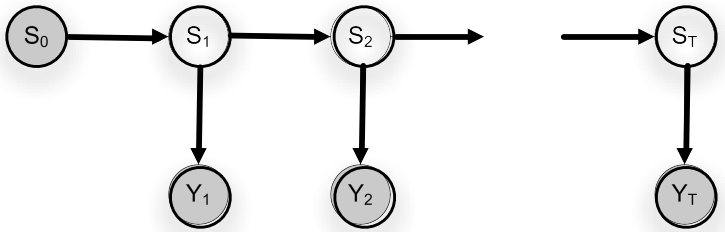


Figure 1: The Hidden Markov Model

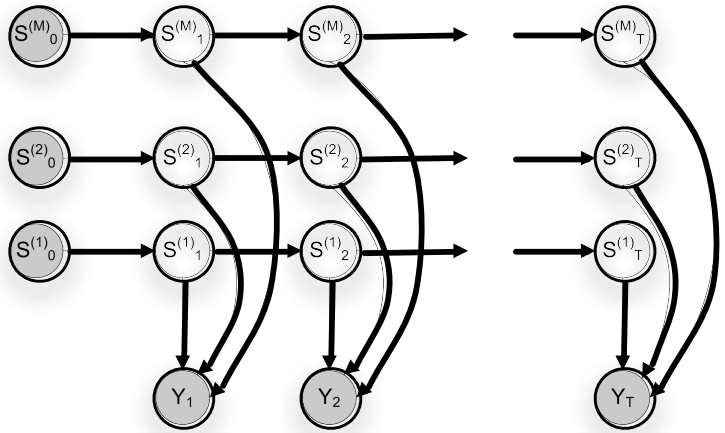


Figure 2: The Factorial Hidden Markov Model

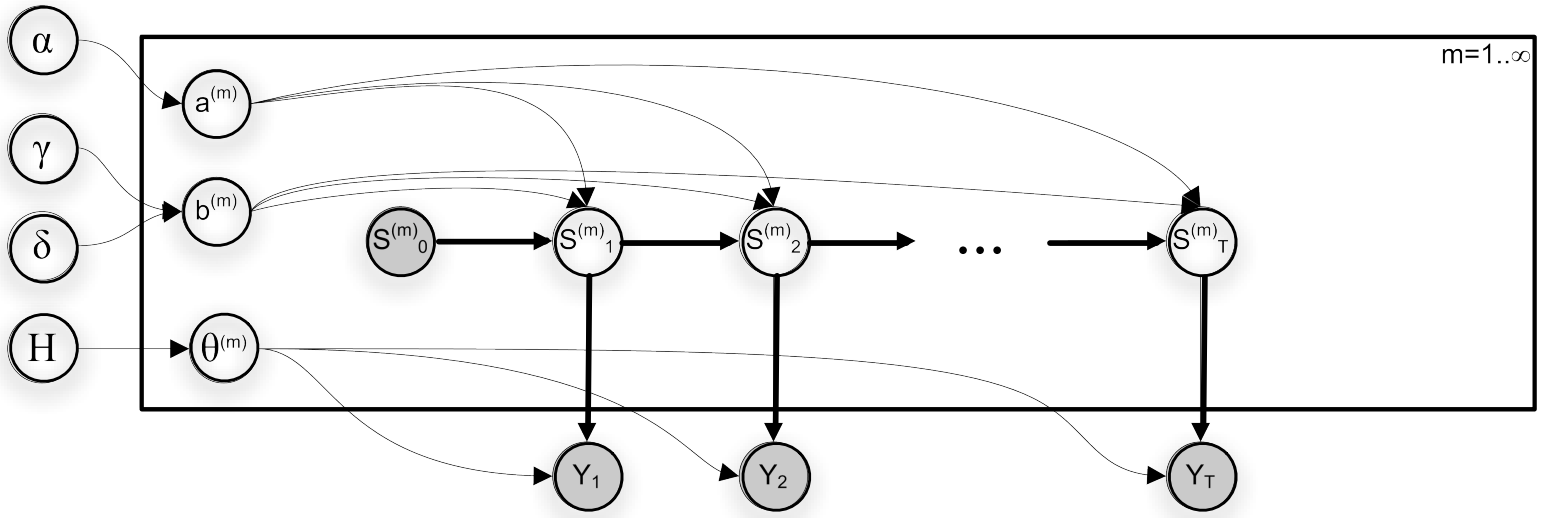
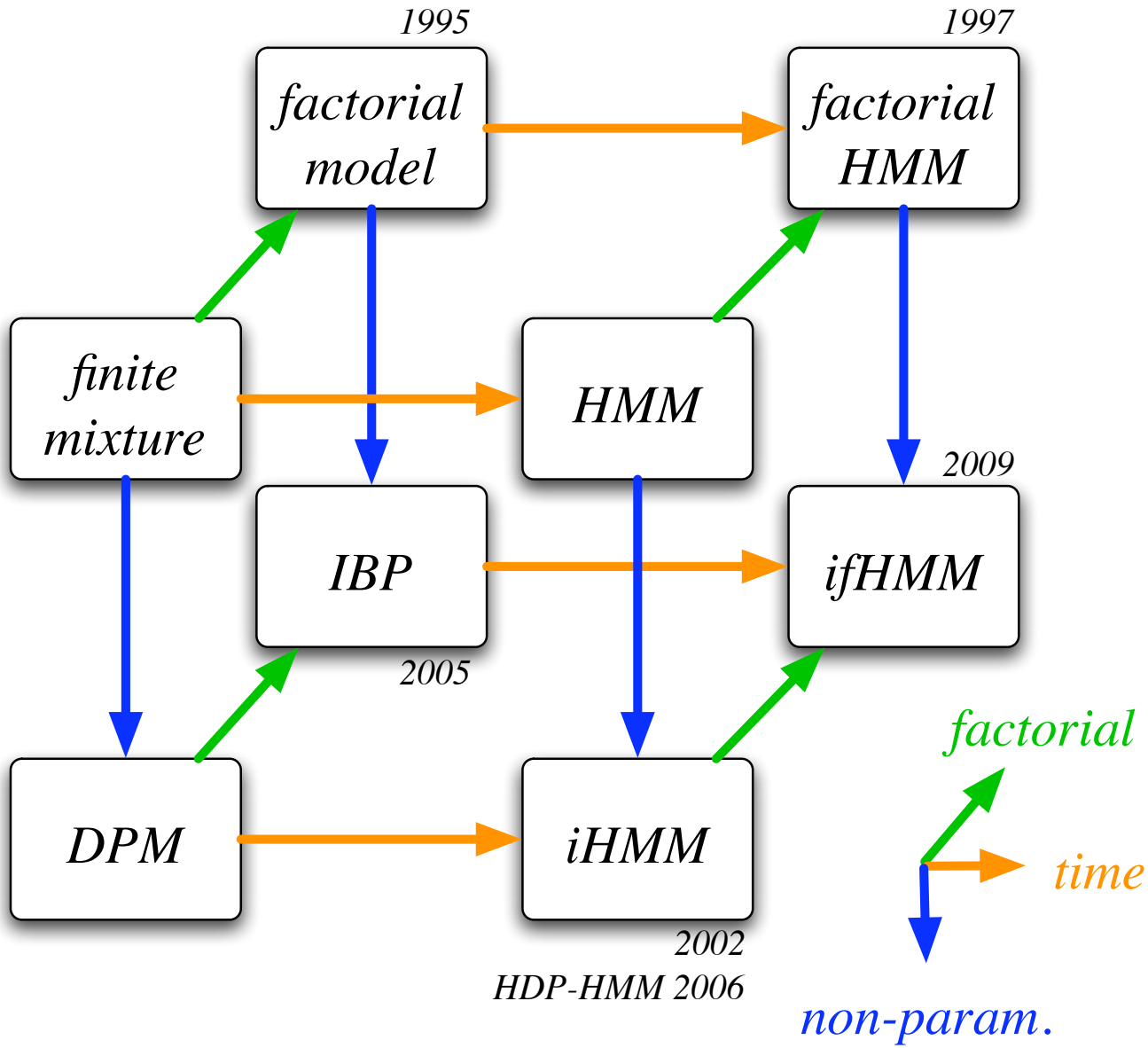


Figure 3: The Infinite Factorial Hidden Markov Model

The Big Picture



Summary

- A distribution on infinite sparse binary matrices that can be used to define many new non-parametric Bayesian models.

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