Bayesian non parametric approaches: an introduction

Pierre CHAINAIS

Bordeaux - nov. 2012

Trajectory

1. **Bayesian non parametric approaches (BNP)**
2. **How to handle an unknown number of clusters ?**
   - Latent class models : **infinite mixture models**
   - Dirichlet Processes
   - Chinese Restaurant Process
   - Segmentation (= clustering $\Rightarrow$ Dir. Proc.)
3. **How to handle an unknown number of features ?**
   - Latent feature models : **infinite sparse binary matrices**
   - Beta & Bernoulli Processes
   - Indian Buffet Process
   - Dictionary learning (= discovering features $\Rightarrow$ Beta Proc.)
4. **Conclusion & perspectives**

Trajectory

1. **Bayesian non parametric approaches (BNP)**
2. **How to handle an unknown number of clusters ?**
   - Latent class models : **infinite mixture models**
   - **Dirichlet Processes**
   - Chinese Restaurant Process
   - Segmentation (= clustering $\Rightarrow$ Dir. Proc.)
3. **How to handle an unknown number of features ?**
   - Latent feature models : **infinite sparse binary matrices**
   - **Beta & Bernoulli Processes**
   - Indian Buffet Process
   - Dictionary learning (= discovering features $\Rightarrow$ Beta Proc.)
4. **Conclusion & perspectives**

On the optimization side

Main problem :

ill posed inverse problems : unknown $\underbrace{\text{complexity}}_{\text{diversity}}$ / $\underbrace{\text{structure}}_{\text{content}}$

Main purpose :

- **to understand** data to propose a suitable model (structure)
- **to discover** the number of degrees of freedom (complexity)

Main tools to promote sparsity :

- a wide range of penalized optimization formulations (L^1...)
- model selection deals with discovering complexity

Main interest :

- efficient tools and algorithms in optimization,
- different approaches to promote (structured) sparsity
- theorems to control convergence properties...

Bayesian non parametric approaches

Main problem :

ill posed inverse problems : unknown $\underbrace{\text{complexity}}_{\text{diversity}}$ / $\underbrace{\text{structure}}_{\text{content}}$

Main purpose :

- **to understand** data to propose a suitable model (structure)
- **to discover** the number of degrees of freedom (complexity)

Main tools to promote sparsity :

- there exists a wide range of "classical" Bayesian models
- **flexible priors on infinite dimensional objects**

Main interest :

- **complexity of the model is adaptive**
- **some non parametric priors promote sparsity**
- **no need for model selection**

The Bayesian framework in brief

$\mathbf{Y}$ = data (observations), $\quad \boldsymbol{\theta}$ = parameters (model)

$$p(\mathbf{Y}, \boldsymbol{\theta}) = p(Y|\boldsymbol{\theta})p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathbf{Y})p(\mathbf{Y})$$

$$\Longrightarrow \quad \underset{\text{posterior}}{p(\boldsymbol{\theta}|\mathbf{Y})} \propto \underset{\text{likelihood}}{p(\mathbf{Y}|\boldsymbol{\theta})} \quad \underset{\text{prior}}{p(\boldsymbol{\theta})}$$

parameter relevance	goodness of fit	model constraints

e.g. $\boldsymbol{\theta} = \boldsymbol{\alpha}$, coeff. on dictionary $\mathbf{D}$

$$p(\boldsymbol{\alpha}|\mathbf{Y}, \mathbf{D}) \propto \underbrace{p(Y|\mathbf{D}, \boldsymbol{\alpha})}_{\text{noise distr.}} \; \underbrace{p(\boldsymbol{\alpha})}_{\text{model}}$$

Optimization : $\quad (\mathbf{X} + \text{Gaussian noise}) \quad + \quad \text{regularizing penalty}$

$$\log p(\boldsymbol{\alpha}|Y) = \quad \|Y - \underbrace{\mathbf{D}\boldsymbol{\alpha}}_{\hat{\mathbf{X}}}\|^2 \quad + \quad \underbrace{\lambda\|\boldsymbol{\alpha}\|_{L1}}_{\textit{Laplace}}$$

Optimization vs Bayesian tools

Optimization

- gradient descent
- proximal operators
- functional analysis
- L_0, L_1, $L_p^q \Rightarrow$ **sparsity**

$\vdots$

Bayesian world

- Maximum Likelihood,
- Maximum A Posteriori (MAP)
- Gibbs sampling, MCMC,
- EM algorithm (hidden variables...)
- prior promoting **sparsity** :
  1. 'heavy tailed' priors (Laplace, Student, Bessel-K...)
  2. **non parametric approaches**

Remark : conjugate priors (w.r.t. likelihood) $\Rightarrow$ easier inference
(importance of the exponential family of distributions)

Bayesian non parametric in Machine Learning
Document classification

Typical application : **unsupervised classification of documents**
unknown number of categories/subcategories, e.g. NIPS sections

1 document $\in$ 1 class (unique)

- $G_0 =$ **multinomial distribution** of words in language
- **Category** $j =$ typical distribution of words β_k^j, $k \in \mathbb{N}$

  prior on $\boldsymbol{\beta} =$ **Dirichlet process** (G_0)
- **Sub-category** $j, \ell =$ typical distribution of words $\pi_k^{j,\ell}$, $k \in \mathbb{N}$

  prior on $\boldsymbol{\pi}^{j,\ell} =$ **Dirichlet process** ($\boldsymbol{\beta}^j$)
- ...

[Teh, Jordan, Beal, Blei'06 ; Teh, Jordan'09]

Typical application : **association between movies and viewers**
(unknown number of features characterizing movies / viewers)

Observations : ratings of movies by viewers

predict ratings $\Rightarrow$ collaborative filtering

1 viewer = several features, 1 movie = several features

- Movies **binary matrix** : "horror", "comedy", "3D"...
- Viewers **binary matrix** : "likes horror", "likes 3D"...
- Weight matrix : links viewers features to movies features

$\Longrightarrow$ **factorization matrix problem : R = V W M**

Beta & Bernoulli Processes

[Meeds et al.'07]

Latent class models : from finite to infinite mixtures

Finite mixtures of distributions

e.g. mixture of Gaussians : $p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}| \underbrace{\mu_k, \Sigma_k}_{\theta_k})$

- π_k, $1 \leq k \leq K$ = multinonomial dist.
  $= $ proportions of the mixture

  prior on $\boldsymbol{\pi}$ = conj. distr. = **Dirichlet distribution (α)**

$$\mathbf{p}(\boldsymbol{\pi}) \propto \prod_{\mathbf{k}=\mathbf{1}}^{\mathbf{K}} \pi_{\mathbf{k}}^{\alpha_{\mathbf{k}}-\mathbf{1}}, \quad \text{typically } \alpha_k = \alpha/K$$

- $\sum_k \pi_k = 1$
- θ_k : **prior** G_0, e.g. Normal-Wishart on (μ_k, Σ_k)
- Inference : EM algorithm thanks to hidden z_k

Latent class models : from infinite to finite mixtures

Infinite mixtures : Dirichlet process

e.g. mixture of Gaussians : $p(\mathbf{x}) = \sum_{k=1}^{\infty} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$

- $\boldsymbol{\pi} = \{\pi_k, k \in \mathbb{N}\}$ **infinite combinatorial distribution**
  prior on $\boldsymbol{\pi}$ = **Dirichlet process** = produces random distr.

$$\mathbf{G} = \sum_{\mathbf{k=1}}^{\infty} \boldsymbol{\pi_k} \boldsymbol{\delta_{\theta_k}}$$

  e.g. $\theta_k = (\mu_k, \Sigma_k)$, where $\theta_k \sim G_0$

  where G_0 = base distribution $\simeq$ prior on parameters θ_k,
  e.g. Normal-Wishart on (μ_k, Σ_k).

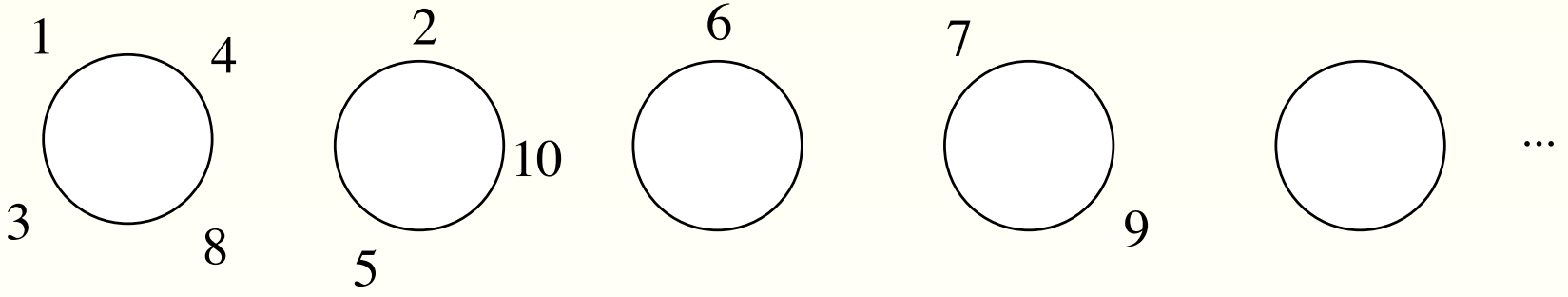
- Gibbs sampler, MCMC, EM algorithm (truncated DP)

[Ferguson'73]

DP and the Chinese Restaurant Process

CRP = Gibbs sampling of the DP posterior

- DP is a clustering prior on the infinite $(\pi_k, \theta_k)_{k \in \mathbb{N}}$ (cf. stick-breaking)



- the **Chinese Restaurant Process** :
  - Objects are customers, classes are tables,
  - Customer n chooses table k with probability :

$$P(z_n = k | z_1, ..., z_{n-1}) = \begin{cases} \dfrac{m_k}{\alpha + n - 1} & k \leq K_+, \\ \dfrac{\alpha}{\alpha + n - 1} & k = K + 1. \end{cases}$$

- m_k = number of customers at table k,
- K_+ = number of classes s.t. $m_k > 0$.

An illustration : mixture of Gaussians

1. K **is known :** $p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k, \Sigma_k)$
   EM algorithm by defining hidden variables z ($x \in \mathcal{C}_k$)...
2. K **is unknown :**
   1. try various K, keep the best (model selection...)
   2. use a 'nice' prior on $\{\pi_k,\ (\mu_k, \Sigma_k),\ k \in \mathbb{N}\}$ (assume $K = \infty$!)

Clustering prior with infinite number of components :

Dirichlet Process

Various inference algorithms :

- Gibbs sampling : **the Chinese Restaurant Process**,
- Variational bayesian inference,
- EM algorithm [Kimura et al.'11]

Remark : Image processing : segmentation, classification...

An illustration : mixture of Gaussians

1. K **is known :** $p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k, \Sigma_k)$
   EM algorithm by defining hidden variables z ($x \in \mathcal{C}_k$)...
2. K **is unknown :**
   1. try various K, keep the best (model selection...)
   2. **use a prior on** $(\pi_k)_{k \in \mathbb{N}}$ **(assume** $K = \infty$ **!)**

Clustering prior with infinite number of components :

Dirichlet Process

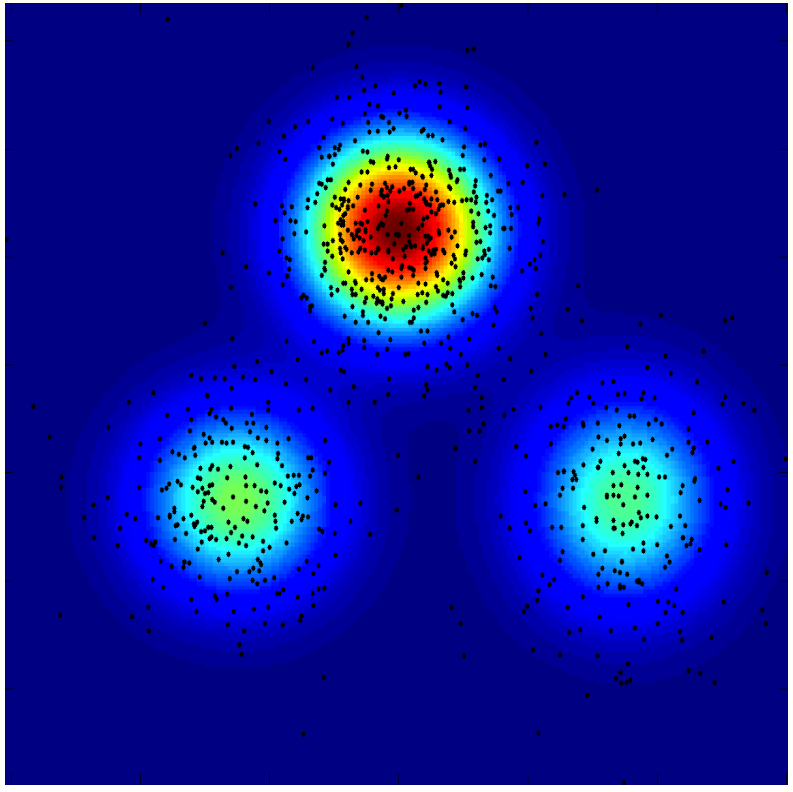
Various inference algorithms :

- Gibbs sampling : **the Chinese Restaurant Process**,
- Variational bayesian inference,
- EM algorithm [Kimura et al.'11]

Remark : Image processing : segmentation, classification...

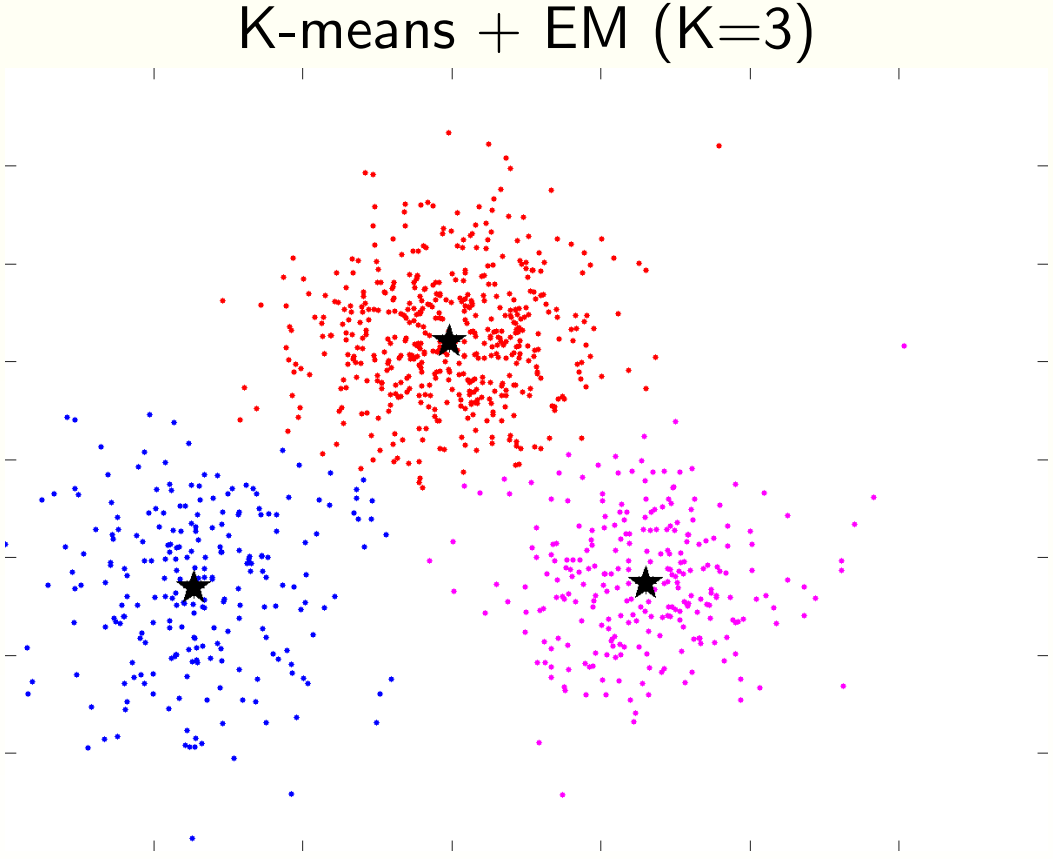
An illustration : mixture of Gaussians

Inference using EM algorithm [Kimura et al'11]



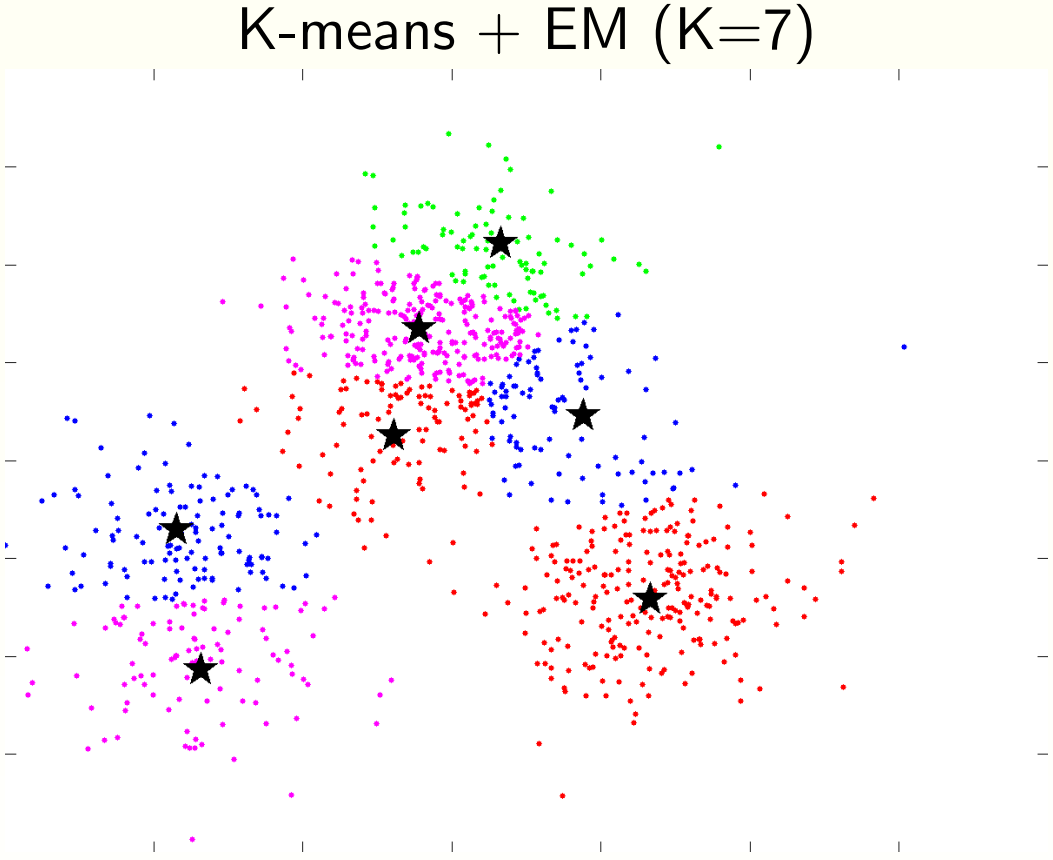
An illustration : mixture of Gaussians

Inference using EM algorithm [Kimura et al'11]



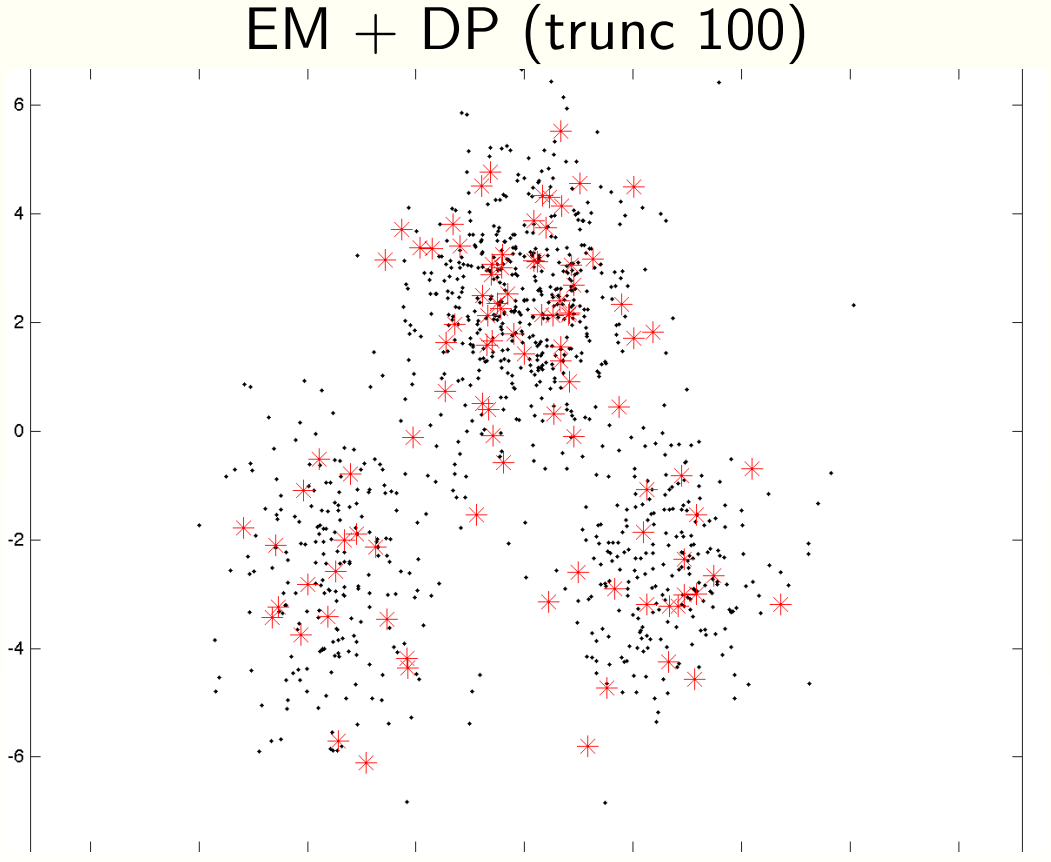
An illustration : mixture of Gaussians

Inference using EM algorithm [Kimura et al'11]



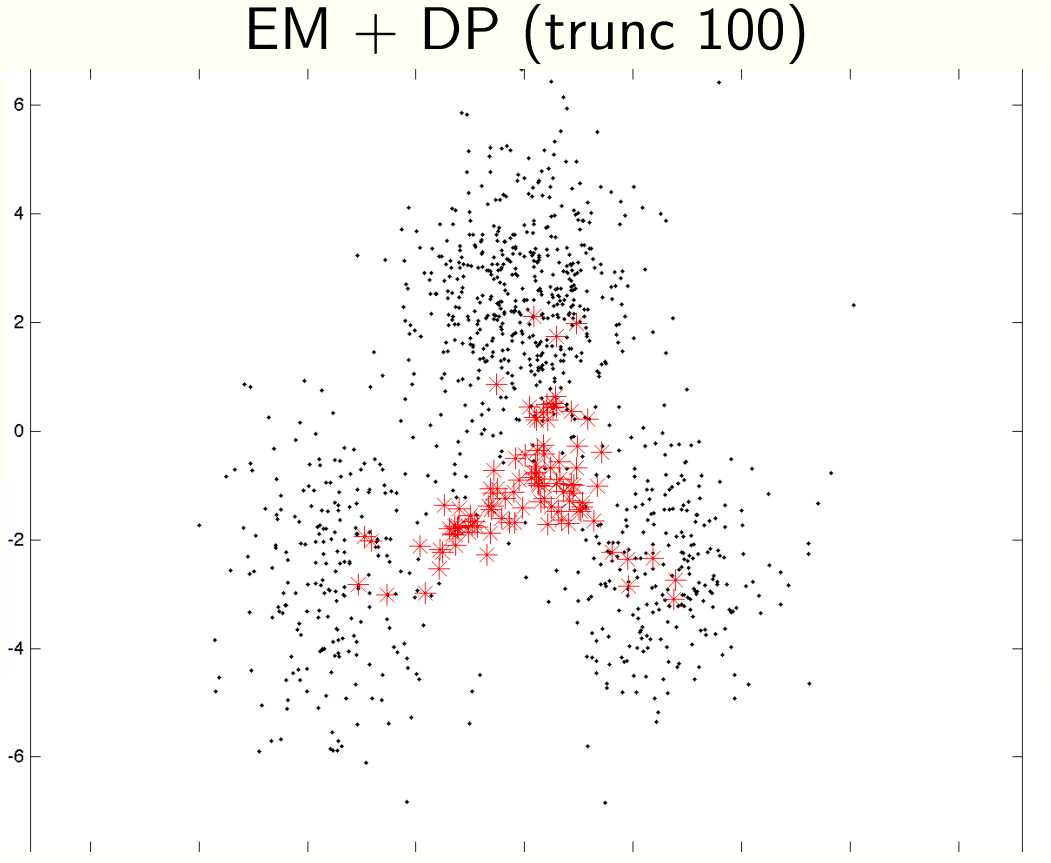
An illustration : mixture of Gaussians

Inference using EM algorithm [Kimura et al'11]



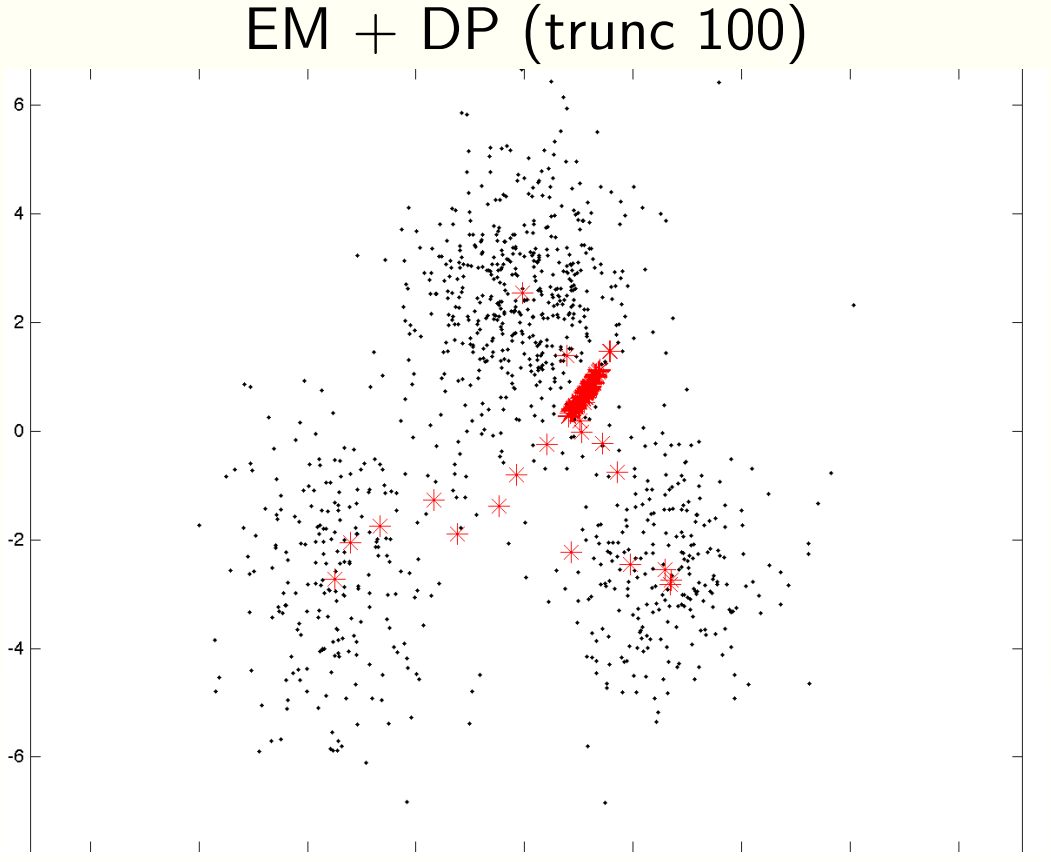
An illustration : mixture of Gaussians

Inference using EM algorithm [Kimura et al'11]



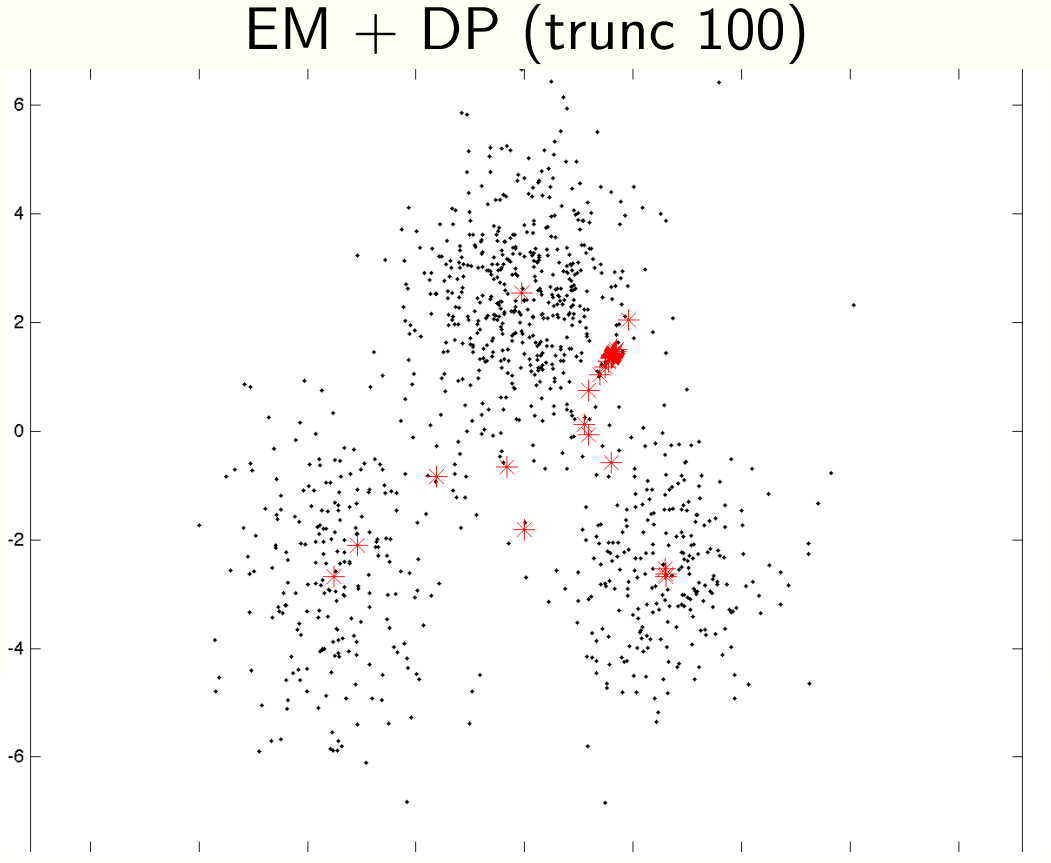
An illustration : mixture of Gaussians

Inference using EM algorithm [Kimura et al'11]



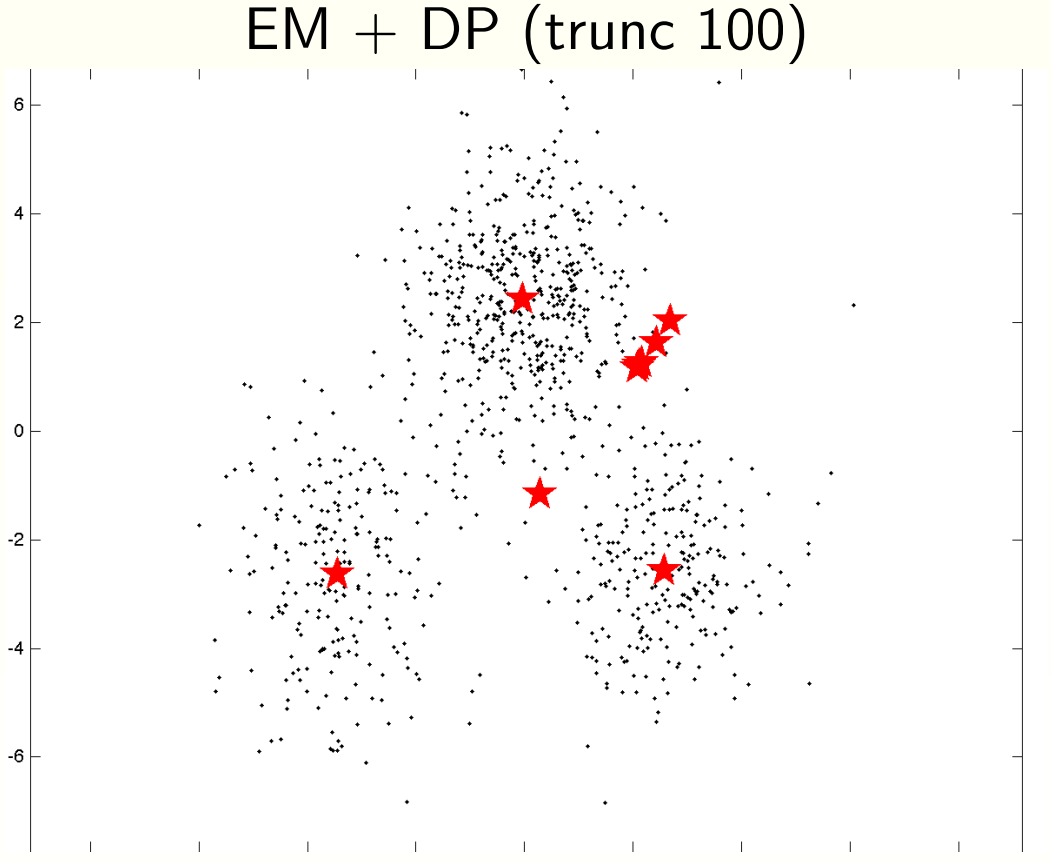
An illustration : mixture of Gaussians

Inference using EM algorithm [Kimura et al'11]



An illustration : mixture of Gaussians

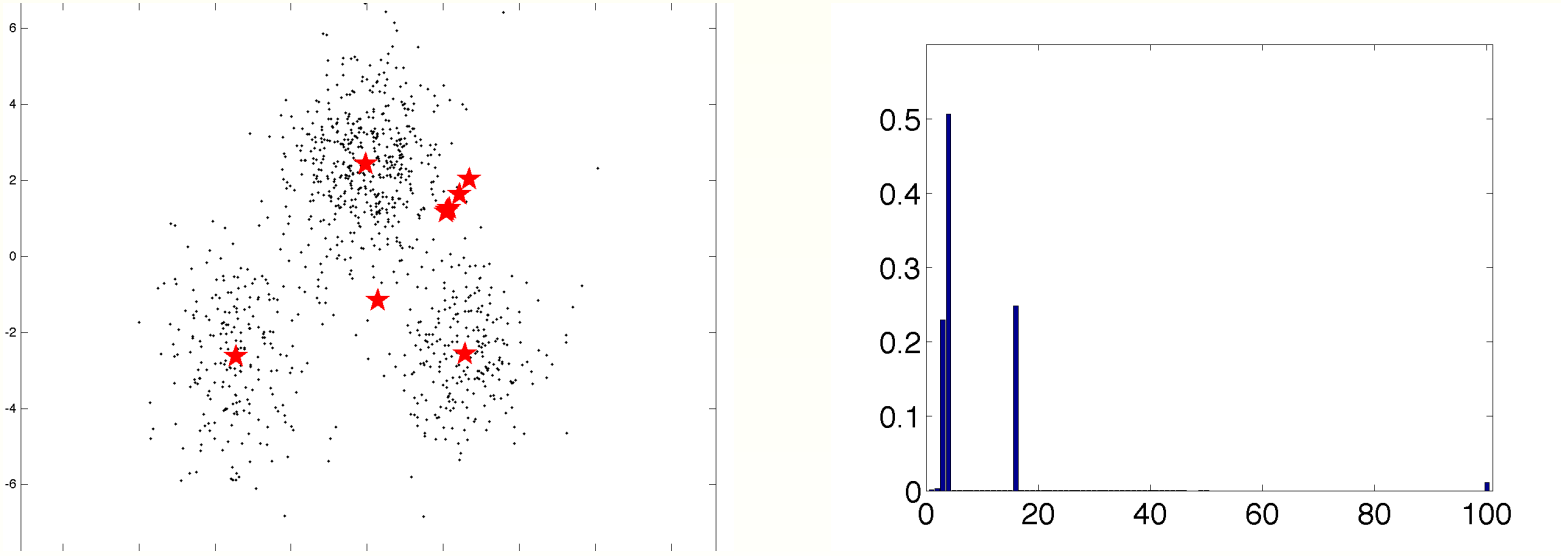
Inference using EM algorithm [Kimura et al'11]



An illustration : mixture of Gaussians

Inference using EM algorithm [Kimura et al'11]

EM + DP (trunc 100)



Expected proportions were [0.23 ; 0.50 ; 0.27]

Image processing : shared segmentation of natural scenes

Pb : how to obtain an unsupervised segmentation of images ?

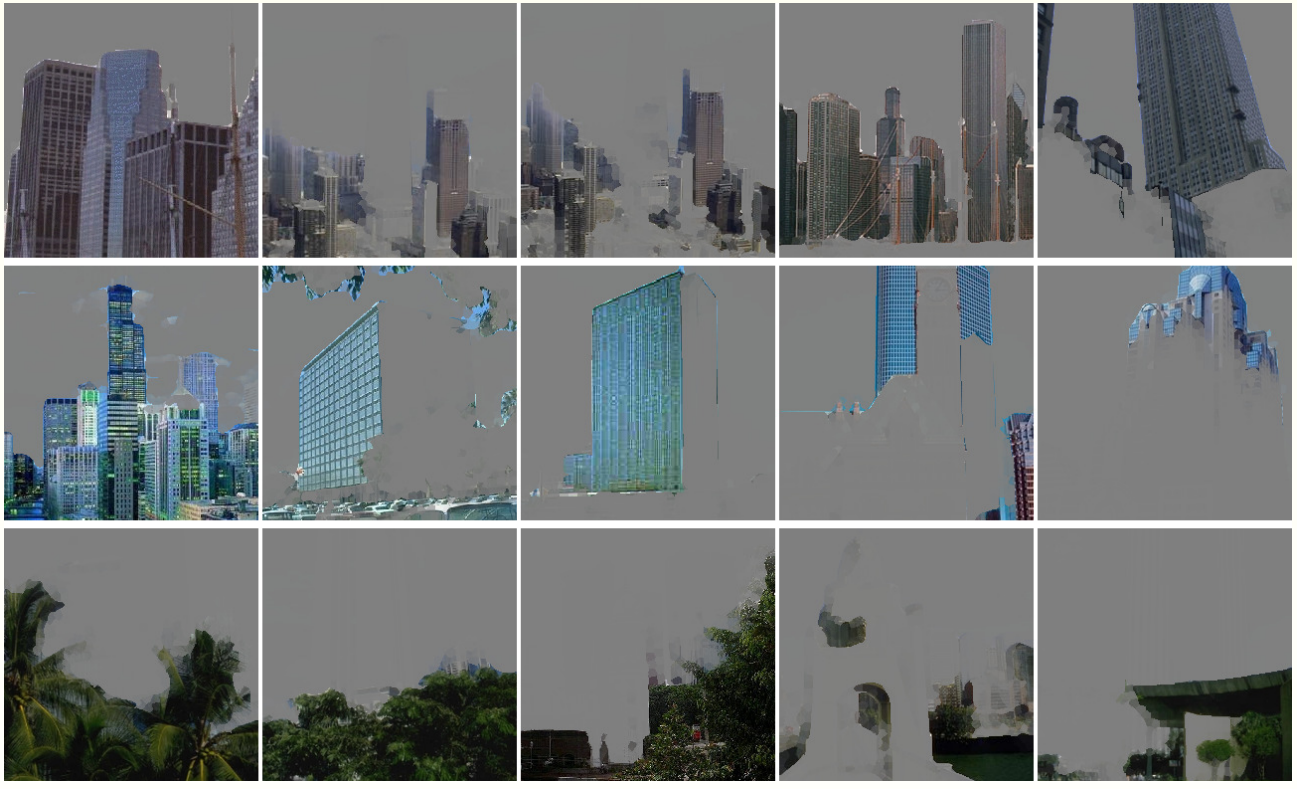
(by simultaneously segmenting a set of images)

Main ideas :

- objects = typical textures + colors + frequency + area
- use some first **over-segmentation** (super-pixels)
- features = texton & color histograms $\Longrightarrow$ DP prior
- each object category k occurs with frequency $\varphi_k \Rightarrow$ PY prior
- each object t = random proportion π_{jt} of image $j \Rightarrow$ PY prior
- PY = Pitman-Yor process, a generalized DP (2 param)
- spatial dependencies : thresholded Gaussian processes (another story...)

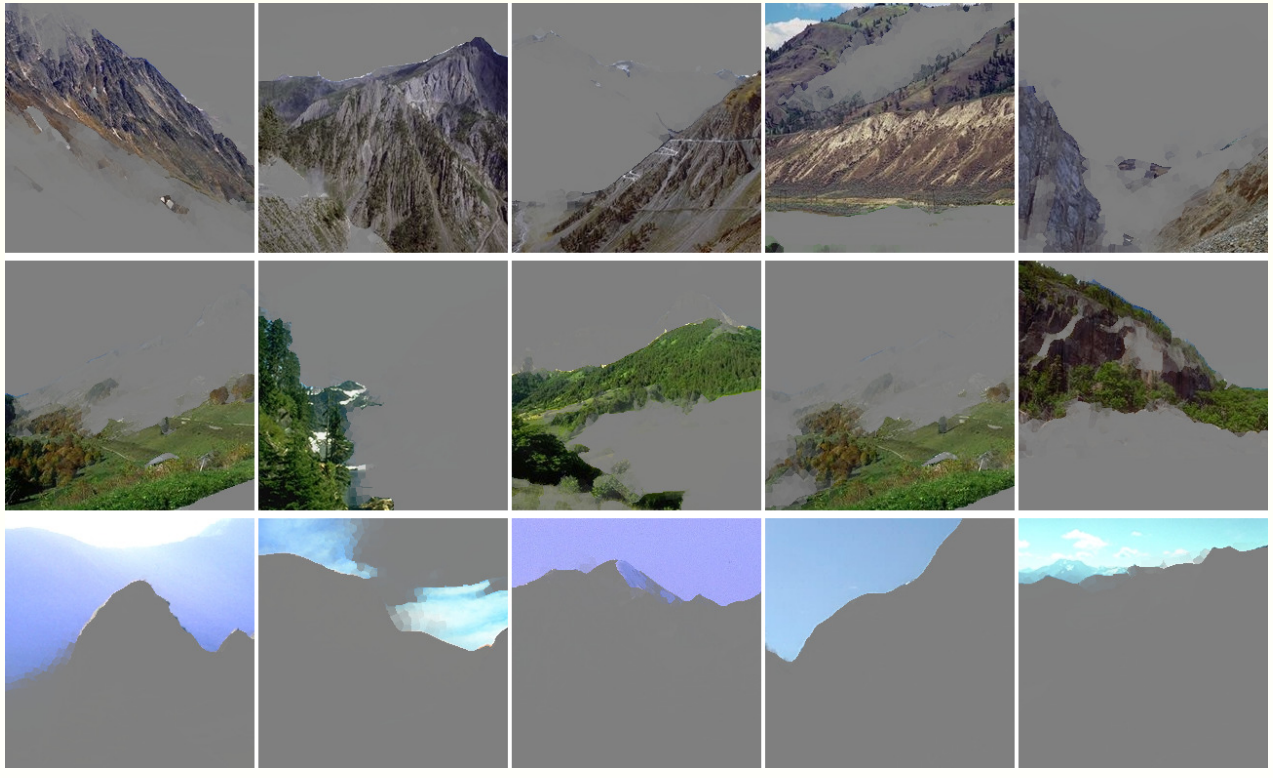
[Sudderth & Jordan'08]

Image processing : shared segmentation of natural scenes



[Sudderth & Jordan'08]

Image processing : shared segmentation of natural scenes



[Sudderth & Jordan'08]

Take home message : BNP is rich and adaptive

1. there exists **infinite latent class models**
2. **Dirichlet processes** (and their generalization) ...
3. ... are adaptive and permits the **discovery of classes**
4. learning **DP mixtures** : non Gaussian distributions
   [Kivinen, Sudderth & Jordan 2007]
5. too simple ? go for **hierarchical** but...
6. **inference** : go to the **Chinese Restaurant** !
   (MCMC, split & merge, variational inference...)

Latent feature models : sparse binary matrices

- latent class models : 1 object $\leftrightarrow$ 1 class
- latent feature models : **1 object $\leftrightarrow$ several features**

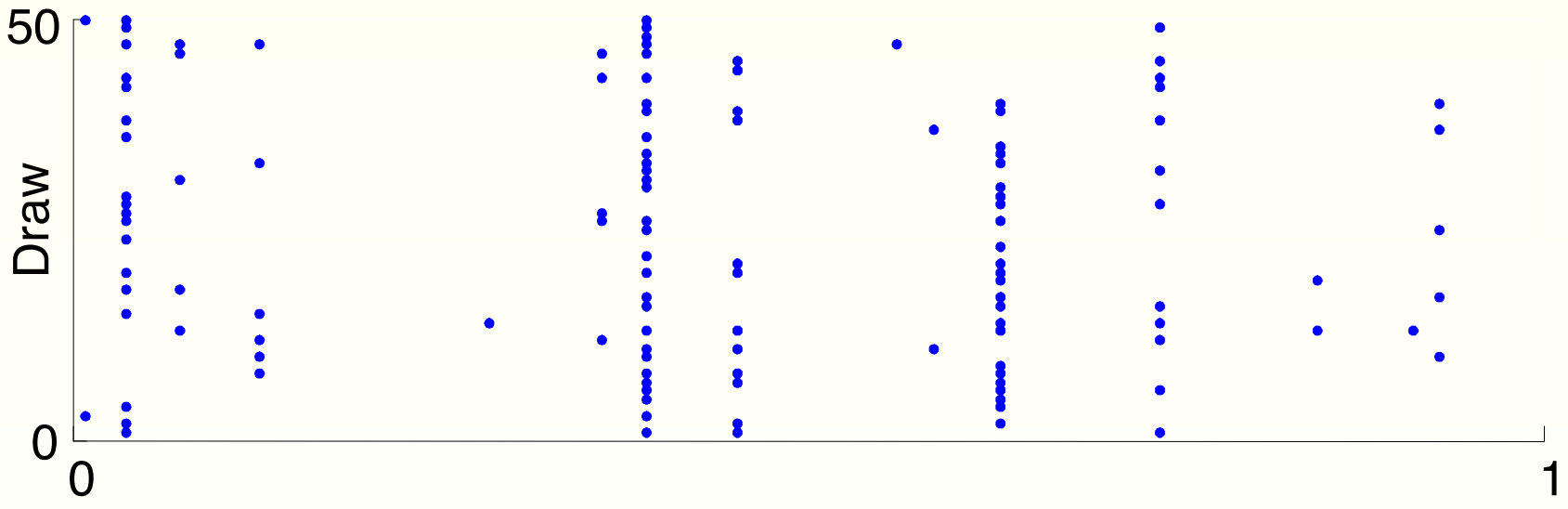
Main ideas :

- build (infinite) **binary matrices** :
  $z_{nk} = 1$ if object n possesses feature k
- **with probability** π_k (but $\sum_k \pi_k \neq 1$)
- each π_k follows a **Beta distribution** $p(\pi_k) \propto \pi_k^{r-1}(1-\pi_k)^{s-1}$
  conjugate to the binomial
- In summary, the probability model is :

$$\pi_k|\alpha \sim Beta(\tfrac{\alpha}{K}, 1) \qquad K \to \infty \,?$$
$$z_{nk}|\pi_k \sim Bernoulli(\pi_k)$$

[Griffiths & Ghahramani'11]

Latent feature models : sparse binary matrices



1. **Beta & Bernoulli processes**
2. **inference : the Indian Buffet Process**

Beta process (BP)

A *beta process* (BP) $B \sim BP(c, B_0)$ on Ω is a positive Lévy process whose Lévy measure depends on :

- the *concentration function* c,
- the *base measure* B_0, a fixed measure on Ω.
- $\gamma = B_0(\Omega) =$ *mass parameter*

$$\nu(d\omega, d\pi) = c(\omega) \underbrace{\pi^{-1}(1-\pi)^{c(\omega)-1} d\pi}_{\text{beta distribution}} B_0(d\omega) \quad \text{on } \Omega \times [0,1]$$

- Draw of a BP :
  Poisson point process (ω_k, π_k) with base measure ν,

$$\mathbf{B} = \sum_{\mathbf{k=1}}^{\infty} \boldsymbol{\pi_k} \boldsymbol{\delta_{\omega_k}} \quad \text{discrete}$$

[Hjort'90, Thibaux & Jordan'07]

Beta process (BP)

A *beta process* (BP) $B \sim BP(c, B_0)$ on Ω is a positive Lévy process whose Lévy measure depends on :

- the *concentration function* c,
- the *base measure* B_0, a fixed measure on Ω.
- $\gamma = B_0(\Omega) =$ *mass parameter*

$$\nu(d\omega, d\pi) = c(\omega) \underbrace{\pi^{-1}(1-\pi)^{c(\omega)-1} d\pi}_{\text{beta distribution}} B_0(d\omega) \quad \text{on } \Omega \times [0,1]$$

- Draw of a BP :
  Poisson point process (ω_k, π_k) with base measure ν,

$$\mathbf{B} = \sum_{\mathbf{k=1}}^{\infty} \boldsymbol{\pi_k} \boldsymbol{\delta_{\omega_k}} \quad \text{discrete}$$

[Hjort'90, Thibaux & Jordan'07]

Bernoulli process

- Ω = potential set of features, ($B_0 \sim$ prior on features)
- random measure B = **proba that X possesses feature ω_k**
- **binary matrix = Bernoulli process from the Beta process**

$$\mathbf{z_{nk} \sim Bernoulli(\pi_k)}$$

for data $\mathbf{X}$, $\sum_k \pi_k \neq 1$,

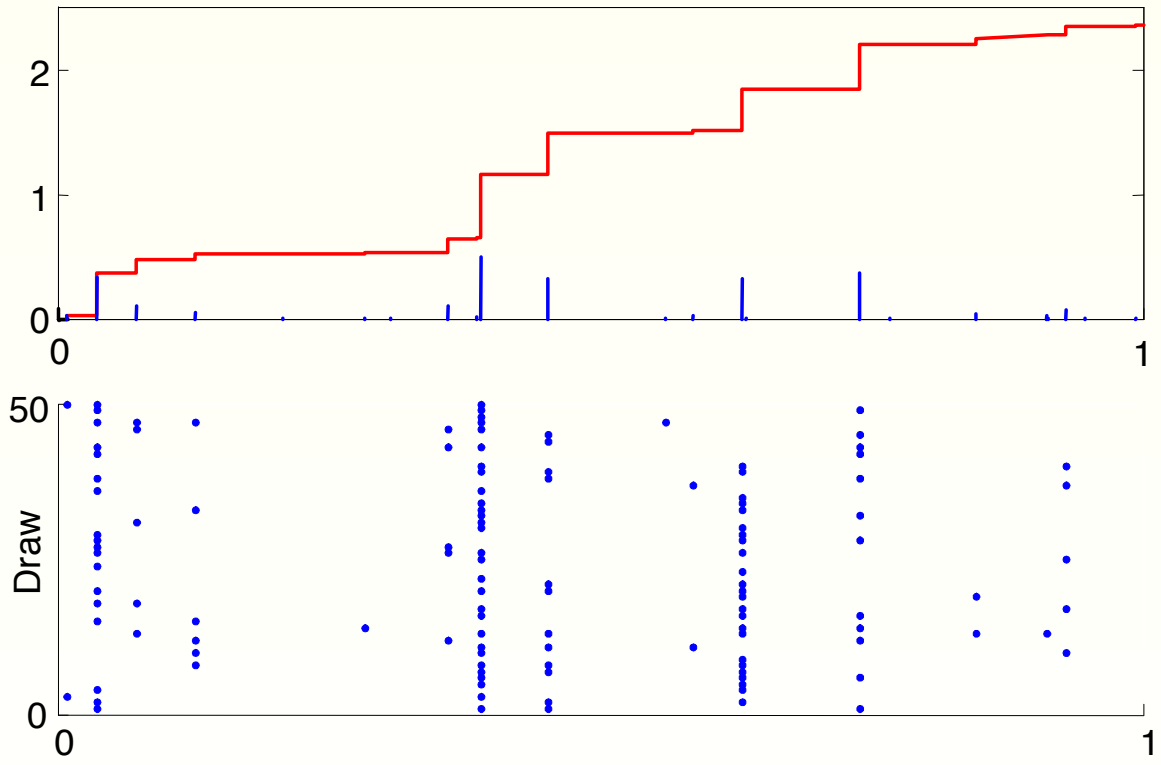
$$X_{nk}|B \sim BeP(B)$$

- **inference $\Rightarrow$ posterior of Bernoulli = Beta process**

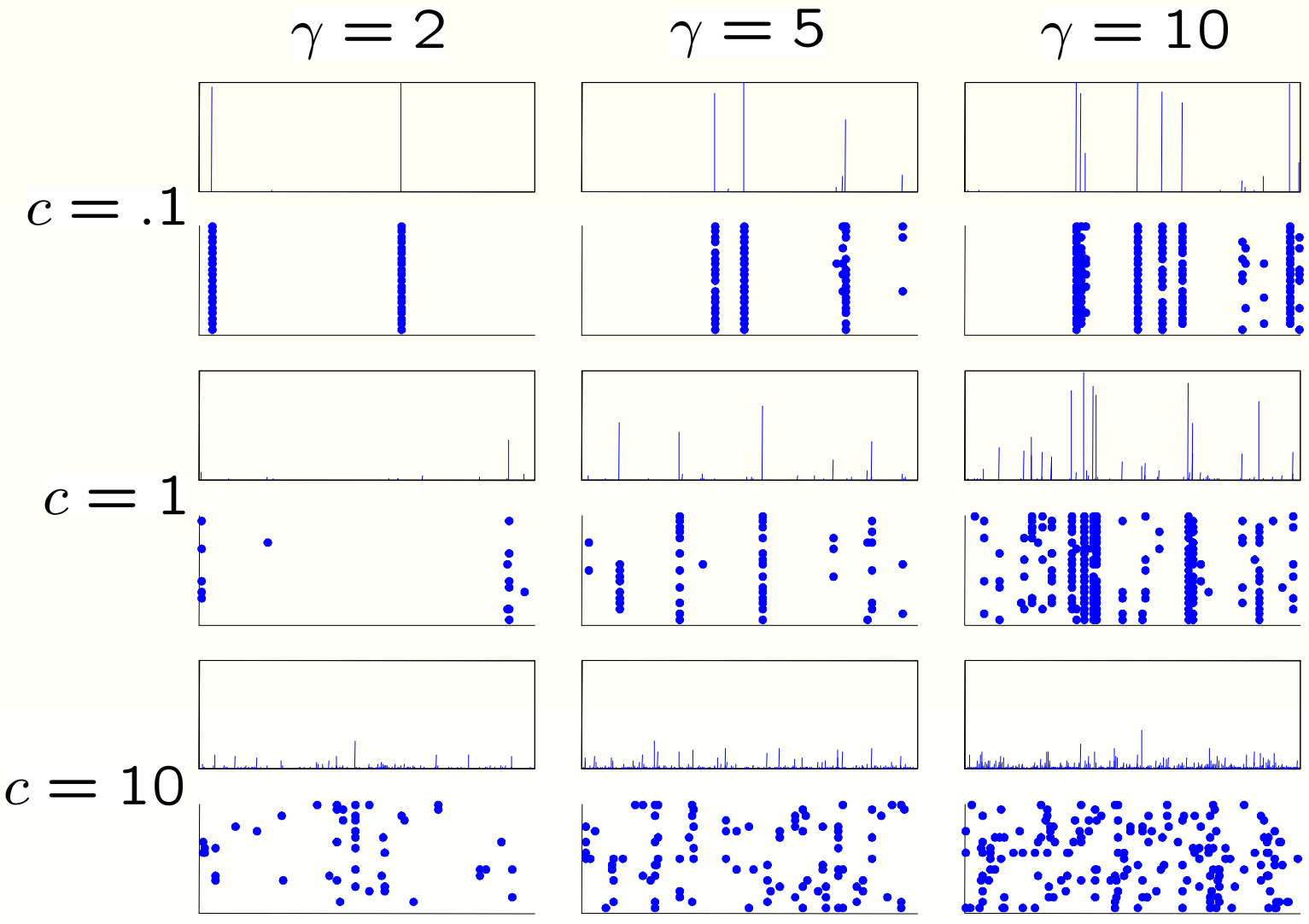
$$B|X_{1,\dots,n} \sim BP\left(c+n, \frac{c}{c+n}B_0 + \frac{1}{c+n}\sum_{i=1}^{n} X_i\right)$$

$$\Longrightarrow \textbf{the Indian Buffet Process}$$

Bernoulli process



Bernoulli process



various sparsity degrees

Inference : the Indian Buffet Process (IBP)

Pb : Beta & Bernoulli processes for simple inference ?

$\Longrightarrow$ **the Indian Buffet Process**

- X_n is a customer choosing dish k with proba π_{nk}

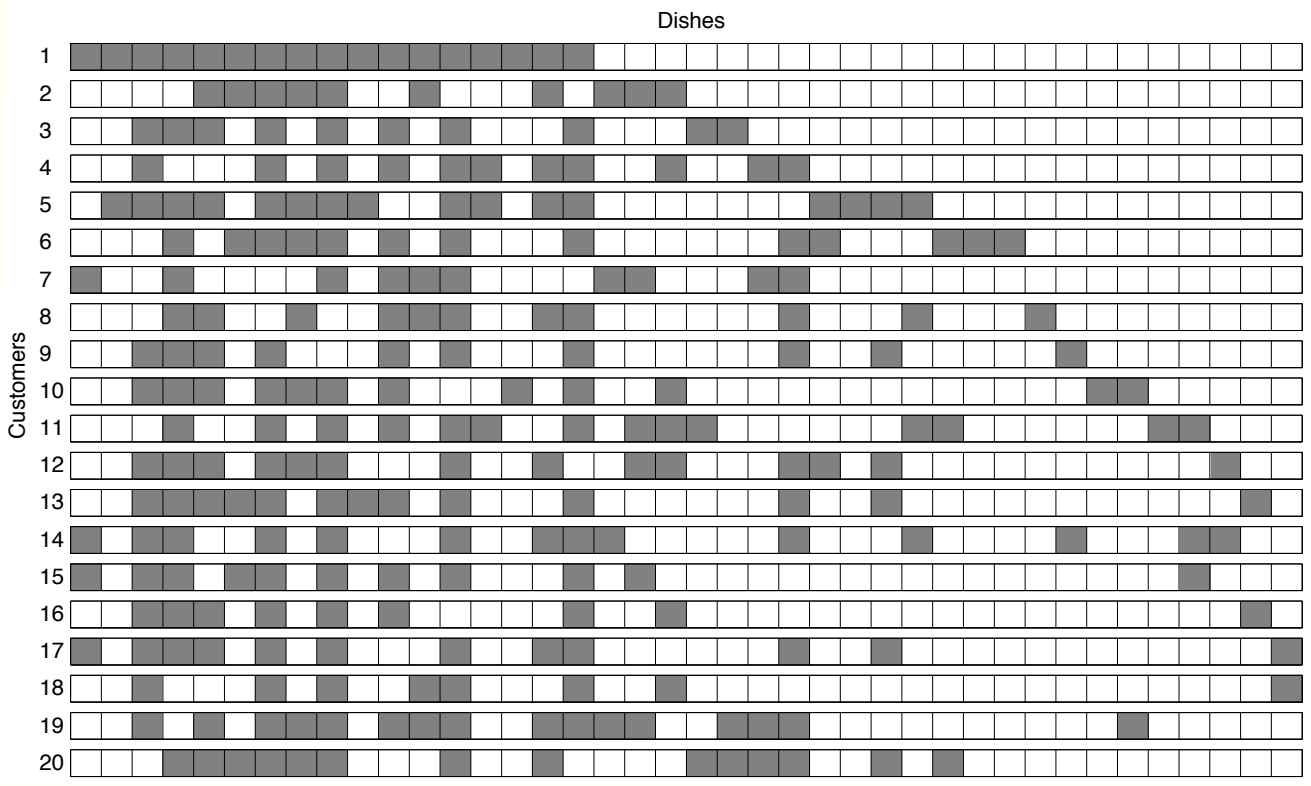
The story is a sequential mechanism
(but variables are *exchangeable* !)

1. 1st customer chooses Poisson(γ) dishes,
2. the nth customer
   - chooses dish k with proba $\dfrac{m_k}{n}$, $\quad m_k = \sum_{\ell<n} z_{\ell k}$
   - tries Poisson($\frac{\gamma}{n}$) new dishes
3. store choices in binary matrix $\mathbf{Z} = (z_{nk})$

- posterior takes likelihood into account

[Griffiths & Ghahramani'11] (worked in London...and had Indian food !)

Inference : the Indian Buffet Process (IBP)



Inference : the Indian Buffet Process (IBP)

For inference : use the exchangeable property of IBP

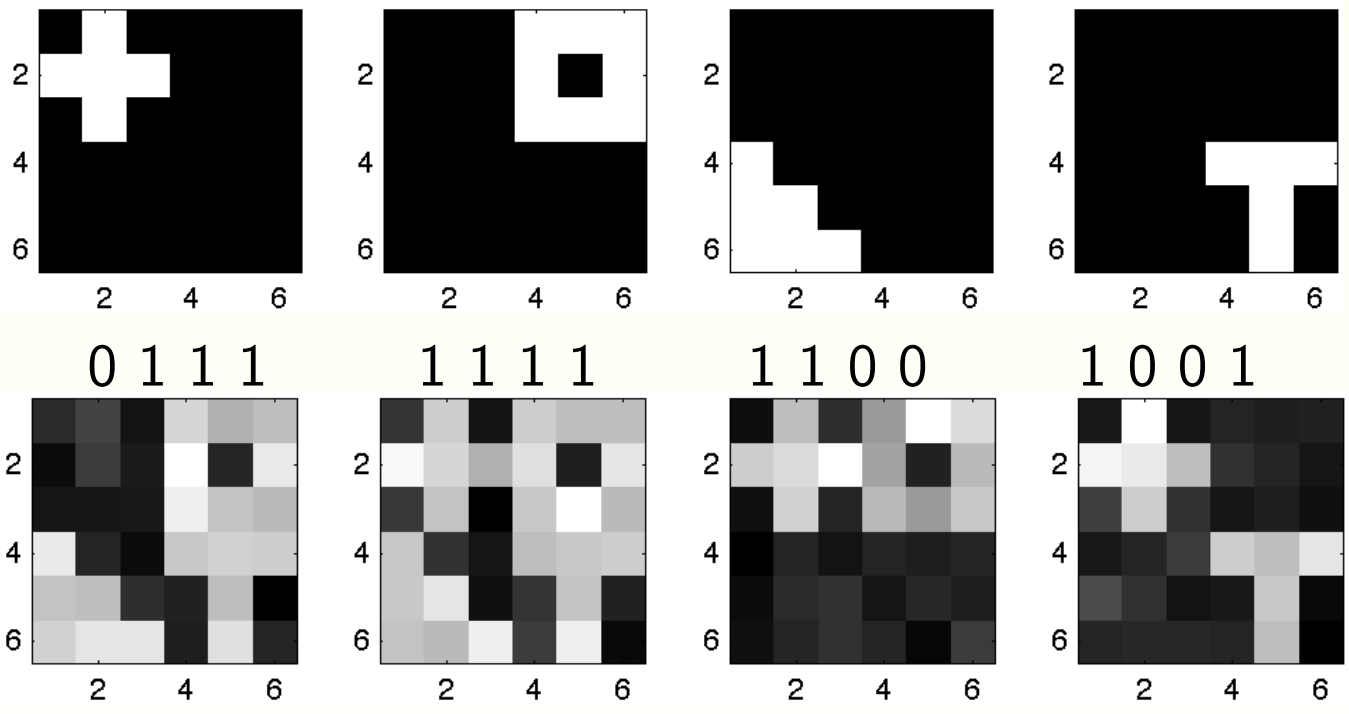
Key relation :

$$P(z_{nk} = 1 | \mathbf{Z}_{-(nk)}, \mathbf{X}) \propto \underbrace{p(\mathbf{X}|\mathbf{Z})}_{\text{likelihood}} \underbrace{P(z_{nk} = 1 | \mathbf{Z}_{-(nk)})}_{\text{Indian Buffet Proc.}}$$

Application : random binary images + noise

random images : binary combination of elements $\mathbf{X} = \mathbf{Dz} + \mathbf{n}$

unknowns : D, z, σ (noise)

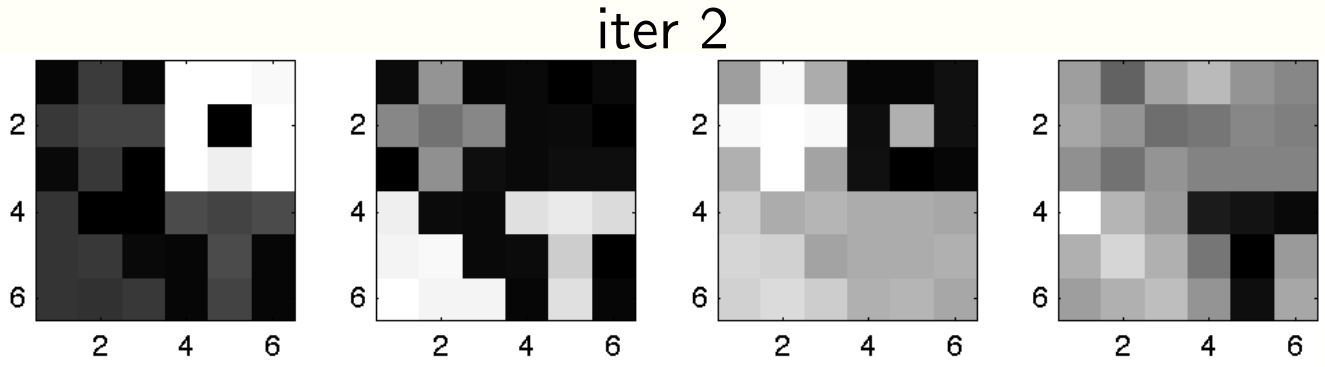


[Griffiths & Ghahramani'11]

Application : random binary images + noise

Gibbs sampling inference using the IBP,

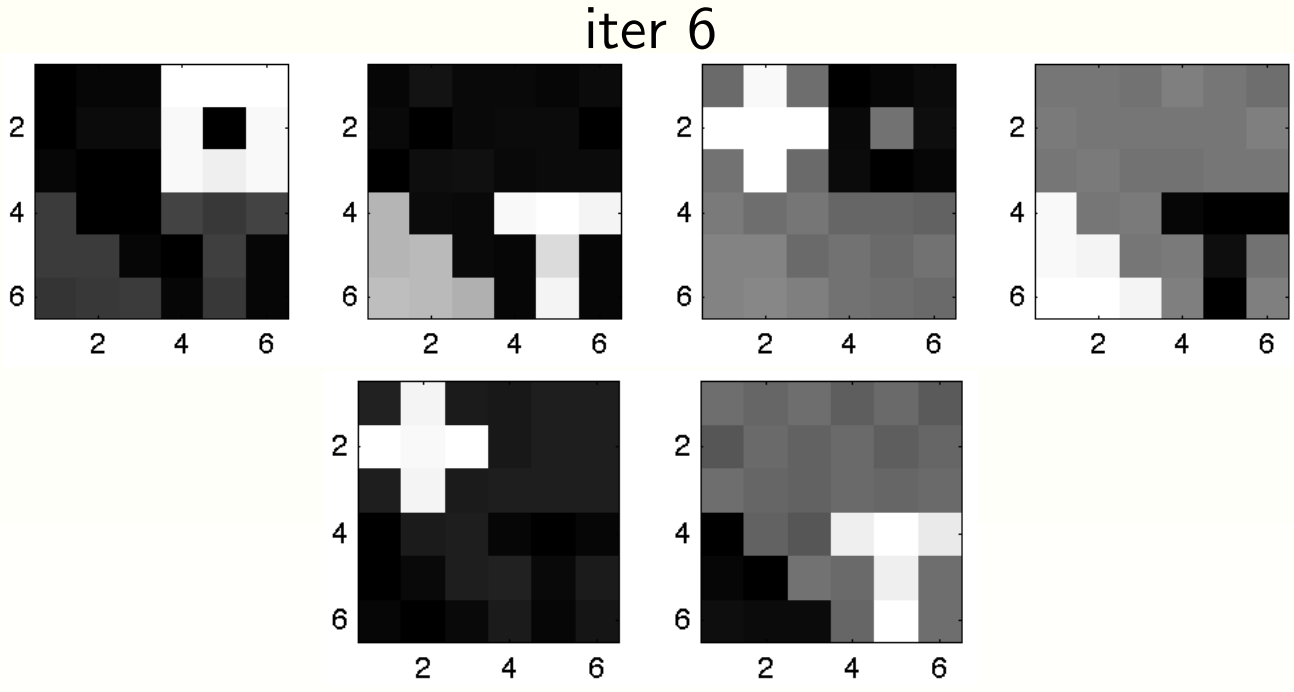
$\mathbf{X} = \mathbf{Dz} + \mathbf{n}$ **unknowns : D, z,** σ (noise)



Application : random binary images + noise

Gibbs sampling inference using the IBP,

$\mathbf{X} = \mathbf{D}\mathbf{z} + \mathbf{n}$ **unknowns : D, z,** σ (noise)

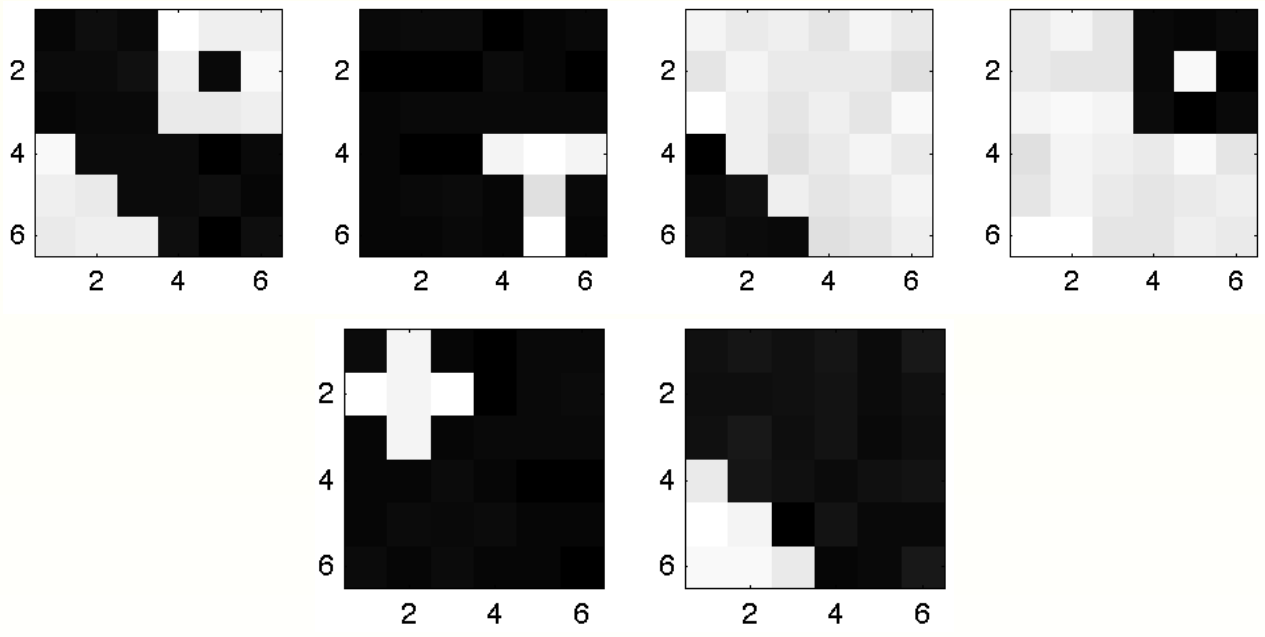


Application : random binary images + noise

Gibbs sampling inference using the IBP,

$\mathbf{X} = \mathbf{D}\mathbf{z} + \mathbf{n}$ **unknowns : D, z,** σ (noise)

iter 15 for initial $\sigma = 0.35$ (true value is $\sigma = 0.15$)



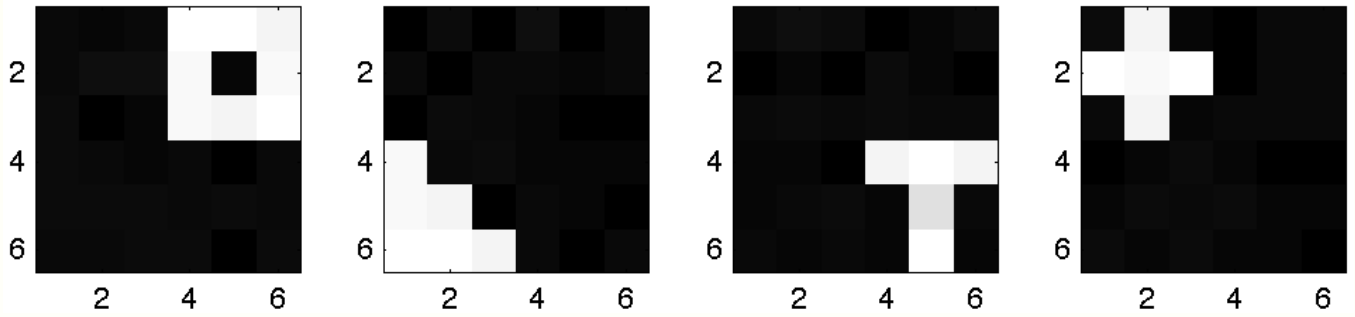
estimated $\boldsymbol{\sigma_{\mathbf{X}} = 0.15}$

Application : random binary images + noise

Gibbs sampling inference using the IBP,

$\mathbf{X} = \mathbf{D}\mathbf{z} + \mathbf{n}$ **unknowns : D, z,** σ (noise)

iter 2 for initial $\sigma = 0.25$ (true value is $\sigma = 0.15$)



estimated $\boldsymbol{\sigma_{\mathbf{X}} = 0.15}$

perfect reconstruction of all 200 images : **D, z**

Take home message : Bayesian Non Param. sparse latent feature models

1. there exists **infinite latent feature models**
2. **Beta & Bernoulli processes** (and their generalization) ...
3. ... are adaptive and permits the **discovery of features**
4. **source separation**, **dictionary learning**...
5. **estimates the unknown noise level**
6. **inference** : **go to the Indian Buffet** !
7. too simple ? go for **hierarchical** but...

Bayesian non parametric dictionary learning

[Zhou, Carin, Paisley et al.'11'12]

Main ingredients :

- a feature is used or not : binary $\mathbf{z}_i$

  Indian Buffet Process

- patch clustering : Dirichlet process
  similar patches use similar features
  (they go to the same Indian Buffet)

  Dirichlet Process on the $\boldsymbol{\pi}_i$

- dictionary learning : Gibbs sampling
- extends to missing pixels...

Bayesian non parametric dictionary learning

The model

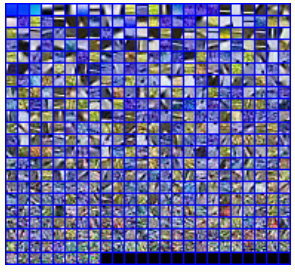
[Zhou, Carin, Paisley et al.'11'12]

$$
\begin{cases}
\mathbf{x}_i &= \mathbf{D}\mathbf{w}_i + \varepsilon_i & \text{patches} \\
\mathbf{w}_i &= \mathbf{z}_i \odot \mathbf{s}_i & 0 \text{ or } s_{ik} \text{ if } z_{ik} = 1 \\
\mathbf{s}_i &\sim \mathcal{N}(0, \gamma_s^{-1}\mathbf{I}_K) & \text{non-zero coefficients} \\
\mathbf{z} &\sim \text{Bernoulli}(\boldsymbol{\pi}) & \text{sparsity : coeff is here or not} \\
\boldsymbol{\pi}_i &\sim \text{BP}(a, b) & \text{favor reuse of same features} \\
\mathbf{d}_k &\sim \mathcal{N}(0, P^{-1}\mathbf{I}_P) & \text{prior dict. features} \\
\varepsilon_i &\sim \mathcal{N}(0, \gamma_\varepsilon^{-1}\mathbf{I}_P) & \text{unknown noise level}
\end{cases}
$$

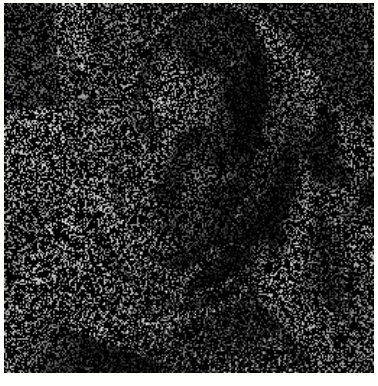
Bayesian non parametric dictionary learning

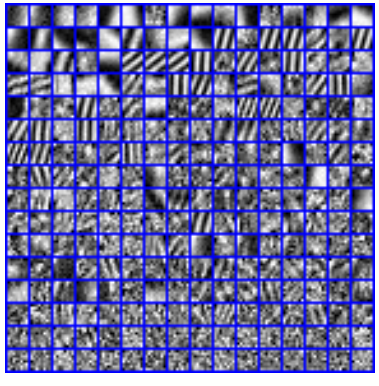
[Zhou, Carin, Paisley et al.'11'12]



infers the number of necessary dictionary elements



In summary : Bayesian non parametric is a rich framework

1. efficient priors for **non parametric clustering** :

   Dirichlet and generalizations (Pitman-Yor process...)

   inference : **Chinese Restaurant Process**

   - segmentation,
   - joint processing,
   - non Gaussian distributions (GSM...)

2. efficient priors for **latent feature models** :

   Beta & Bernoulli processes and generalizations

   inference : **Indian Buffet Process**

   - source separation,
   - mixture of components (multi/hyper spectral),
   - dictionary learning...

In summary : Bayesian non parametric is a rich framework

1. **unknown noise level taken into account**
2. complexity of the model governed by the data :

   bypasses model selection
3. **inference algorithms** :
   - Gibbs sampling (CRP & IBP...)
   - variational Bayesian approximation,
   - EM (truncation)

Perspectives

1. many models to explore, including analysis approaches,
2. revisiting inverse problems (blind deconvolution...)
3. non Gaussian noise, non Gaussian models,
4. dictionary learning : still much work to be done...
5. multi-component systems (multi-spectral...)
6. progress to expect on algorithms,
7. ...

Dirichlet processes (DP)

- a DP is an 'infinitely decimated' Dirichlet distribution : as the limit of Dirichlet$(\alpha/K, ..., \alpha/K)$ as $K \to \infty$
- a DP is a distribution over probability measures
- a DP has two parameters :
  - Base distribution G_0, like the *mean* of the DP,
  - Strength parameter α, *concentration* of the DP.
- $G \sim DP(\alpha, G_0)$ are discrete distributions

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

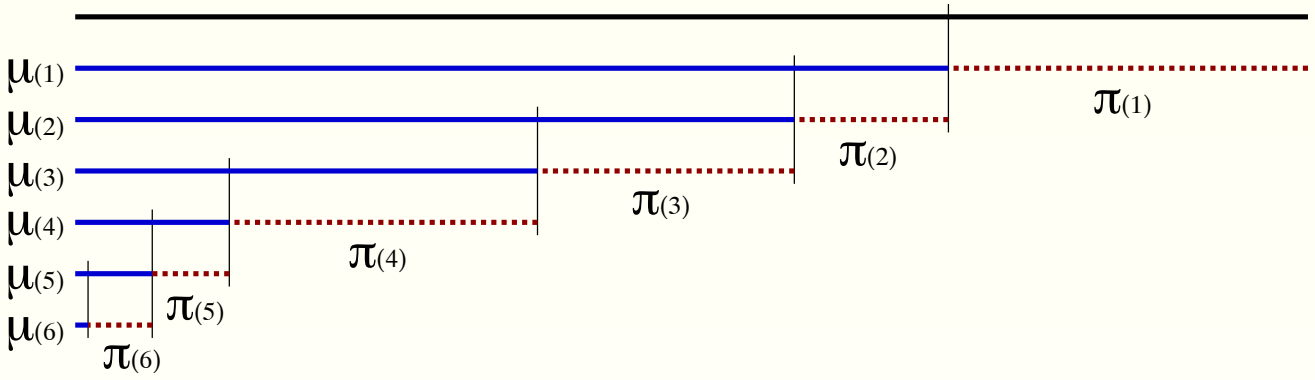
  - $\theta_k \sim G_0$
  - $\pi_k = \nu_k \prod_{j=1}^{k-1}(1 - \nu_j) =$ stick-breaking
  - $\nu_j \sim \text{Beta}(1, \alpha)$ where $\text{Beta}(x|1, \alpha) \propto (1 - x)^{\alpha - 1}$

[Ferguson'73]

Posterior Dirichlet processes (DP)

Inference

$\pi_k = \nu_k \prod_{j=1}^{k-1}(1-\nu_j)$



Fundamental motivation : **posterior updating** (see CRP)

$$G|\theta \sim DP\left(\alpha+1, \frac{\alpha G_0+\delta_\theta}{\alpha+1}\right)$$

or more generally

$$\text{if } \theta_1,...,\theta_n|G \sim G \text{ and } G \sim DP(\alpha, G_0)$$
$$\text{then } G|\theta_1,...,\theta_n \sim DP(\alpha+n, G_0+\textstyle\sum_{i=1}^{n}\delta_{\theta_i})$$

$\Rightarrow$ **inference : G clusters on previous estimates**

The key : de Finetti's theorem and exchangeable variables

Hidden variables

$$\forall \sigma, p(X_1, ..., X_n) = p(X_{\sigma(1)}, ..., X_{\sigma(n)})$$

$$\Longleftrightarrow$$

$$p(X_1, ..., X_n|\theta) = \prod_i p(X_i|\theta)$$