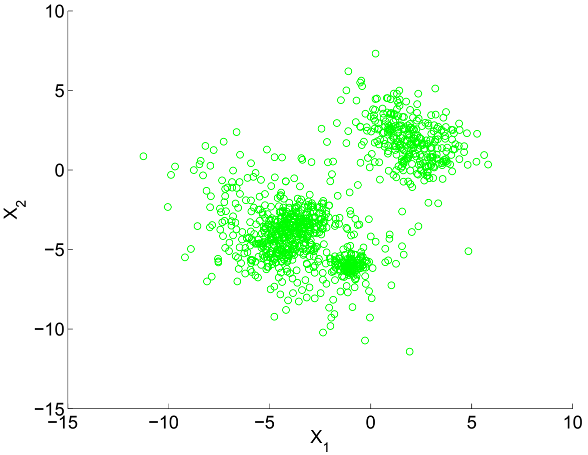
The Chinese Restaurant Process: Bayesian Inference of Mixture Models and Applications in Computational Biology

Ivan Gesteira Costa Filho

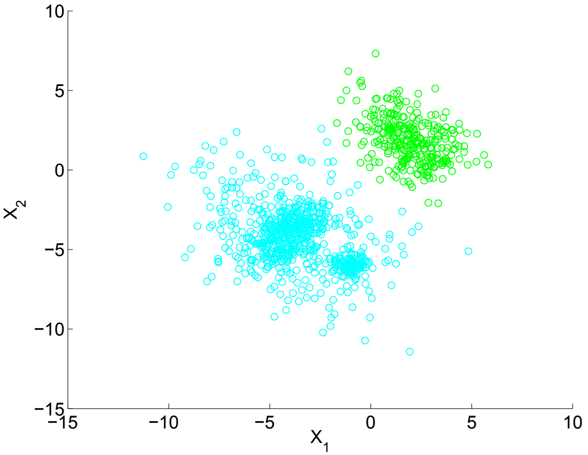
Max Planck Institute for Molecular Genetics

The clustering problem



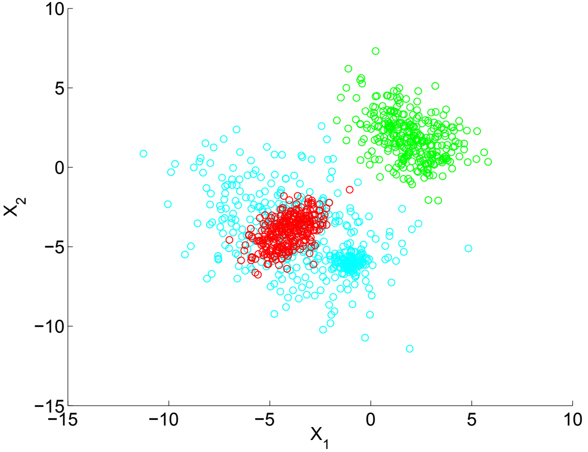
- K^N partition of N data points in K clusters.
- **number of clusters** (or K)?

The clustering problem



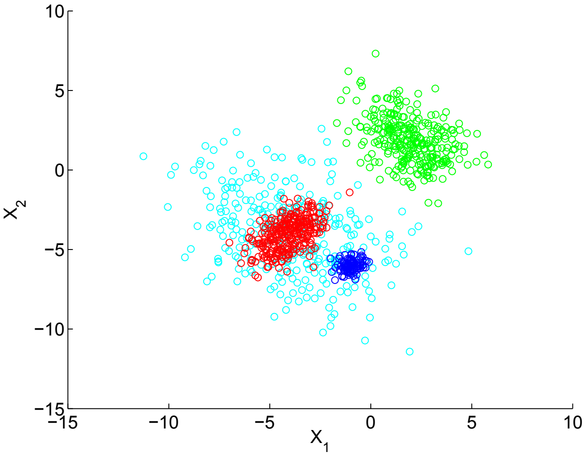
- K^N partition of N data points in K clusters.
- **number of clusters** (or K)?

The clustering problem



- K^N partition of N data points in K clusters.
- **number of clusters** (or K)?

The clustering problem



- K^N partition of N data points in K clusters.
- **number of clusters** (or K)?

Mixture Model

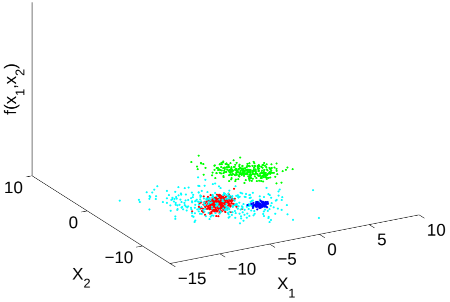
Definition

Convex combination of K probability distributions

$$\mathbf{P}(x|\Theta) = \sum_{k=1}^{K} \pi_k \mathbf{P}_k(x|\theta_k)$$

where x is an L-dimensional variable
$(\pi_1, ..., \pi_K)$ - mixing coefficients $\sum_1^K \pi_k = 1$ and $\pi_k > 0$
$\Theta = (\pi_1, ..., \pi_K, \theta_1, ..., \theta_K)$ - the model parameters.

Mixture Model and Clustering



Mixture Example

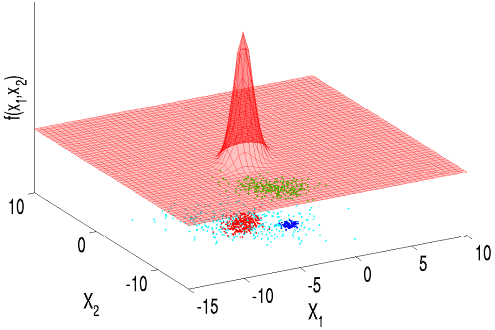
Using Normals,

$$\mathbf{P}(x|\Theta) \quad =$$

Clustering Interpretation

- $\pi_k \propto$ size of cluster k
- $y_i = j$ - cluster from x_i where $j \in \{1, ..., K\}$
- $\mathbf{P}(y_i = j | x_i, \Theta) \propto \pi_j \mathbf{P}(x_i | \theta_j)$.

Mixture Model and Clustering



Mixture Example

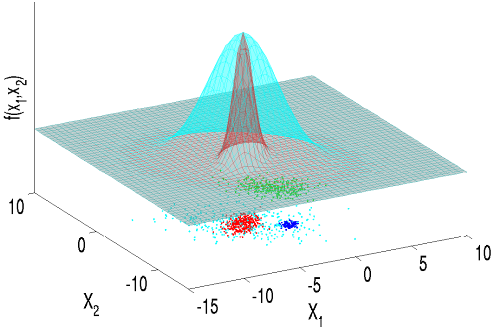
Using Normals,

$$\mathbf{P}(x|\Theta) = \pi_1 * \mathbf{N}(x|\mu_1, \Sigma_1)$$

Clustering Interpretation

- $\pi_k \propto$ size of cluster k
- $y_i = j$ - cluster from x_i where $j \in \{1, \ldots, K\}$
- $\mathbf{P}(y_i = j | x_i, \Theta) \propto \pi_j \mathbf{P}(x_i | \theta_j)$.

Mixture Model and Clustering



Mixture Example

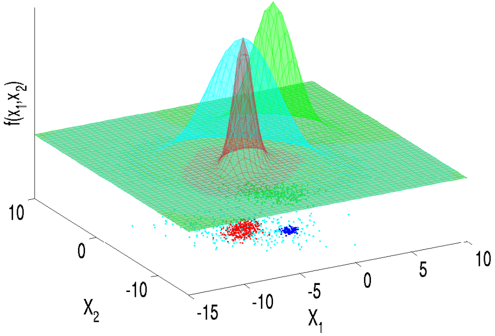
Using Normals,

$$\mathbf{P}(x|\Theta) = \pi_1 * \mathbf{N}(x|\mu_1,\Sigma_1) + \pi_2 * \mathbf{N}(x|\mu_2,\Sigma_2)$$

Clustering Interpretation

- $\pi_k \propto$ size of cluster k
- $y_i = j$ - cluster from x_i where $j \in \{1, ..., K\}$
- $\mathbf{P}(y_i = j|x_i,\Theta) \propto \pi_j \mathbf{P}(x_i|\theta_j)$.

Mixture Model and Clustering



Mixture Example

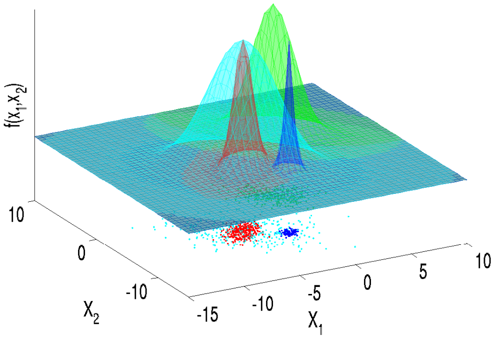
Using Normals,

$$\begin{aligned}\mathbf{P}(x|\Theta) &= \pi_1 * \mathbf{N}(x|\mu_1,\Sigma_1) \\ &+ \pi_2 * \mathbf{N}(x|\mu_2,\Sigma_2) \\ &+ \pi_3 * \mathbf{N}(x|\mu_3,\Sigma_3)\end{aligned}$$

Clustering Interpretation

- $\pi_k \propto$ size of cluster k
- $y_i = j$ - cluster from x_i where $j \in \{1, ..., K\}$
- $\mathbf{P}(y_i = j | x_i, \Theta) \propto \pi_j \mathbf{P}(x_i|\theta_j)$.

Mixture Model and Clustering



Mixture Example

Using Normals,

$$\begin{aligned} \mathbf{P}(x|\Theta) &= \pi_1 * \mathbf{N}(x|\mu_1, \Sigma_1) \\ &+ \pi_2 * \mathbf{N}(x|\mu_2, \Sigma_2) \\ &+ \pi_3 * \mathbf{N}(x|\mu_3, \Sigma_3) \\ &+ \pi_4 * \mathbf{N}(x|\mu_4, \Sigma_4) \end{aligned}$$

Clustering Interpretation

- $\pi_k \propto$ size of cluster k
- $y_i = j$ - cluster from x_i where $j \in \{1, ..., K\}$
- $\mathbf{P}(y_i = j | x_i, \Theta) \propto \pi_j \mathbf{P}(x_i | \theta_j)$.

Estimation of Mixture Models

Problem - Maximum Likelihood Estimation (MLE)

find Θ maximizing $\mathbf{P}(\mathbf{X}|\Theta)$

$$\arg\max_{\Theta} \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k \cdot \mathbf{P}_k(x_i|\theta_k).$$

where $\mathbf{X}$ are N data points x_i

Solution

The complete data likelihood

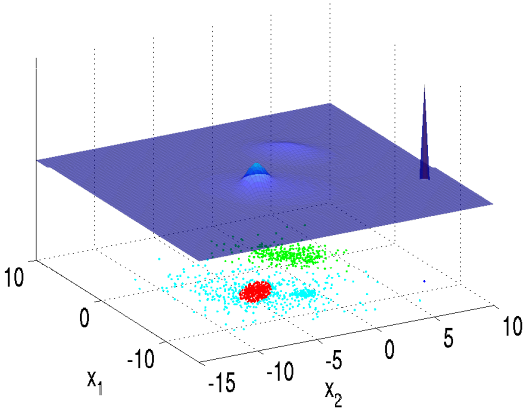
$$\mathbf{P}(\mathbf{X},\mathbf{Y}|\Theta) = \prod_{k=1}^{K} \prod_{i=1}^{N} (\pi_k \cdot \mathbf{P}_k(x_i|\theta_k))^{\mathbf{1}\{y_i=k\}}$$

Expectation-Maximization (EM) algorithm maximizes $\mathbf{P}(\mathbf{X}|\Theta)$ locally.

where $\mathbf{Y}$ are N cluster assignments y_i.

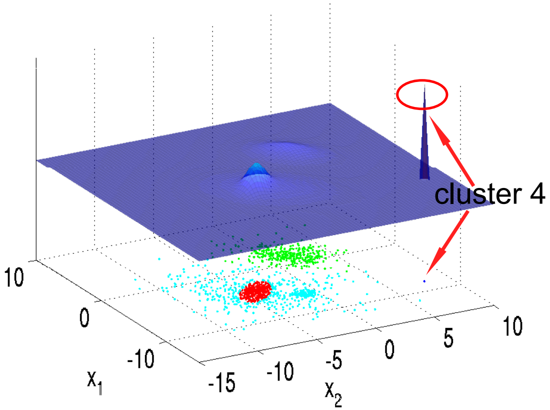
MLE Problems - Over-fitting

A MLE solution with 4 clusters



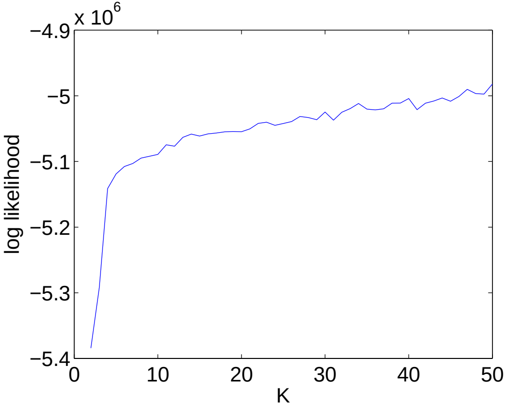
MLE Problems - Over-fitting

A MLE solution with 4 clusters



MLE Problems - Number of Clusters

Likelihood against **number of clusters** (K).



- Alternative - model selection via penalized likelihood methods (e.g. BIC and AIC).

Motivation

- Mixture models
  - statistical formalism to clustering.
- Problems: Maximum likelihood estimation of mixture models
  - prone to **over-fitting**.
  - do not determine **number of clusters**.

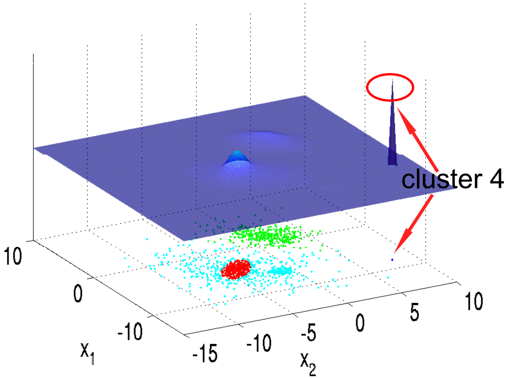
Bayesian Estimation of Mixture Models

- Use prior beliefs about model parameters $\mathbf{P}(\Theta)$ to reduce uncertainty in model estimation.
- By Bayes rule,

$$\mathbf{P}(\Theta|\mathbf{X},\mathbf{Y}) = \frac{\mathbf{P}(\mathbf{X},\mathbf{Y}|\Theta)\mathbf{P}(\Theta)}{\mathbf{P}(\mathbf{X},\mathbf{Y})}$$

- The posterior distribution
  - mode, mean - most likely parameters given the data and prior distribution.
  - variance - confidence intervals on parameters.

Bayesian Analysis - Including Prior Knowledge

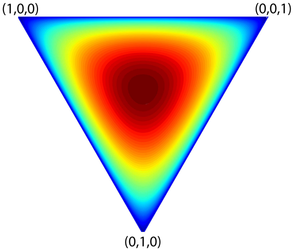


Very small clusters → **over-fitting**.

Bayesian Analysis - Prior on Cluster Sizes

Example of prior on cluster sizes: Symmetric-Dirichlet $(\pi_1, \pi_2, \pi_3 | \alpha)$

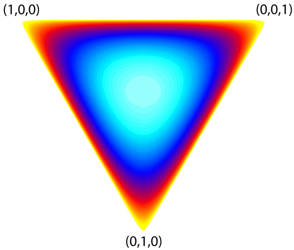
$\alpha > 3$



Bayesian Analysis - Prior on Cluster Sizes

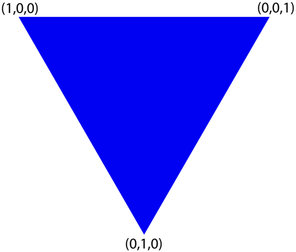
Example of prior on cluster sizes: Symmetric-Dirichlet $(\pi_1, \pi_2, \pi_3 | \alpha)$

$\alpha < 3$



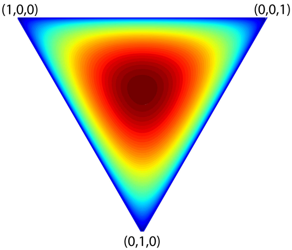
unequal size clusters

$\alpha = 3$



any size clusters

$\alpha > 3$

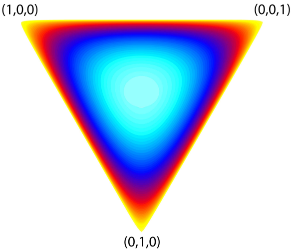


equal size clusters

Bayesian Analysis - Prior on Cluster Sizes

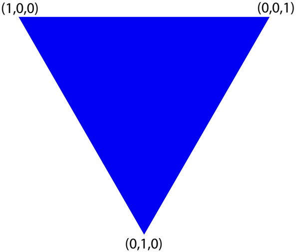
Example of prior on cluster sizes: Symmetric-Dirichlet $(\pi_1, \pi_2, \pi_3 | \alpha)$

$\alpha < 3$



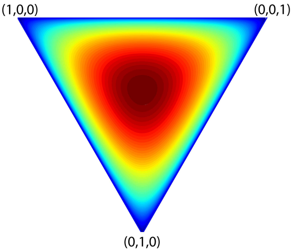
unequal size clusters

$\alpha = 3$



any size clusters

$\alpha > 3$



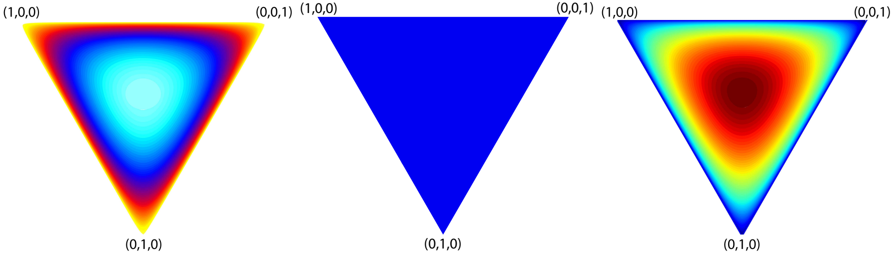
equal size clusters

$$\mathbf{P}(\Theta) = \mathbf{P}(\pi_1, \dots, \pi_K)$$

Bayesian Analysis - Prior on Cluster Sizes

Example of prior on cluster sizes: Symmetric-Dirichlet $(\pi_1, \pi_2, \pi_3 | \alpha)$

$\alpha < 3$ | $\alpha = 3$ | $\alpha > 3$



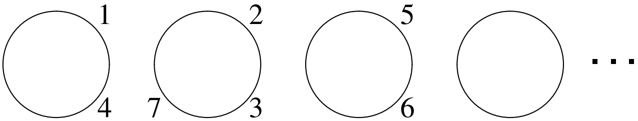
unequal size clusters | any size clusters | equal size clusters

$$\mathbf{P}(\Theta) = \mathbf{P}(\pi_1, \ldots, \pi_K) \underbrace{\mathbf{P}(\theta_1, \ldots, \theta_K)}_{\sum_{k=1}^{K} \mathbf{P}(\mu_k | \Sigma_k) \mathbf{P}(\Sigma_k)}$$

Bayesian Analysis - Prior on Number of Clusters?

- Distribution over **number of clusters**?

Chinese Restaurant Process (CRP)

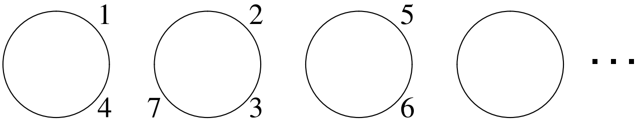


CRP is the process on how a customer n choose a table $1, ..., \infty$ to sit.

Analogy to the clustering problem:

- tables are clusters.
- customers are data points.
- a seating configuration is a partition.

Chinese Restaurant Process (CRP)



The nth customer chooses

- an occupied table $j \in \{1, ..., K\}$

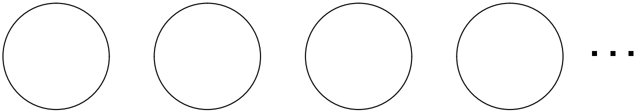
$$\mathbf{P}(y_n = j | y_1, ..., y_{n-1}) \propto \underbrace{\text{\#customers at table } j}_{n_j}$$

- an empty table

$$\mathbf{P}(y_n > K | y_1, ..., y_{n-1}) \propto \alpha$$

where K is the number of occupied tables, y_n is the table customer n sits and $\alpha \geq 0$.

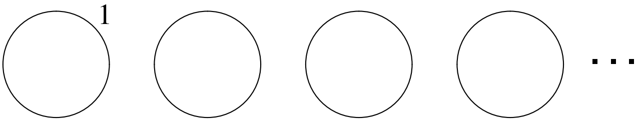
Chinese Restaurant Process (CRP)



CRP - probability distribution over seating configurations (or partitions)

$$\mathbf{P}(y_1 = 1, y_2 = 2, ..., y_7 = 3) \quad =$$

Chinese Restaurant Process (CRP)

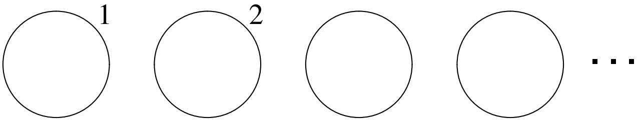


CRP - probability distribution over seating configurations (or partitions)

$$\mathbf{P}(y_1 = 1, y_2 = 2, ..., y_7 = 3) = \mathbf{P}(y_1 = 1)$$

$$= \frac{\alpha}{\alpha}$$

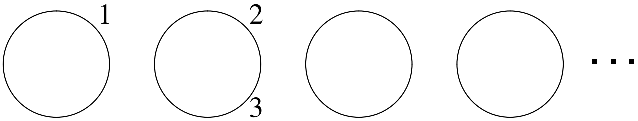
Chinese Restaurant Process (CRP)



CRP - probability distribution over seating configurations (or partitions)

$$
\begin{aligned}
\mathbf{P}(y_1 = 1, y_2 = 2, ..., y_7 = 3) &= \mathbf{P}(y_1 = 1) \cdot \mathbf{P}(y_2 = 2 | y_1) \cdot \\
&= \frac{\alpha}{\alpha} \cdot \frac{\alpha}{1 + \alpha}
\end{aligned}
$$

Chinese Restaurant Process (CRP)

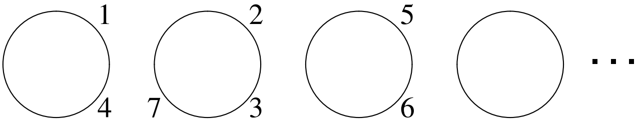


CRP - probability distribution over seating configurations (or partitions)

$$\mathbf{P}(y_1 = 1, y_2 = 2, ..., y_7 = 3) = \mathbf{P}(y_1 = 1) \cdot \mathbf{P}(y_2 = 2 | y_1) \cdot \mathbf{P}(y_3 = 2 | y_1, y_2)$$

$$= \frac{\alpha}{\alpha} \cdot \frac{\alpha}{1+\alpha} \cdot \frac{1}{2+\alpha}$$

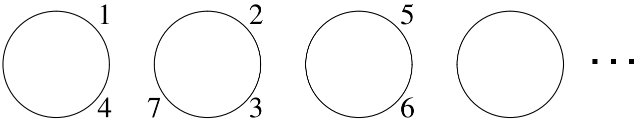
Chinese Restaurant Process (CRP)



CRP - probability distribution over seating configurations (or partitions)

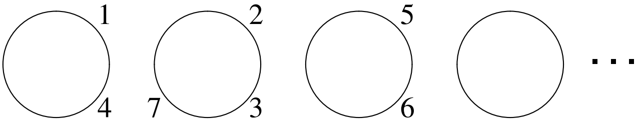
$$\begin{aligned}\mathbf{P}(y_1 = 1, y_2 = 2, ..., y_7 = 3) &= \mathbf{P}(y_1 = 1) \cdot \mathbf{P}(y_2 = 2 | y_1) \cdot \mathbf{P}(y_3 = 2 | y_1, y_2) \\ &\quad \cdot ... \cdot \mathbf{P}(y_7 = 3 | y_1, ..., y_6) \\ &= \frac{\alpha}{\alpha} \cdot \frac{\alpha}{1+\alpha} \cdot \frac{1}{2+\alpha} \cdot \frac{1}{3+\alpha} \cdot \frac{\alpha}{4+\alpha} \cdot \frac{1}{5+\alpha} \cdot \frac{2}{6+\alpha}\end{aligned}$$

Chinese Restaurant Process (CRP)



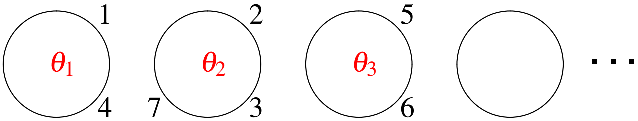
- Clustering Problem: data points in a cluster are similar.
- CRP: only table sizes (and α) are considered.
- Solution ...

Weighted Chinese Restaurant (WCRP)



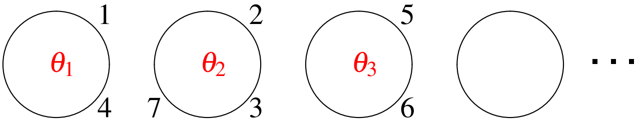
- Customers seat at tables with similar occupants.

Weighted Chinese Restaurant (WCRP)



- Customers seat at tables with similar occupants.
- Table j has parameter θ_j.

Weighted Chinese Restaurant (WCRP)



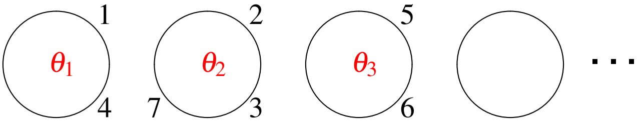
The nth customer chooses

- an occupied table $j \in \{1, ..., K\}$

$$\mathbf{P}(y_n = j | \mathbf{Y}_{n-1}, x_n, \Theta) \propto n_j \mathbf{P}(x_n | \theta_j)$$

where $\Theta = (\theta_1, ..., \theta_K)$, $\mathbf{P}(\theta)$ is a prior distribution over table parameters θ, x_n are a variable describing customer n, n_j is the number of customers in table j.

Weighted Chinese Restaurant (WCRP)



The nth customer chooses

- an occupied table $j \in \{1, ..., K\}$

$$\mathbf{P}(y_n = j | \mathbf{Y}_{n-1}, x_n, \Theta) \propto n_j \mathbf{P}(x_n | \theta_j)$$

- an empty table

$$\mathbf{P}(y_n > K | \mathbf{Y}_{n-1}, x_n, \Theta) \propto \alpha \underbrace{\int_\theta \mathbf{P}(x_n|\theta)\mathbf{P}(\theta)d(\theta)}_{\mathbf{P}(x_n)}$$

where $\Theta = (\theta_1, ..., \theta_K)$, $\mathbf{P}(\theta)$ is a prior distribution over table parameters θ, x_n are a variable describing customer n, n_j is the number of customers in table j.

WCRP and Infinite Mixtures

Integrating out y, the WCRP defines an infinite mixture.

$$\mathbf{P}(x_n|\Theta) = \sum_{k=1}^{K} \frac{n_k}{n-1+\alpha}\mathbf{P}(x_n|\theta_k) + \frac{\alpha}{n-1+\alpha}\int_{\theta}\mathbf{P}(x_n|\theta)\mathbf{P}(\theta)d(\theta)$$

WCRP and Infinite Mixtures

Integrating out y, the WCRP defines an infinite mixture.

$$\mathbf{P}(x_n|\Theta) = \underbrace{\sum_{k=1}^{K} \frac{n_k}{n-1+\alpha} \mathbf{P}(x_n|\theta_k)}_{\text{occupied tables}} + \underbrace{\frac{\alpha}{n-1+\alpha} \int_{\theta} \mathbf{P}(x_n|\theta) \mathbf{P}(\theta) d(\theta)}_{\text{empty tables}}$$

WCRP and Infinite Mixtures

Integrating out y, the WCRP defines an infinite mixture.

$$\mathbf{P}(x_n|\Theta) = \underbrace{\sum_{k=1}^{K} \frac{n_k}{n-1+\alpha}\mathbf{P}(x_n|\theta_k)}_{\text{occupied tables}} + \underbrace{\frac{\alpha}{n-1+\alpha}\int_{\theta}\mathbf{P}(x_n|\theta)\mathbf{P}(\theta)d(\theta)}_{\text{empty tables}}$$

- Problem: inference infeasible for most choices of $\mathbf{P}(\theta)$ and $\mathbf{P}(x|\theta)$

WCRP and Infinite Mixtures

Integrating out y, the WCRP defines an infinite mixture.

$$\mathbf{P}(x_n|\Theta) = \underbrace{\sum_{k=1}^{K} \frac{n_k}{n-1+\alpha} \mathbf{P}(x_n|\theta_k)}_{\text{occupied tables}} + \underbrace{\frac{\alpha}{n-1+\alpha} \int_{\theta} \mathbf{P}(x_n|\theta)\mathbf{P}(\theta)d(\theta)}_{\text{empty tables}}$$

- Problem: inference infeasible for most choices of $\mathbf{P}(\theta)$ and $\mathbf{P}(x|\theta)$
- WCRP: how to sample from $\mathbf{P}(y_n|\mathbf{Y}_{n-1}, x_n, \Theta)$

WCRP and Gibbs Sampling

Gibbs Sampler: define conditional distributions of $\mathbf{Y}$ and $\mathbf{\Theta}$.

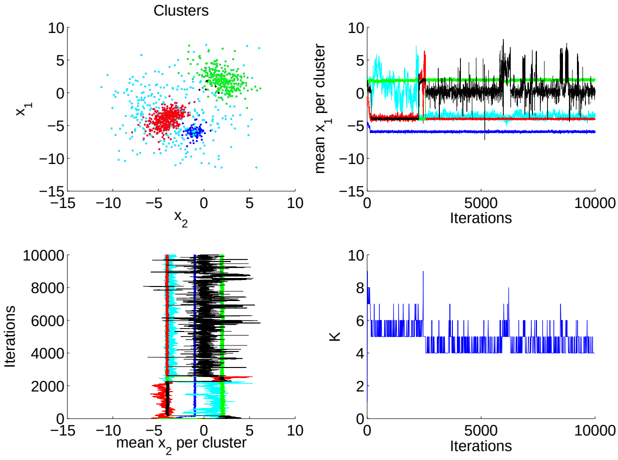
Algorithm

for t from 1 to T

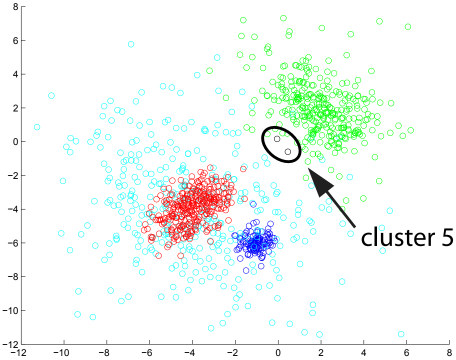
- Draw $\mathbf{Y}^t$ from $\mathbf{P}(y_i|\mathbf{Y}^{t-1},\mathbf{X},\Theta^{t-1}) =$ seating prob. from WCRP
- Draw Θ^t from $\mathbf{P}(\theta_k|\mathbf{Y}^t,\mathbf{X},\Theta^{t-1}) \propto \mathbf{P}(\mathbf{X},\mathbf{Y}^t|\theta_k^{t-1})\mathbf{P}(\theta_k^{t-1})$

After convergence, samples come from the posterior $\mathbf{P}(\mathbf{\Theta}|\mathbf{X},\mathbf{Y})$.

WCRP and Gibbs Sampler - Example

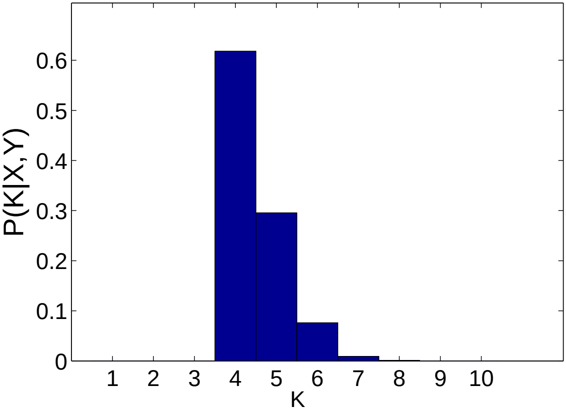


Example WCRP - Maximum-a-posteriori Solution



$$\underset{t \in \{1,\ldots,T\}}{\arg\max}\, \mathbf{P}(\mathbf{X}, \mathbf{Y}^t | \Theta^t) \mathbf{P}(\Theta^t)$$

Example WCRP - Posterior Analysis

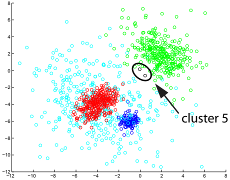


$$\mathbf{P}(K|\mathbf{X},\mathbf{Y}) = \int_{\Theta} \mathbf{P}(K|\Theta,\mathbf{X},\mathbf{Y})\mathbf{P}(\Theta|\mathbf{X},\mathbf{Y})d(\Theta)$$

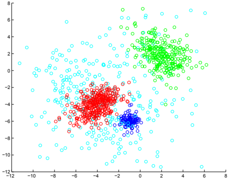
$$\rightarrow \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}(\max(\mathbf{Y}^t) = K)$$

Example WCRP - Posterior on Cluster Means

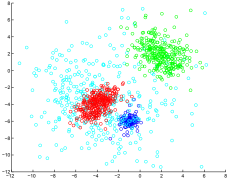
Some solutions from $\mathbf{Y}^1, \ldots, \mathbf{Y}^T$.



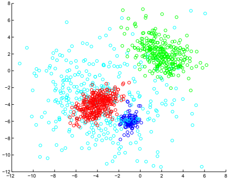
K=5



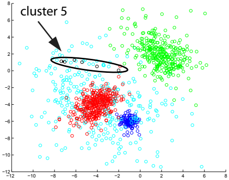
K=4



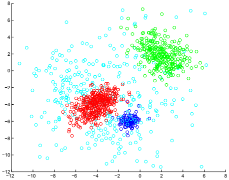
K=4



K=4



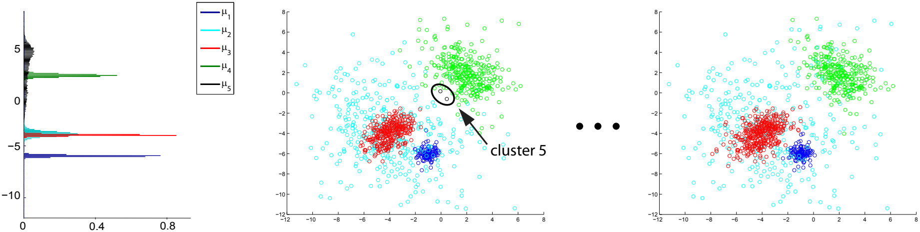
K=5



K=4

Example WCRP - Posterior on Cluster Means

Posterior of cluster means (μ_k) over all solutions $\mathbf{Y}^1, ..., \mathbf{Y}^T$.



$\mathbf{P}(\mu_k|\mathbf{X},\mathbf{Y})$ $\mathbf{Y}^1$ ... $\mathbf{Y}^N$

$$\mathbf{P}(\mu_k|\mathbf{X},\mathbf{Y}) = \int_\Theta \mathbf{P}(\mu_k|\Theta,\mathbf{X},\mathbf{Y})\mathbf{P}(\Theta|\mathbf{X},\mathbf{Y})d(\Theta)$$

where $\pi_1 > \pi_2 > ... > \pi_K$.

Mixture Models and Computational Biology

- Mixture model based clustering has several applications in Computational Biology.
    - haplotypes reconstruction from single nucleotide polymorphisms (SNP) [Excoffier and Slatkin, 1995, Xing et al., 2007].
    - finding common mutagenic events in HIV patients [Beerenwinkel et al., 2004].
    - finding clusters of co-expressed genes from transcription data [Bar-Joseph et al., 2003, Schliep et al., 2003].
- **Number of clusters** usually unknown.

Problem - Haplotype Reconstruction

SNPs in a Diploid Individual

$\mathbf{C}_{father}$- CGTCACGGACATG

$\mathbf{C}_{mother}$- CGCCACTGACATG

Problem - Haplotype Reconstruction

SNPs in a Diploid Individual

C$_{father}$- CGTCACGGACATG

C$_{mother}$- CGCCACTGACATG

$\longrightarrow$

Haplotype pair (h_1, h_2)

h_1={TGA}

h_2={CTA}

Problem - Haplotype Reconstruction

SNPs in a Diploid Individual

C$_{father}$- CGTCACGGACATG

C$_{mother}$- CGCCACTGACATG

$\longrightarrow$

Haplotype pair (h_1, h_2)

h_1={TGA}

h_2={CTA}

$\downarrow$

A SNP genotype can be read with sequencing.

g={(C,T),(G,T),(A,A)}

Problem - Haplotype Reconstruction

SNPs in a Diploid Individual

C$_{father}$- CGTCACGGACATG

C$_{mother}$- CGCCACTGACATG

→

Haplotype pair (h_1, h_2)

h_1={TGA}

h_2={CTA}

↓

A SNP genotype can be read with sequencing.

g={(C,T),(G,T),(A,A)}

→

How to resolve (h_1, h_2) from g?

h_1={TGA} h_1 ={CGA}

X

h_2={CTA} h_2 ={TTA}

Problem - Haplotype Reconstruction

SNPs in a Diploid Individual

C$_{father}$- CGTCACGGACATG

C$_{mother}$- CGCCACTGACATG

$\longrightarrow$

Haplotype pair (h_1, h_2)

h_1={TGA}

h_2={CTA}

$\downarrow$

$\uparrow$?

A SNP genotype can be read with sequencing.

g={(C,T),(G,T),(A,A)}

$\longrightarrow$

How to resolve (h_1, h_2) from g?

h_1={TGA} h_1 ={CGA}

X

h_2={CTA} h_2 ={TTA}

Haplotype Reconstruction and Mixture Models

For a set of genotypes $\mathbf{G}$

$$g_1=\{(C,T),(G,T),(A,A)\}$$
$$g_2=\{(C,T),(G,G),(A,A)\}$$
$$g_3=\{(C,C),(T,T),(A,A)\}$$
$$\vdots$$
$$g_n=\{(C,C),(G,T),(A,A)\}$$

$\longrightarrow$

Find their haplotypes from a set $\mathbf{H}$

$$h_1=\{TGA\}\ h_2=\{CTA\}$$
$$h_1=\{TGA\}\ h_3=\{CGA\}$$
$$h_2=\{CTA\}\ h_2=\{CTA\}$$
$$\vdots$$
$$h_3=\{CGA\}\ h_2=\{CTA\}$$

Definition

A genotype g is a mixture over the set of haplotypes $\mathbf{H}$ in a population.

$$\mathbf{P}(g|\mathbf{H}) = \sum_{h_i,h_j\in\mathbf{H}} \mathbf{P}(h_i)\mathbf{P}(h_j)\mathbf{P}(g|h_i,h_j)$$

(assuming Hardy-Weinberg equilibrium $\mathbf{P}(h_i,h_j) = \mathbf{P}(h_i)\mathbf{P}(h_j)$)

Haplotype Reconstruction and Mixture Models

For a set of genotypes $\mathbf{G}$ — Find their haplotypes from a set $\mathbf{H}$

$$
\begin{array}{lcl}
g_1 = \{(C,T),(G,T),(A,A)\} & & h_1 = \{TGA\}\ h_2 = \{CTA\} \\
g_2 = \{(C,T),(G,G),(A,A)\} & \longrightarrow & h_1 = \{TGA\}\ h_3 = \{CGA\} \\
g_3 = \{(C,C),(T,T),(A,A)\} & & h_2 = \{CTA\}\ h_2 = \{CTA\} \\
\vdots & & \vdots \\
g_n = \{(C,C),(G,T),(A,A)\} & & h_3 = \{CGA\}\ h_2 = \{CTA\}
\end{array}
$$

Definition

A genotype g is a mixture over the set of haplotypes $\mathbf{H}$ in a population.

$$
\underbrace{\mathbf{P}(g|\mathbf{H})}_{\mathbf{P}(x|\Theta)} = \sum_{h_i,h_j \in \mathbf{H}} \underbrace{\mathbf{P}(h_i)}_{\pi_i} \underbrace{\mathbf{P}(h_j)}_{\pi_j} \underbrace{\mathbf{P}(g|h_i,h_j)}_{\mathbf{P}(x|\theta_i,\theta_j)}
$$

(assuming Hardy-Weinberg equilibrium $\mathbf{P}(h_i,h_j) = \mathbf{P}(h_i)\mathbf{P}(h_j)$)

Bayesian Analysis in Haplotype Reconstruction

- DP-Haplotyper [Xing et al., 2007]
  - Bayesian Mixture model based Haplotype Reconstruction.
  - Weighted Chinese Restaurant Process as prior distribution.
- Prior knowledge
  - few haplotypes in a population of related individuals
  - small probability of opening a new table ($\alpha = N/100$)

Results - Haplotype Reconstruction

Number of haplotypes over samples 1 to T.

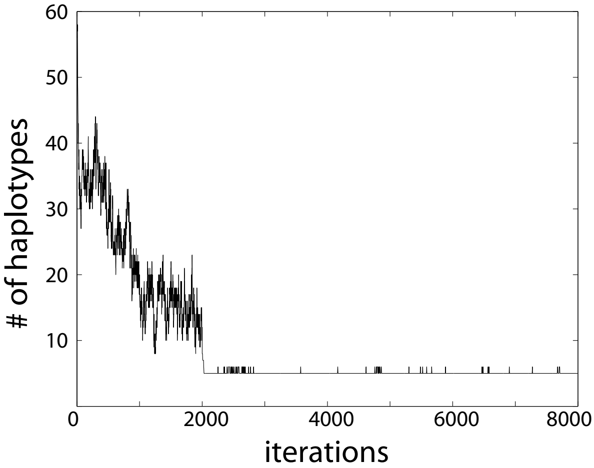


Image from Xing et al. [2007]

Results - Haplotype Reconstruction

Error on SNPs positions for state of the art methods

Chromosome region	**DP-Hap.** Xing et al. [2007]	PHASE Stephens et al. [2001]	HAP Eskin et al. [2003]
1	**0**	0.003	0.007
2	**0.007**	**0.007**	0.036
3	**0**	**0**	**0**
4	**0**	**0**	**0**
5	**0.011**	**0.011**	0.027
6	0.005	**0**	0.018
7	**0.005**	**0.005**	0.068
8	**0**	**0**	**0**
9	**0.012**	**0.012**	0.057
10	**0.007**	0.008	0.042
11	**0.005**	0.011	0.033
12	**0**	**0**	**0**
Average	**0.004**	0.005	0.025
# Haplotypes	**5-12**	60-100	NA

Human Haplotype data from Dale et al. 2001

Discussion

- The weighted Chinese restaurant process (WCRP):
  - Bayesian estimation over infinite mixture models.
  - inclusion of prior knowledge regarding number of clusters.
- In haplotype reconstruction, the WCRP was more accurate than competing methods.
  - prior distribution favored solution with few haplotypes.

References

Z. Bar-Joseph, G. K. Gerber, D. K. Gifford, T. S. Jaakkola, and I. Simon. Continuous representations of time-series gene expression data. *J Comput Biol*, 10(3-4):341–356, 2003.

N. Beerenwinkel, J. Rahnenführer, M. Daumer, D. Hoffmann, R. Kaiser, J. Selbig, and T. Lengauer. Learning multiple evolutionary pathways from cross-sectional data. In *RECOMB 2004: Proceedings of the eighth annual international conference on Research in computational molecular biology*, pages 36–44. ACM Press, 2004.

E. Eskin, E. Halperin, and R. M. Karp. Efficient reconstruction of haplotype structure via perfect phylogeny. *J Bioinform Comput Biol*, 1(1):1–20, Apr 2003.

L. Excoffier and M. Slatkin. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol*, 12(5):921–927, Sep 1995.

A. Schliep, A. Schönhuth, and C. Steinhoff. Using Hidden Markov models to analyze gene expression time course data. *Bioinformatics*, 19 Suppl 1:i255–i263, 2003.

M. Stephens, N. J. Smith, and P. Donnelly. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet*, 68(4):978–989, Apr 2001.

E. P. Xing, M. I. Jordan, and R. Sharan. Bayesian haplotype inference via the dirichlet process. *J Comput Biol*, 14(3): 267–284, Apr 2007.

Conjugate Priors

$$\begin{aligned}
\mathbf{P}(\mu, \Sigma) &= \text{NIW}(\mu, \Sigma | \mu_0, \sigma_0, \Lambda_0, \nu_0) \\
&= \frac{1}{Z} |\Sigma|^{-(\nu_0+L)2+1} \exp\left(-\frac{1}{2}\text{trace}(\Lambda_0 \Sigma^{-1}) - \frac{\sigma_0}{2}(\mu - \mu_0)^T \Sigma^{-1} (\mu - \mu_0) \right)
\end{aligned}$$

$$\begin{aligned}
\mathbf{P}(\pi_1, ..., \pi_K) &= \text{Symm-Dirichlet}(\pi_1, ..., \pi_K | \alpha) \\
&= \frac{\Gamma(\alpha)}{\prod_{j=1}^K \Gamma(\alpha/K)} \prod_{j=1}^K \pi_j^{\alpha/K-1}
\end{aligned}$$