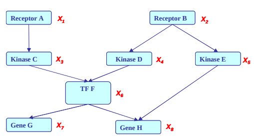
School of Computer Science
Carnegie Mellon

Infinite Mixture and Dirichlet Process

Probabilistic Graphical Models (10-708)

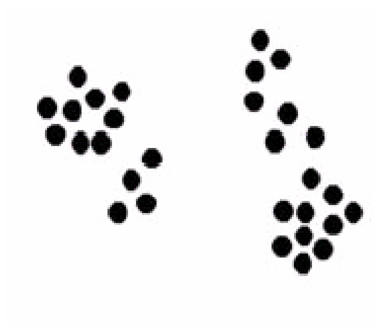
Lecture 20, Nov 28, 2007



Eric Xing

Reading:

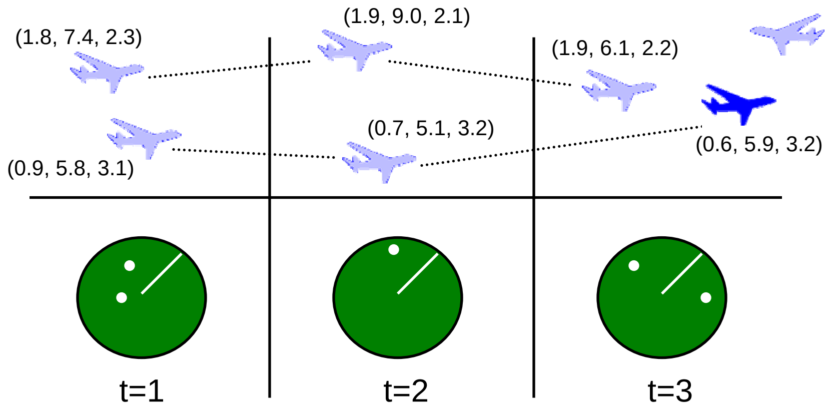
Clustering



K?

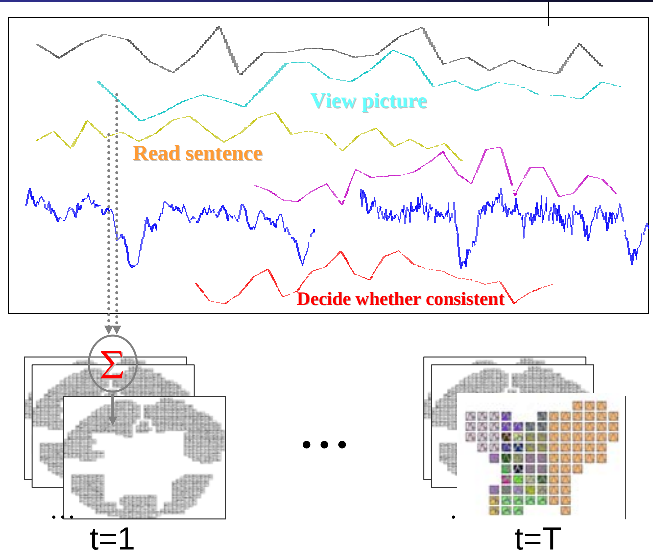
P(K)

Object Recognition and Tracking

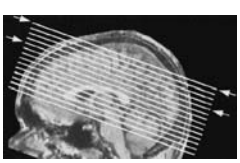


Modeling The Mind …

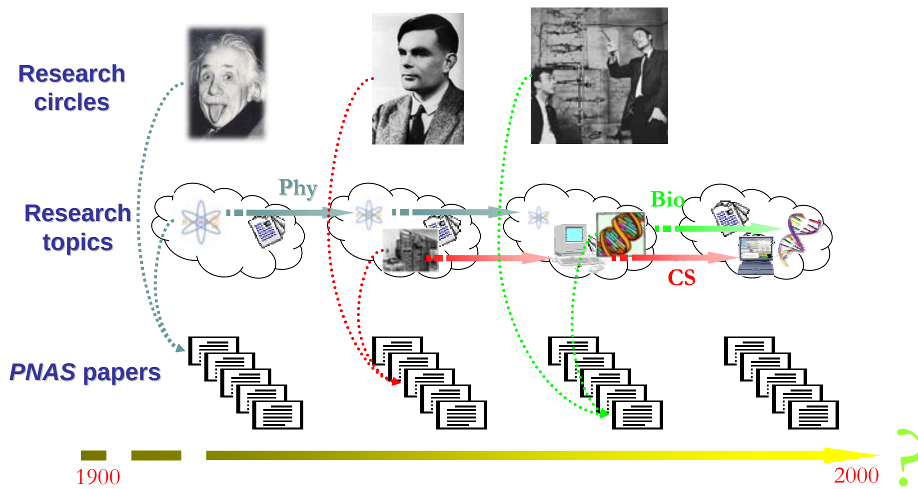
Latent brain processes:



fMRI scan:



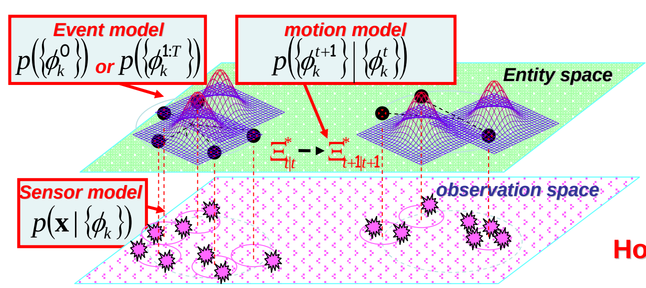
The Evolution of Science



Partially Observed, Open and Evolving Possible Worlds

- Unbounded # of objects/trajectories
- Changing attributes
- Birth/death, merge/split
- Relational ambiguity

- The parametric paradigm:

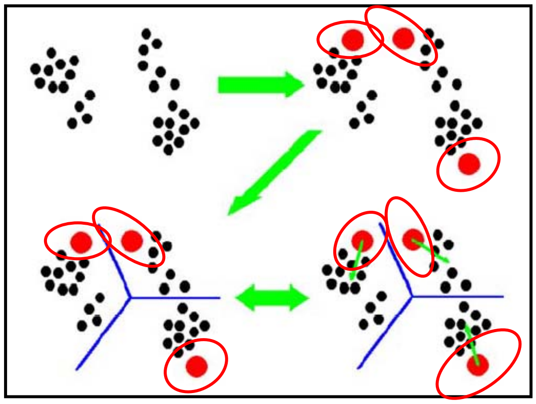


- Finite
- Structurally unambiguous

How to open it up?

A Classical Approach

- Clustering as Mixture Modeling



- Then "model selection"

Model Selection vs. Posterior Inference

- Model selection
  - "intelligent" guess: ???
  - cross validation: data-hungry ☹
  - information theoretic:
    - AIC
    - TIC
    - MDL :

    $$\left.\begin{matrix}\\ \\ \end{matrix}\right\} \quad \arg\min KL\left(f(\cdot) \mid g(\cdot \mid \hat{\theta}_{ML}, K)\right)$$

    Parsimony, Ockam's Razor
  - Bayes factor: need to compute data likelihood

- Posterior inference:

  we want to handle uncertainty of model complexity explicitly

$$p(M \mid D) \propto p(D \mid M) p(M)$$

$$M \equiv \{\theta, K\}$$

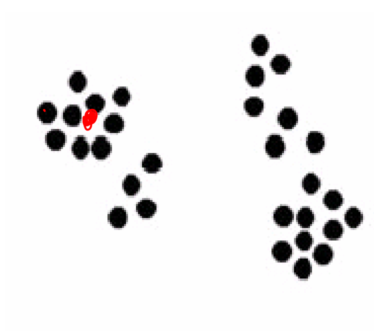
  - we favor a distribution that does not constrain M in a "closed" space!

Two "Recent" Developments

- First order probabilistic languages (FOPLs)
  - Examples: PRM, BLOG …
  - Lift graphical models to "open" world (#rv, relation, index, lifespan …)
  - Focus on complete, consistent, and operating rules to instantiate possible worlds, and formal language of expressing such rules
  - Operational way of defining distributions over possible worlds, via sampling methods

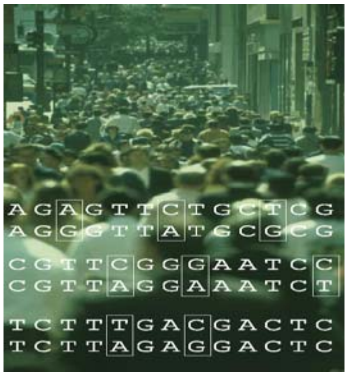
- Bayesian Nonparametrics
  - Examples: Dirichlet processes, stick-breaking processes …
  - From finite, to infinite mixture, to more complex constructions (hierarchies, spatial/temporal sequences, …)
  - Focus on the laws and behaviors of both the generative formalisms and resulting distributions
  - Often offer explicit expression of distributions, and expose the structure of the distributions --- motivate various approximate schemes

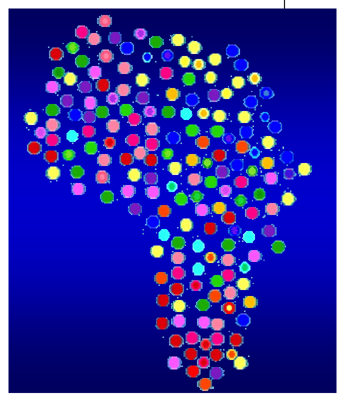
Clustering



- How to label them ?
- How many clusters ???

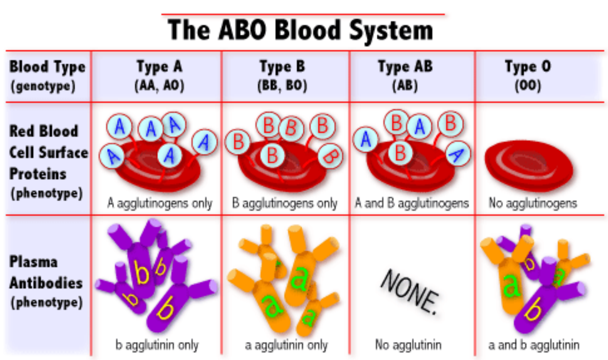
Genetic Demography





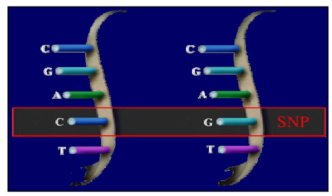
- Are there genetic prototypes among them ?
- What are they ?
- How many ? (how many ancestors do we have ?)

Genetic Polymorphisms



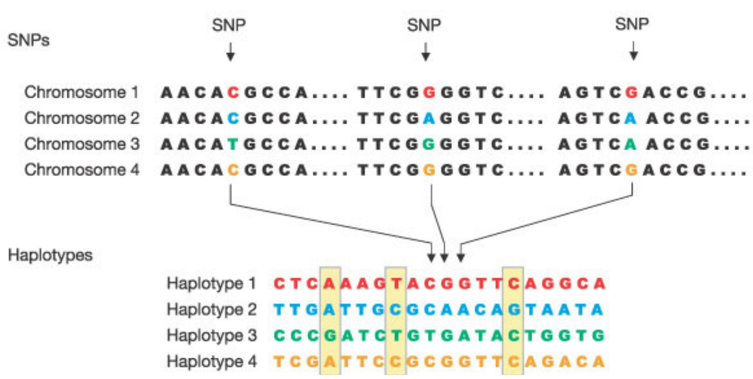
Biological Terms

- Genetic polymorphism: a difference in DNA sequence among individuals, groups, or populations
- **Single Nucleotide Polymorphism (SNP):** DNA sequence variation occurring when a single nucleotide - A, T, C, or G - differs between members of the species
  - Each variant is called an “allele”
  - Almost always bi-allelic
  - Account for most of the genetic diversity among different (normal) individuals, e.g. drug response, disease susceptibility



From SNPs to Haplotypes

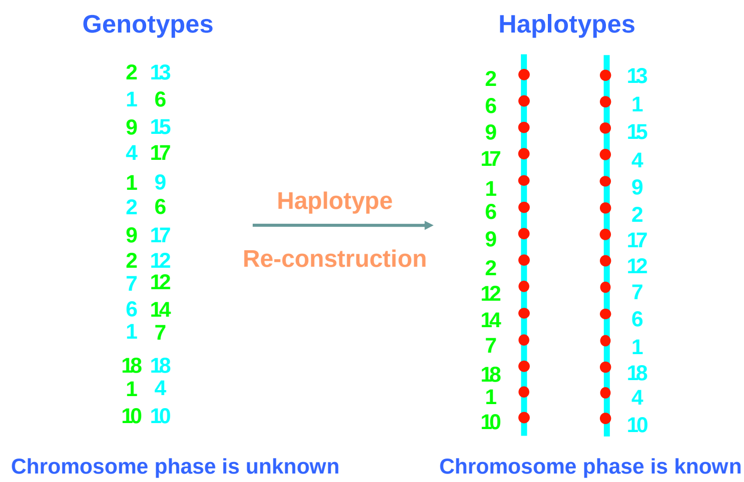
- Alleles of adjacent SNPs on a chromosome form **haplotypes**



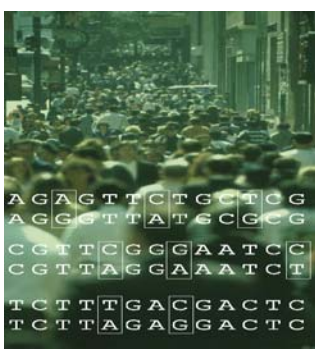
- Powerful in the study of disease association or genetic evolution

Haplotype and Genotype

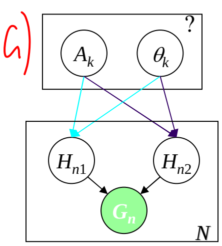
- A collection of alleles derived from the same chromosome



Ancestral Inference



$p(A\ \theta | G)$

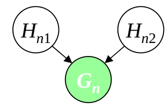


Essentially a clustering problem, but …

- Better recovery of the ancestors leads to better haplotyping results (because of more accurate grouping of common haplotypes)
- True haplotypes are obtainable with high cost, but they can validate model more subjectively (as opposed to examining saliency of clustering)
- Many other biological/scientific utilities

A Finite (Mixture of) Allele Model

- The probability of a genotype *g*:



$$p(g) = \sum_{h_1, h_2 \in \mathcal{H}} \underbrace{p(h_1, h_2)}_{\text{Haplotype model}} \underbrace{p(g \mid h_1, h_2)}_{\text{Genotyping model}}$$

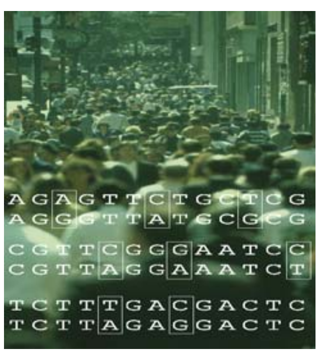
Population haplotype pool

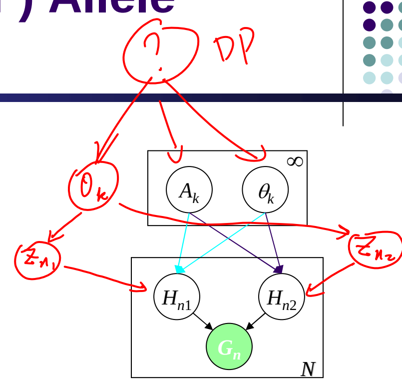
Haplotype model

Genotyping model

- Standard settings:
  - $|\mathcal{H}| = K << 2^J$ fixed-sized population haplotype pool
  - $p(h_1,h_2)= p(h_1)p(h_2)=f_1 f_2$ Hardy-Weinberg equilibrium
- Problem: K ? $\mathcal{H}$?

A Infinite (Mixture of) Allele Model





- How?
  - Via a nonparametric hierarchical Bayesian formalism !

Stick-breaking Process

v_i

$\beta_k \sim P(\cdot \mid v_i)$

$G \sim \mathrm{DP}(\alpha, G_0)$

$$G = \sum_{k=1}^{\infty} \pi_k \delta(\theta_k)$$

$$\theta_k \sim G_0$$

Location

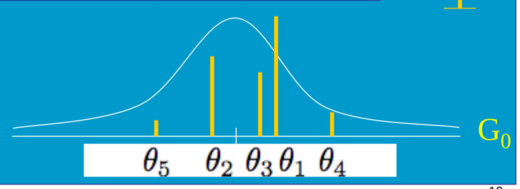
$$\sum_{k=1}^{\infty} \pi_k = 1$$

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

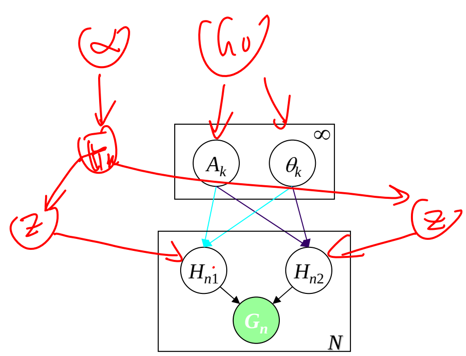
Mass

$$\beta_k \sim \mathrm{Beta}(1, \alpha)$$

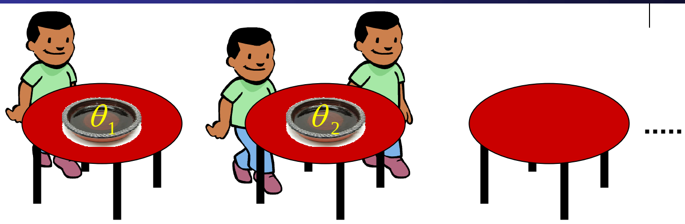
$\prod_{j=1}^{k-1}(1-\beta_j)$	β_k	π_k
0	0.4	0.4
0.6	0.5	0.3
0.3	0.8	0.24



Graphical Model



Chinese Restaurant Process



$P(c_i = k \mid \mathbf{c}_{-i}) =$

1	0		0
$\frac{1}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$		0
$\frac{1}{2+\alpha}$	$\frac{1}{2+\alpha}$		$\frac{\alpha}{2+\alpha}$
$\frac{1}{3+\alpha}$	$\frac{2}{3+\alpha}$		$\frac{\alpha}{3+\alpha}$
$\frac{m_1}{i+\alpha-1}$	$\frac{m_2}{i+\alpha-1}$	....	$\frac{\alpha}{i+\alpha-1}$

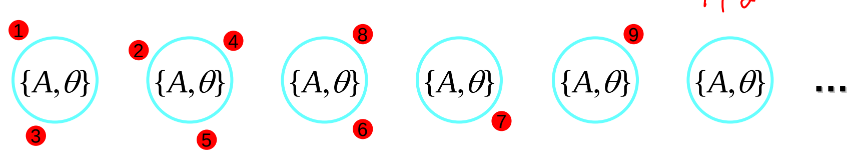
CRP defines an exchangeable distribution on partitions over an (infinite) sequence of samples, such a distribution is formally known as the Dirichlet Process (DP)

The DP Mixture of Ancestral Haplotypes

N

- The customers around a table form a cluster
  - associate a mixture component (*i.e.*, a population haplotype) with a table
  - sample $\{a, \theta\}$ at each table from a base measure G_0 to obtain the population haplotype and nucleotide substitution frequency for that component

$\frac{\alpha}{i+\alpha}$

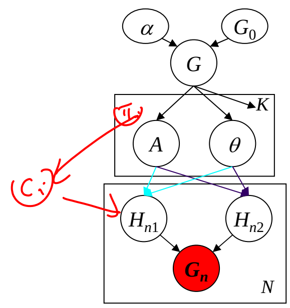


  - With $p(h|\{A, \theta\})$ and $p(g|h_1,h_2)$, the CRP yields a posterior distribution on the number of population haplotypes (and on the haplotype configurations and the nucleotide substitution frequencies)

DP-haplotyper

$\frac{m_i}{i+\alpha}$

$\frac{\alpha}{i+\alpha}$



DP

infinite mixture components
(for population haplotypes)

Likelihood model
(for individual
haplotypes and genotypes)

- Inference: Markov Chain Monte Carlo (MCMC)
  - Gibbs sampling
  - Metropolis Hasting

Model components

- Choice of base measure:

$$G_0 \sim \mathrm{Unif}(a) \cdot \prod_j \mathrm{Beta}(\theta_j)$$

- Nucleotide-substitution model:

$$p(h_i \mid \{a,\theta\}_k) = \prod_j p(h_{i,j} \mid a_{k,j}, \theta_{k,j})$$

$$\text{where} \quad p(h_{i,j} \mid a_{k,j}, \theta_{k,j}) = \begin{cases} \theta_{k,j} & \text{if} \quad h_{i,j} = a_{k,j} \\ 1-\theta_{k,j} & \text{if} \quad h_{i,j} = a_{k,j} \end{cases}$$

- Noisy genotyping model:

$$p(g_i \mid h_{i_1}, h_{i_2}) = \prod_j p(g_{i,j} \mid h_{i_1,j}, h_{i_2,j})$$

$$\text{where} \quad p(g_{i,j} \mid h_{i_1,j}, h_{i_2,j}) = \begin{cases} \gamma & \text{if} \quad h_{i_1,j} \oplus h_{i_2,j} = g_{i,j} \\ \frac{1-\gamma}{2} & \text{if} \quad h_{i_1,j} \oplus h_{i_2,j} \neq g_{i,j} \end{cases}$$

Gibbs sampling

Starting from some initial haplotype reconstruction $H^{(0)}$, pick a first table with an arbitrary $a_1^{(0)}$, and form initial population-hap pool $\mathbf{A}^{(0)} = \{a_1^{(0)}\}$:

i) Choose an individual i and one of his/her two haplytopes t, uniformly and at random, from all ambiguous individuals;

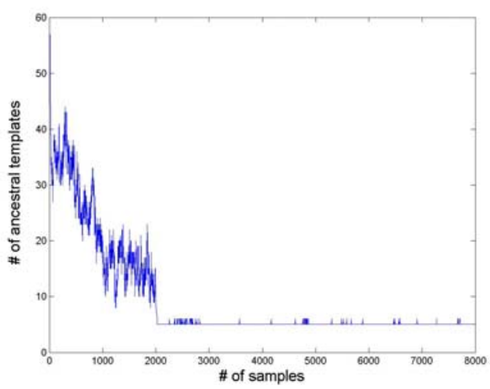
ii) Sample $c_{i_t}^{(t+1)}$ from $p(c_{i_t}^{(t+1)} \mid c_{-i_t}^{(t)}, H^{(t)}, \mathbf{A}^{(t)})$, update $c^{(t+1)}$;

$\propto P_{CRP}(C_i)P(H_i \mid A_{c_i})$

iii) Sample $a_k^{(t+1)}$, where $k = c_{i_t}^{(t+1)}$, from $p(a_k^{(t+1)} \mid \forall h_{-i'_{t'}}^{(t)} \text{ s.t. } c_{i'_{t'}}^{(t+1)} = k)$; update $\mathbf{A}^{(t+1)}$;

iii) Sample $h_{i_t}^{(t+1)}$ from $p(h_{i_t}^{(t+1)} \mid c_{i_t}^{(t+1)}, H_{-i_t}^{(t)}, \mathbf{A}^{(t+1)})$, update $H^{(t+1)}$.

Convergence of Ancestral Inference



Haplotyping Error

The Gabriel data