Topic models – to infinity and beyond!
6.874 Lecture 6

David Gifford
February 16, 2010
Topic models

Topic model

\[ p(d) = \prod_{w \in d} \left( \sum_{k \in \text{topics}} \theta_{w|k} \pi_k \right)^{n_w(d)} \]
How is the genome sequence interpreted by these regulators?
Chromatin and Nucleosome Organization


Green -H3, yellow - H4, red - H2A, pink - H2B.
Dark and light blue - DNA

Nucleosome

DNA - 146 base pairs, wrapped 1.7 times in a left-handed superhelix

Proteins - two copies of each Histones H2A, H2B, H3 and H4. Higher organisms have linker H1 histone

Histone variants

H3 variants:  
- H3.3 - transcribed
- CENP-A - centromeres

H2A variants:  
- H2A.X - DNA damage
- macroH2A - X chromosome
- H2A.Z - transcribed regions
Histone marks encode state

- Activation
- Repression
- Methylation

146 bp

Image: http://www.gtsm.com/
Histone Tail Modifications

Sims III et al., 2003
How many topics are required to represent this chromatin state information?

<table>
<thead>
<tr>
<th></th>
<th>H3K4me3</th>
<th>H3K27me3</th>
<th>H3K79me2</th>
<th>Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>GENE1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GENE2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GENE3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GENE4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GENE5</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GENE6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GENE7</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GENE8</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>GENE9</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>GENE10</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>GENE11</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GENE12</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Topic models and Mixture models

Topic model

\[ p(d) = \prod_{w \in d} \left( \sum_{k \in \text{topics}} \theta_{w|k} \pi_k \right)^{n_w(d)} \]
Topic models and Mixture models

Topic model

\[ p(d) = \prod_{w \in d} \left( \sum_{k \in \text{topics}} \theta_{w|k} \pi_k \right)^{n_w(d)} \]

Mixture model

\[ p(d) = \sum_{k \in \text{components}} \pi_k \left( \prod_{w \in d} \theta_{w|k}^{n_w(d)} \right) \]
Topic models and Mixture models

Topic model – mixture indicator selected once per word

\[ p(d) = \prod_{w \in d} \left( \sum_{k \in \text{topics}} \theta_{w|k} \pi_k \right)^{n_w(d)} \]

Mixture model – mixture indicator is selected once per document

\[ p(d) = \sum_{k \in \text{components}} \pi_k \left( \prod_{w \in d} \theta_{w|k}^{n_w(d)} \right) \]
Samples from a Dirichlet Process Mixture of Gaussians

Notice that more structure (clusters) appear as you draw more points.

(figure inspired by Neal)

http://learning.eng.cam.ac.uk/zoubin/
Non-parametric methods

- Non-parametric does not mean no parameters
- Non-parametric means that the number of parameters grows in response to the complexity of observed data
- The exemplar based affinity propagation algorithm we discussed is an example of a non-parametric method
- Model complexity is determined in part by observed data
- “Hyper parameters” can be used to set priors that influence model complexity
The Chinese Restaurant Process (CRP)

- A random process in which $n$ customers sit down in a Chinese restaurant with an infinite number of tables
  - first customer sits at the first table
  - $m$th subsequent customer sits at a table drawn from the following distribution:

\[
\begin{align*}
P(\text{previously occupied table } i \mid F_{m-1}) & \propto n_i \\
P(\text{the next unoccupied table } \mid F_{m-1}) & \propto \alpha_0
\end{align*}
\]  

(1)

where $n_i$ is the number of customers currently at table $i$ and where $F_{m-1}$ denotes the state of the restaurant after $m - 1$ customers have been seated
The Chinese Restaurant Process (CRP)

- Data points are customers; tables are clusters
  - the CRP defines a prior distribution on the partitioning of the data and on the number of tables

- This prior can be completed with:
  - a likelihood—e.g., associate a parameterized probability distribution with each table
  - a prior for the parameters—the first customer to sit at table $k$ chooses the parameter vector for that table ($\phi_k$) from the prior

- So we now have a distribution—or can obtain one—for any quantity that we might care about in the clustering setting
Infinite mixture model – definitions

\( n \)  number of observations
\( y_i \)  observation \( i \)
\( c_i \)  component of observation \( i \) (latent)
\( c_{-i} \)  components of all observations except \( i \) (latent)
\( n_{-i,j} \)  number of observations assigned to component \( j \) not counting observation \( I \)
\( \theta_j \)  parameters of component \( j \)

We do not define the number of components; there is potentially an infinite number of them
Prior on component membership

\[ p(c_i = j \mid c_{-i}, \alpha) \propto \frac{n_{-i,j}}{n-1+\alpha} \]

\[ p(c_i = \text{new} \mid c_{-i}, \alpha) \propto \frac{\alpha}{n-1+\alpha} \]
Conditional indicator posteriors

\[ p(c_i = j \mid c_{-i}, \alpha, \theta_j, y_i) \propto p(y_i \mid \theta_j, c_i = j) p(c_i = j \mid c_{-i}, \alpha) \]

\[ p(c_i = \text{new} \mid c_{-i}, \alpha, \beta, y_i) \]

\[ \propto p(c_i = \text{new} \mid c_{-i}, \alpha) \int_\theta p(y_i \mid \theta) p(\theta \mid \beta) d\theta \]

\[ \propto p(c_i = \text{new} \mid c_{-i}, \alpha) p\left(y_i \mid \theta_{\text{sampled}}\right) \]
Using MCMC sampling to find a solution

- Assign all observations to one component
- For each sweep of the sampler
  - For each observation sample its posterior assignment. This may change its component assignment, possibly creating or deleting a component.
  - Update the parameter vector for all occupied components
- Iterate the sampler until the data likelihood stabilizes after some number of iterations (burn in)
- Take record samples at intervals
- Output consensus of samples
50 observations sampled from 4 components
400 observations sampled from 4 components
From the infinite to the hierarchical…

• We have seen how to create infinite models using the Chinese Restaurant Process (CRP)
• Now let’s explore how we can build interpretable models with hierarchical structure
Plates

• A plate is a “macro” that allows subgraphs to be replicated:

• Shading denotes observed variables; i.e., conditioning

• Note that this graph represents the following marginal probability for the observations \((x_1, x_2, \ldots, x_N)\):

\[
p(x_1, x_2, \ldots, x_N) = \int \left( \prod_{i=1}^{N} p(x_i \mid \theta) \right) dP(\theta)
\]
Finite Mixture Model

\[ G = \sum_{k=1}^{K} \pi_k \delta_{\phi_k} \]
\[ \theta_i \sim G \]
\[ x_i \sim p(\cdot | \theta_i) \]
Bayesian Finite Mixture Model

\[
\begin{align*}
\phi_k & \sim G_0 \\
\pi_k & \sim \text{Dir}(\alpha_0/K, \ldots, \alpha_0/K) \\
G & = \sum_{k=1}^{K} \pi_k \delta_{\phi_k} \\
\theta_i & \sim G \\
x_i & \sim p(\cdot | \theta_i)
\end{align*}
\]

Note that \(G\) is now a random measure

www.cs.berkeley.edu/~jordan/nips-tutorial05.ps
Infinite Mixture Model

www.cs.berkeley.edu/~jordan/nips-tutorial05.ps
Dirichlet process mixture model

\[ G \sim \text{DP}(\alpha G_0) \]
\[ \theta_i | G \sim G \quad i \in 1, \ldots, n \]
\[ x_i | \theta_i \sim F(x_i | \theta_i) \quad i \in 1, \ldots, n \]
Chinese restaurant franchises (CRF)

- To each group there corresponds a restaurant, with an unbounded number of tables in each restaurant
- There is a global menu with an unbounded number of dishes on the menu
- The first customer at a table selects a dish for that table from the global menu
- Reinforcement effects—customers prefer to sit at tables with many other customers, and prefer to choose dishes that are chosen by many other customers
Hierarchical Dirichlet Process

\[ G_0 | \gamma, H \sim \text{DP}(\gamma H) \]
\[ G_i | \alpha, G_0 \sim \text{DP}(\alpha_0 G_0) \]
\[ \theta_{ij} | G_i \sim G_i \]
\[ x_{ij} | \theta_{ij} \sim F(x_{ij}, \theta_{ij}) \]
Collapsing a hierarchical Dirichlet process (HDP)

- First integrate out the $G_i$, then integrate out $G_0$
Topic models can be described by HDPs

(Blei, Ng, & Jordan, 2003)

- **Random variables:**
  - A *word* is represented as a *multinomial* random variable $w$
  - A *topic* is represented as a *multinomial* random variable $z$
  - A *document* is represented as a *Dirichlet* random variable $\theta$

- **Plates:**
  - *repeated* sampling of Dirichlet document variable within corpus
  - *repeated* sampling of multinomial topic variable within documents

www.cs.berkeley.edu/~jordan/nips-tutorial05.ps
Hierarchical topic models

- Each node in the tree is a Chinese restaurant

- Each table in every restaurant has an associated distribution on words (a "topic") drawn from a prior

- Sitting at a table in a given restaurant also selects an outgoing branch, which provides access to further restaurants and further topics
  - we hope for more specialized topics to emerge as we descend the tree

www.cs.berkeley.edu/~jordan/nips-tutorial05.ps
Example from Psychology Today

response; stimulus; reinforcement
speech; reading; words
action; social; self
group; iq; intelligence
hippocampus; growth; hippocampal
numerals; catastrophe; stream
rod; categorizer; child

sex; emotions; gender
reasoning; attitude; consistency
self; social; psychology
 genetic; scenario; adaptations

color; image; monocular
 motion; visual; binocular
conditioning; stress; behavioral
 drug; food; brain
GeneProgram model review
Background: topic models

- **topic** = probability distribution over word frequencies

- **document groups** = shared mixtures of topics

- **documents** = bags of words = mixtures of topics

**Diagram:**
- **Root:**
  - DNA
  - RNA
  - Membrane
  - Mitochondria
  - ...

- **Document Groups:**
  - Biology
  - Medicine
  - Biophysics
  - Physics
  - Romance

- Each group contains documents with histograms indicating word frequencies.
GeneProgram is based on topic models
GeneProgram extends the Hierarchical Dirichlet Process topic model

distributions over change patterns (+ = induction, - = repression)

nodes = Dirichlet Processes (mixtures over programs)

tissues = mixtures of expression programs

expression program = probabilities of expressing genes (and temporal pattern)

tissue picks how to use program

tissues = expressions of tissues

expression data magnitudes pattern types

select genes for expression in tissue

selected genes pattern

select genes expression program

APOH APOA1 APOA2 ITIH3 ALB
The GeneProgram algorithm steps involve preprocessing, inference and summarization.

1) combine replicates and normalize
2) discretize
3) model inference
4) summarize inference/produce map
Synthetic microarray data so could compare GeneProgram’s performance to other algorithms

- 150 genes in 4 sets of 40 (10 in sets 3 and 4 overlap)
- 40 tissues equally among 4 groups
- Not generated from Gene-Program itself
Hierarchical clustering

Only GeneProgram was capable of recovering all 4 gene sets from the synthetic data

Singular Value Decomposition (SVD)

Non-negative Matrix Factorization (NMF)

GeneProgram w/o tissue groups

Full GeneProgram model
GeneProgram outperformed popular biclustering algorithms in discovery of biologically meaningful gene sets from real microarray data.

<table>
<thead>
<tr>
<th>Data source</th>
<th>Algorithm</th>
<th>Gene dimension (GO category enrichment)</th>
<th>Tissue dimension (manually derived category enrichment)</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>GeneProgram</td>
<td>93%</td>
<td>76%</td>
</tr>
<tr>
<td>N</td>
<td>NMF</td>
<td>35%</td>
<td>29%</td>
</tr>
<tr>
<td>N</td>
<td>Samba</td>
<td>53%</td>
<td>9%</td>
</tr>
<tr>
<td>S</td>
<td>GeneProgram</td>
<td>66%</td>
<td>53%</td>
</tr>
<tr>
<td>S</td>
<td>NMF</td>
<td>28%</td>
<td>19%</td>
</tr>
<tr>
<td>S</td>
<td>Samba</td>
<td>51%</td>
<td>28%</td>
</tr>
</tbody>
</table>

N = Novartis Tissue Atlas v2 (141 mouse and human tissues)
S = Shyamsundar et al. (115 human tissues)
GeneProgram application:

temporal map of expression programs in human infections

• Compendium 62 short gene expression time-series:
  – 6 publications; 347 experiments
  – different durations, sampling rates
• Primary and cell-culture derived human cells:
  – macrophages (primary and cultured)
  – peripheral blood mononuclear cells [PBMCs]
  – epithelial cells
  – whole blood
• Exposed to infectious agents and immune-modulating substances
Each gene in each time-series → 1 of 6 simple temporal patterns (time earliest induction/repression)
GeneProgram produced a map of 5 tissue groups and 104 expression programs in the infection data.

<table>
<thead>
<tr>
<th>A</th>
<th>m0 control1</th>
<th>70</th>
<th>77</th>
<th>79</th>
<th>90</th>
<th>95</th>
<th>98</th>
<th>101</th>
<th>102</th>
<th>104</th>
</tr>
</thead>
<tbody>
<tr>
<td>m0 control2</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
</tr>
<tr>
<td>m0 control3</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
</tr>
<tr>
<td>m0 control4</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
</tr>
<tr>
<td>m0 LPS</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
</tr>
<tr>
<td>m0 EHEC</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
</tr>
<tr>
<td>m0 S. typhimurium</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
</tr>
<tr>
<td>m0 S. aureus</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
</tr>
<tr>
<td>m0 L. monocytogenes</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
</tr>
<tr>
<td>m0. M. tuberculosis</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
</tr>
<tr>
<td>m0 M. bosei DCG</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
</tr>
<tr>
<td>m0 LPS E</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
</tr>
<tr>
<td>m0 LPS S</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
<td>M+</td>
</tr>
<tr>
<td>m0 LTA</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
</tr>
<tr>
<td>m0 MrpB</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
</tr>
<tr>
<td>m0 TB hag72</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
</tr>
<tr>
<td>m0 BGC tag65</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
</tr>
<tr>
<td>m0 IFNα</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
<td>E+</td>
</tr>
</tbody>
</table>

GeneProgram produced a map of 5 tissue groups and 104 expression programs in the infection data.
GeneProgram produced a map of 5 tissue groups and 104 expression programs in the infection data.
GeneProgram produced a map of 5 tissue groups and 104 expression programs in the infection data (sorted by generality score).
GeneProgram produced a map of 5 tissue groups and 104 expression programs in the infection data.
Global map analysis: discovered expression programs significantly enriched for 1) biological processes and pathways; 2) binding by the key transcriptional regulator NF-κB

- 50% significantly enriched for GO categories; 59% for KEGG pathways
  - many processes/pathways associated w/ infection
  - some surprising pathways/processes
- Programs significantly enriched for NF-κB binding = high generality scores ("common response")
Low generality example: epithelial cells exposed to *H. pylori* activate cytoskeletal regulation pathways

<table>
<thead>
<tr>
<th>Host-cell type</th>
<th>Agent</th>
<th>Temporal pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gastric epithelial</td>
<td><em>H. pylori</em> bacteria</td>
<td><img src="image" alt="temporal pattern" /></td>
</tr>
</tbody>
</table>

Significantly enriched pathway/process: *Regulation of actin cytoskeleton*
Genes from program #22 are found throughout the regulation of actin cytoskeleton signaling pathway.

*red* = protein coded for by gene in program #22
Several intermediate generality programs implicated surprising pathways or receptor types in the response to infection.
Major contributions: GeneProgram

- New framework - maps of gene expression programs from large compendia human expression data
- Extended Hierarchical Dirichlet Process topic models - group uncertainty and patterns of change in data
- Discovered novel biology in human response to infection