Applications of admixture models
Fall 2016

Sriram Sankararaman
Acknowledgments: Fei Sha, Ameet Talwalkar, Alkes Price
Outline

Admixture models

Population structure and GWAS
Mixture model for genetic data

- $K$: the number of populations (mixture components)
- $\pi_k$: mixture weights – they represent how much each population contributes to the final distribution
- $f_{m,k}$: allele frequency in each of $K$ populations.

$$p(x, z) = p(z)p(x|z)$$
Mixture model for genetic data

Denote

\[ p(z = k | \pi) = \prod_{k=1}^{K} \pi_k 1\{z=k\} \]

Now, assume the conditional distributions are independent Binomial.

\[ p(x | z = k, (f_{m,k})_{m=1}^{M}) = \prod_{m} p(x_m | z = k) = \prod_{m} \text{Bin}(2, f_{m,k}) \]

Then, the marginal distribution of \( x \) is

\[ p(x | \theta) = \sum_{k=1}^{K} p(z = k | \pi)p(x | z = k, f_k) \]

The parameters \( \theta = (\pi_k, f_k)_{k=1}^{K} \).
Mixture model for genetic data

- Given $N$ individuals over $M$ SNPs: $x_n, n \in \{1, \ldots, N\}$, write the log likelihood $LL(\theta)$.
- Estimate the maximum likelihood parameters $\hat{\theta}$ using EM.
Supervised mixture models

| Allele frequency | POP1   | 0.25 | 0.57 | 0.29 | 0.38 |
|                 | POP2   | 0.40 | 0.32 | 0.84 | 0.22 |

| Individual x    | 2 0 1 1 |

Does individual x belong to population 1 or 2?

\[
P(Data|x \text{ is in population 1}) = (0.25)^2(0.75)^0(0.57)^0(0.43)^2 \ldots = 0.0006
\]

\[
P(Data|x \text{ is in population 2}) = (0.40)^2(0.60)^0(0.32)^0(0.68)^2 \ldots = 0.0017
\]
Mixture model for genetic data: Example

### Supervised mixture models

<table>
<thead>
<tr>
<th>SNPs</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>POP1</td>
<td>0.25</td>
<td>0.57</td>
<td>0.29</td>
<td>0.38</td>
</tr>
<tr>
<td>POP2</td>
<td>0.40</td>
<td>0.32</td>
<td>0.84</td>
<td>0.22</td>
</tr>
</tbody>
</table>

Allele frequency

| Individual x | 2   | 0   | 1   | 1   |

Does individual x belong to population 1 or 2?

\[
P(Data|x \text{ is in population 1})
= (0.25)^2(0.75)^0(0.57)^0(0.43)^2 \ldots = 0.0006
\]

\[
P(Data|x \text{ is in population 2})
= (0.40)^2(0.60)^0(0.32)^0(0.68)^2 \ldots = 0.0017
\]
Unsupervised mixture models

- What if allele frequencies are not known?
- Use EM to infer parameters (HW problem).
• Clustering: sample belongs to exactly one cluster.

• In genetics:
  • Cluster $\equiv$ population
  • Individuals could belong to more than one population.
Admixture models

- Individual can now have fractional memberships in each population.
- Each SNP can have different ancestry.
• Admixed population is one that has ancestry from multiple distinct populations.
Admixture better reflects human biology

Black? White? Asian? More Young Americans Choose All of the Above

From left: Shannon Palmer, Japanese/Irish; Vasco Mateus, Portuguese/African-American/Haitian; Laura Wood, Black/White. More Photos

By SUSAN SAULNY
Published: January 29, 2011
Admixture better reflects human biology

A genetic atlas of human admixture history


Applications of admixture models

Hellenthal et al. Science 2014
Examples of admixed populations

- **African-Americans:**
  - African and European ancestry.
  - 10% of US population
- **Latino Americans (Hispanics):**
  - European, Native American and African
  - 15% of US population
  - Mexican Americans, Puerto Ricans
- **Hawaiians**
- **South Asians**
- **Middle Easterners**
Admixture and ancestry
PCA on genetic data

CHB+JPT

CEU

YRI

CHB

JPT

AA

AA: 21% ± 14%
European ancestry
Admixture leads to variation in proportions of European ancestry in African American populations.
PCA on HapMap Phase 3
PCA on HapMap Phase 3

Applications of admixture models

Admixture models
Each individual $n$ has a parameter $g_n = (g_{n,1}, \ldots, g_{n,K})$ where $g_{n,k} \geq 0$ and $\sum_k g_{n,k} = 1$. Each population has a parameter for a SNP $f_k = (f_{1,k}, \ldots, f_{M,k})$.

\[
z_{n,m,l} | g_n \sim \text{Mult}(g_n), \quad l \in \{1, 2\}
\]
\[
x_{n,m} | z_{n,m,1}, z_{n,m,2}, f_k \sim \text{Ber} (f_{m,z_{n,m,1}}) + \text{Ber} (f_{m,z_{n,m,2}})
\]
Inference in the admixture model

- Parameters: \( \theta = (g_n, f_k) \).
- Use EM to estimate parameters.

  - E-step: Compute \( r_{n,m,a,b}^{(t)} \equiv p(z_{n,m} = (a, b) | x_{n,m}, g_n^{(t)}, f_{m,k}^{(t)}) \).
  - M-step: Update estimates of the parameters. **Work out the updates!**
• Parameters: \( \theta = (g_n, f_k) \).
• Use EM to estimate parameters.
• E-step: Compute \( r_{n,m,a,b}^{(t)} \equiv p(z_{n,m} = (a, b) | x_{n,m}, g_{n}^{(t)}, f_{m,k}^{(t)}) \).
• M-step: Update estimates of the parameters. Work out the updates!
Supervised admixture models

<table>
<thead>
<tr>
<th>Allele frequency</th>
<th>POP1</th>
<th>POP2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.25</td>
<td>0.40</td>
</tr>
<tr>
<td>2</td>
<td>0.57</td>
<td>0.32</td>
</tr>
<tr>
<td>3</td>
<td>0.29</td>
<td>0.84</td>
</tr>
<tr>
<td>4</td>
<td>0.38</td>
<td>0.22</td>
</tr>
</tbody>
</table>

Individual x has ancestry $\alpha$ from population 1 and $(1 - \alpha)$ from population 2. Find $\alpha$.

$$P(Data|\alpha) = [0.25\alpha + 0.40(1 - \alpha)]^2[(1 - 0.25)\alpha + (1 - 0.40)(1 - \alpha)]^0$$

$$[0.57\alpha + 0.32(1 - \alpha)]^0[(1 - 0.57)\alpha + (1 - 0.32)(1 - \alpha)]^2$$

... 

Maximum value of $P$ attained at $\alpha = 0.22$. 
## Supervised admixture models

<table>
<thead>
<tr>
<th>Allele frequency</th>
<th>POP1</th>
<th>POP2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.25</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td>0.57</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>0.29</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>0.38</td>
<td>0.22</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Individual x</th>
<th>2</th>
<th>0</th>
<th>1</th>
<th>1</th>
</tr>
</thead>
</table>

Individual x has ancestry $\alpha$ from population 1 and $(1 - \alpha)$ from population 2. Find $\alpha$.

\[
P(Data|\alpha) = [0.25\alpha + 0.40(1 - \alpha)]^2[(1 - 0.25)\alpha + (1 - 0.40)(1 - \alpha)]^0
\]
\[
[0.57\alpha + 0.32(1 - \alpha)]^0[(1 - 0.57)\alpha + (1 - 0.32)(1 - \alpha)]^2
\]

Maximum value of $P$ attained at $\alpha = 0.22$. 
Applying admixture models to HGDP

Human Genome Diversity Project
Applying admixture models to HGDP

Human Genome Diversity Project

- 938 HGDP individuals (118 related individuals removed)
- 51 world populations (N. Han and S. Han merged)
- Illumina 650K chip

FRAPPE results at K=7:

Li et al. Science 2008
Admixture models outside of genetics

- Also known as topic models or LDA (Latent Dirichlet Allocation).
- Used to model topics in documents.
  - Genotypes = words
  - Individual = document
  - Population = topic
- Each document has different distributions over topics.
- Each topic specifies distribution over words.
Admixture models outside of genetics

A generalized\textsuperscript{3} fundamental\textsuperscript{146} theorem\textsuperscript{267} of natural\textsuperscript{250} selection\textsuperscript{250} is derived\textsuperscript{233} for populations\textsuperscript{250} incorporating\textsuperscript{149} both genetic\textsuperscript{250} and cultural\textsuperscript{250} transmission\textsuperscript{25}. The phenotype\textsuperscript{3} is determined\textsuperscript{17} by an arbitrary\textsuperscript{3} number\textsuperscript{257} of multiallelic\textsuperscript{3} loci\textsuperscript{3} with two\textsuperscript{271} -factor\textsuperscript{60} epistasis\textsuperscript{250} and an arbitrary\textsuperscript{149} linkage\textsuperscript{3} map\textsuperscript{3}, as well as by cultural\textsuperscript{250} transmission\textsuperscript{25} from the parents\textsuperscript{250}. Generations\textsuperscript{250} are discrete\textsuperscript{69} but partially\textsuperscript{273} overlapping\textsuperscript{146}, and mating\textsuperscript{250} may be nonrandom\textsuperscript{250} at either the genotypic\textsuperscript{250} or the phenotypic\textsuperscript{250} level\textsuperscript{199} (or both). I show\textsuperscript{25} that cultural\textsuperscript{250} transmission\textsuperscript{25} has several\textsuperscript{173} important\textsuperscript{17} implications\textsuperscript{17} for the evolution\textsuperscript{250} of population\textsuperscript{250} fitness\textsuperscript{250}, most notably\textsuperscript{230} that there is a time\textsuperscript{72} lag\textsuperscript{72} in the response\textsuperscript{213} to selection\textsuperscript{250} such that the future\textsuperscript{257} evolution\textsuperscript{250} depends\textsuperscript{105} on the past selection\textsuperscript{250} history\textsuperscript{250} of the population\textsuperscript{250}.

Griffiths and Steyvers, PNAS 2004
Outline

Admixture models

Population structure and GWAS
Population structure can lead to false discoveries

Applications of admixture models

Population structure and GWAS
Population structure can lead to false discoveries

Applications of admixture models

Population structure and GWAS
Structured association

- Cluster individuals into populations.
- Do GWAS in each population.
- Combine results.
Principal Components

- Include Principal Components in the model.
\[ n = 200 \]
\[ m = 1000 \]
\[ Z_n \in \{1, 2\} \]
\[ Z_n = 1, \; n \leq 100 \]
\[ Z_n = 2, \; n > 100 \]
\[ Y_n|Z_n \sim \begin{cases} 
\mathcal{N}(10, 1), & Z_n = 1 \\
\mathcal{N}(0, 1), & Z_n = 2 
\end{cases} \]
\[ X_{n,m}|Z_n \sim \text{Ber} \left( f_{Z_n,m} \right) \]
How well does the model fit?

True ancestry $Z$ known
How well does the model fit?

**True ancestry \( \mathcal{Z} \) unknown**

We find 222 SNPs that are statistically significant (p-value < .05/1000)
How well does the model fit?

Visualize these associations

![Box plot](image-url)
How well does the model fit?

Visualize these associations in each population

![Box plots showing phenotype distribution for different alleles in P1 and P2 populations.](image)
How well does the model fit?

Infer PCs (PC scores for first PC)
How well does the model fit?

Infer PCs (PC1 vs PC2)
How well does the model fit?

Fraction of variance explained

About 6% variance explained by PC1
How well does the model fit?

Correct for PCs

No association is significant!

Applications of admixture models
Population structure and GWAS
Summary

• PCA is an example of a latent variable model with continuous latent variable. Unlike clustering, where the latent variable is discrete.
• Probabilistic model corresponding to PCA.
• Admixture models or topic models or LDA are generalizations of clustering.
• Applications to infer ancestry and correct for population structure.
• Question: When do we include PCs in our regression?