Lecture 5: Population Structure Inference

Timothy Thornton and Michael Wu

Summer Institute in Statistical Genetics 2016
Background: Population Structure

- Humans originally spread across the world many thousand years ago.
- Migration and genetic drift led to genetic diversity between isolated groups.
Inference on genetic ancestry differences among individuals from different populations, or population structure, has been motivated by a variety of applications:

- population genetics
- genetic association studies
- personalized medicine
- forensics

Advancements in array-based genotyping technologies have largely facilitated the investigation of genetic diversity at remarkably high levels of detail.

A variety of methods have been proposed for the identification of genetic ancestry differences among individuals in a sample using high-density genome-screen data.
Inferring Population Structure with PCA

- Principal Components Analysis (PCA) is the most widely used approach for identifying and adjusting for ancestry difference among sample individuals.
- PCA applied to genotype data can be used to calculate principal components (PCs) that explain differences among the sample individuals in the genetic data.
- The top PCs are viewed as continuous axes of variation that reflect genetic variation due to ancestry in the sample.
- Individuals with ”similar” values for a particular top principal component will have similar ancestry for that axes.
Standard Principal Components Analysis (sPCA)

- sPCA is an unsupervised learning tool for dimension reduction in multivariate analysis.
- Widely used in genetics community to infer population structure from genetic data.
  - Belief that top principal components (PCs) will reflect population structure in the sample.
- Orthogonal linear transformation to a new coordinate system
  - Sequentially identifies linear combinations of genetic markers that explain the greatest proportion of variability in the data
  - These define the axes (PCs) of the new coordinate system
  - Each individual has a value along each PC
- EIGENSOFT (Price et al. 2006) is a popular implementation of PCA.
Data Structure

- Sample of \( n \) individuals, indexed by \( i = 1, 2, \ldots, n \).
- Genome screen data on \( m \) genetic autosomal markers, indexed by \( l = 1, 2, \ldots, m \).
- At each marker, for each individual, we have a genotype value, \( x_{il} \).
- Here we consider SNP data, so \( x_{il} \) takes values 0, 1, or 2, corresponding to the number of reference alleles.
- We center and standardize these genotype values:

\[
z_{il} = \frac{x_{il} - 2\hat{p}_l}{\sqrt{2\hat{p}_l(1 - \hat{p}_l)}}
\]

where \( \hat{p}_l \) is an estimate of the reference allele frequency for marker \( l \).
Genetic Correlation Estimation

▶ Create an \( n \times m \) matrix, \( Z \), of centered and standardized genotype values, and from this, a genetic correlation matrix (GRM):

\[
\hat{\Psi} = \frac{1}{m}ZZ^T
\]

▶ \( \hat{\Psi}_{ij} \) is an estimate of the genome wide average genetic correlation between individuals \( i \) and \( j \).

▶ sPCA relies on individuals from the same ancestral population being more genetically correlated than individuals from different ancestral populations.
Standard Principal Components Analysis (sPCA)

- PCA is performed by obtaining the eigendecomposition $\hat{\Psi}$
- Orthogonal axes of variation, i.e. linear combinations of SNPs, that best explain the genotypic variability amongst the $n$ sample individuals are identified.
- The result is:
  - a set of $n$ length $n$ eigenvectors, $(V_1, V_2, \ldots V_n)$, where $V_d$ is a column vector of coordinates of each individual along axis $d$
  - and a corresponding set of $n$ eigenvalues, $(\lambda_1 > \lambda_2 > \ldots > \lambda_n)$, in decreasing order.
  - The $d^{th}$ principal component (eigenvector) corresponds to eigenvalue $\lambda_d$, where $\lambda_d$ is proportional to the percentage of variability in the genome-screen data that is explained by $V_d$.
- These eigenvectors (PCs) are used as surrogates for population structure
The top principal components are viewed as continuous axes of variation that reflect genetic variation due to ancestry in the sample.

Individuals with "similar" values for a particular top principal component will have "similar" ancestry for that axes.

A application of principal components to genetic data from European samples (Novembre et al., Nature 2008) showed that among Europeans for whom all four grandparents originated in the same country, the first two principal components computed using 200,000 SNPs could map their country of origin quite accurately in the plane.
PCA of Europeans

[Diagram showing a principal component analysis (PCA) plot of European populations, with countries represented by colored dots along the PC1 and PC2 axes.]
There can be population structure in all populations, even those that appear to be relatively "homogenous"

An application of principal components to genetic data from Finland samples (Sabatti et al., 2009) identified population structure that corresponded very well to geographic regions in this country.
PCA in Finland
Relatedness Confounds sPCA

- Recall that the GRM used by sPCA, $\hat{\Psi}_{ij}$, is an estimate of the genome wide average genetic correlation between individuals $i$ and $j$.
- It can be shown:

$$\Psi_{ij} = 2[\phi_{ij} + (1 - \phi_{ij})A_{ij}]$$

- $\phi_{ij}$: kinship coefficient - a measure of familial relatedness
- $A_{ij}$: a measure of ancestral similarity
- PCA is an unsupervised method; in related samples we don’t know the correlation structure each eigenvector is reflecting
  - If the only genetic correlation structure among individuals is due to ancestry, $\Psi$ and the top PCs will capture this.
  - If there is relatedness in the sample, the top PCs may reflect this or some combination of ancestry and relatedness.
- Association studies have known or cryptic relatedness!
The PC-AiR method was developed for performing a Principal Components Analysis in Related samples. The algorithm has the following steps:
The PC-AiR Algorithm

1. Estimate the relatedness of all pairs of individuals in the sample.
2. Partition the sample into an unrelated set and a related set of individuals.
3. Perform standard PCA on the set of unrelated individuals.
4. Predict PC values for the set of related individuals based on genetic similarities with the unrelated set.
Admixed Populations

- Several recent and ongoing genetic studies have focused on **admixed populations**: populations characterized by ancestry derived from two or more ancestral populations that were reproductively isolated.

- Admixed populations have arisen in the past several hundred years as a consequence of historical events such as the transatlantic slave trade, the colonization of the Americas and other long-distance migrations.

- Examples of admixed populations include:
  - African Americans and Hispanic Americans in the U.S
  - Latinos from throughout Latin America
  - Uyghur population of Central Asia
  - Cape Verdeans
  - South African ”Coloured” population
The chromosomes of an admixed individual represent a mosaic of chromosomal blocks from the ancestral populations.
Admixed Populations

- Can be substantial genetic heterogeneity among individuals in admixed populations
- Admixed populations are ancestrally admixed and thus have population structure.
- Statistical method for estimating admixture proportions using genetic data are available
Supervised Learning for Ancestry Admixture

Methods, such as ADMIXTURE and FRAPPE, have recently been developed for supervised learning of ancestry proportions for an admixed individuals using high-density SNP data.

Most use either a hidden Markov model (HMM) or an Expectation-Maximization (EM) algorithm to infer ancestry.

Example: Suppose we are interested in identifying the ancestry proportions for an admixed individual.

Observed sequence on a chromosome for an admixed individual:

…TATACGTGCACCTGATTACAGATTACAGATTACAGATTACAGATTACATTCATCGATCGAA…

Observed sequence on a chromosome for samples selected from a “homogenous” reference population:

…TGATCCTGAACCTAGATTACAGATTACAGATTACAGATTACAGATTACATTCATCGATCGAA…

…AGATCCTGAACCTAGATTACAGATTACAGATTACAGATTACAGATTACATTCATCGATCGAA…

…CGATCCTGAACCTAGATTACAGATTACAGATTACATTCATCGATCGATCGAA…
HapMap ASW and MXL Ancestry

- Genome-screen data on 150,872 autosomal SNPs was used to estimate ancestry
- Estimated genome-wide ancestry proportions of every individual using the ADMIXTURE (Alexander et al., 2009) software
- A supervised analysis was conducted using genotype data from the following reference population samples for three "ancestral" populations
  - HapMap YRI for West African ancestry
  - HapMap CEU samples for northern and western European ancestry
  - HGDP Native American samples for Native American ancestry.
Supervised ADMIXTURE Estimated Ancestry for HapMap MXL + ASW

Unsupervised ADMIXTURE Estimated Ancestry for HapMap MXL + ASW
## Table: Average Estimated Ancestry Proportions for HapMap African Americans and Mexican Americans

<table>
<thead>
<tr>
<th>Population</th>
<th>Estimated Ancestry Proportions (SD)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>European</td>
<td>African</td>
<td>Native American</td>
</tr>
<tr>
<td>MXL</td>
<td>49.9% (14.8%)</td>
<td>6% (1.8%)</td>
<td>44.1% (14.8%)</td>
</tr>
<tr>
<td>ASW</td>
<td>20.5% (7.9%)</td>
<td>77.5% (8.4%)</td>
<td>1.9% (3.5%)</td>
</tr>
</tbody>
</table>
**Figure: HapMap MXL + ASW Sample**

A. **Top 2 PCs with PC−AIR**

B. **Top 2 PCs with EIGENSOFT**

C. **EIGENSOFT PCs 2 and 3**

D. **EIGENSOFT PCs 4 and 5**

E. **EIGENSOFT PCs 6 and 7**

F. **EIGENSOFT PCs 8 and 9**
HapMap MXL Cryptic Relatedness
Genetic Diversity and Association Studies in US Hispanic/Latino Populations: Applications in the Hispanic Community Health Study/Study of Latinos

Matthew P. Conomos,1,14,* Cecelia A. Laurie,1,14 Adrienne M. Stilp,1,14 Stephanie M. Gogarten,1,14 Caitlin P. McHugh,1 Sarah C. Nelson,1 Tamar Sofer,1 Lindsay Fernández-Rhodes,2 Anne E. Justice,2 Mariaelisa Graff,2 Kristin L. Young,2 Amanda A. Seyerle,2 Christy L. Avery,2 Kent D. Taylor,3 Jerome I. Rotter,3 Gregory A. Talavera,4 Martha L. Daviglus,5 Sylvia Wassertheil-Smoller,6 Neil Schneiderman,7 Gerardo Heiss,2 Robert C. Kaplan,6 Nora Franceschini,2 Alex P. Reiner,8 John R. Shaffer,9 R. Graham Barr,10 Kathleen F. Kerr,1 Sharon R. Browning,1 Brian L. Browning,11 Bruce S. Weir,1 M. Larissa Avilés-Santa,12 George J. Papanicolaou,12 Thomas Lumley,13 Adam A. Szpiro,1 Kari E. North,2 Ken Rice,1 Timothy A. Thornton,1 and Cathy C. Laurie1,*

PCA-AiR: Hispanic Community Health Study
PC-AiR: Hispanic Community Health Study
Genetic differentiation among individuals is associated with the geography of their countries of grandparental origin.

Individuals for whom all four grandparents were born in a specific country in Central or South America were used.
References


References