We reviewed properties of the SVD. Currently no slides for this part of the lecture. We also saw Kaileigh's presentation on an application of principal components analysis to a problem in population genetics. Her slides come next.

Principal Components Analysis (PCA) for Population Genetics

Presented by Kaileigh Ahlquist

Goal

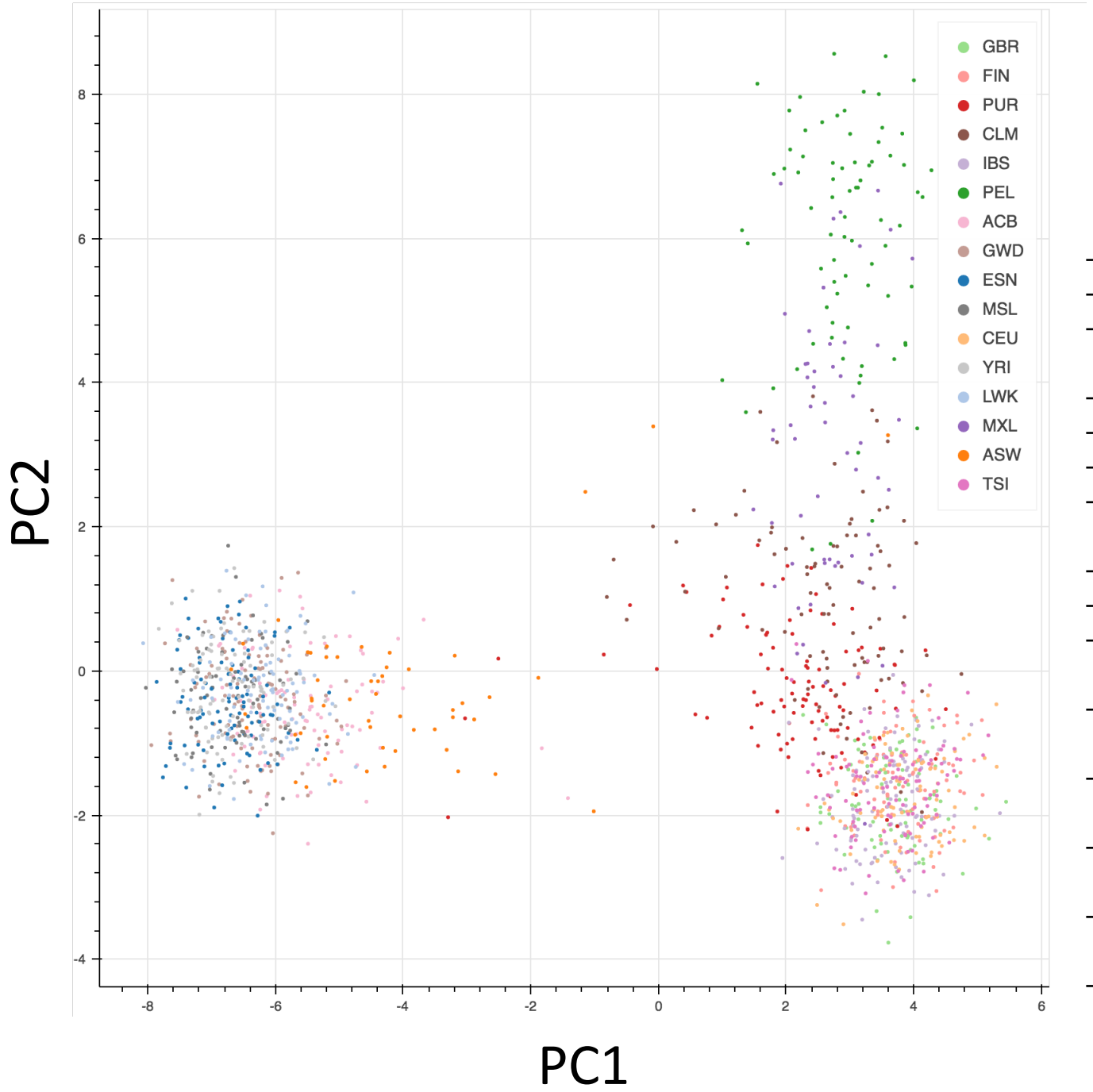
- Visualize the data in two dimensions from a perspective that reveals important aspects of population structure. May be able to predict:
  - Geographic patterns of migration, trade and travel
  - Heritage of unknown or admixed individuals

- Use the resulting principal components to filter data for further analysis, removing locations that are not informative or redundant.

PCA using SVD

```
def PCA(data):
    C = len(data.D[1])
    mn = vec_of_row_means(data)
    data = data - coldict2mat({i: mn for i in range(C)})
    Y = transpose(data) * (1 / sqrt(C))
    u, w, v = svd(Y)
    PC_dict = mat2coldict(v)
    return PC_dict
```

Results



CEU	Utah Residents (CEPH) with Northern and Western European Ancestry
TSI	Toscani in Italia
FIN	Finnish in Finland
GBR	British in England and Scotland
IBS	Iberian Population in Spain
YRI	Yoruba in Ibadan, Nigeria
LWK	Luhya in Webuye, Kenya
GWD	Gambian in Western Divisions in the Gambia
MSL	Mende in Sierra Leone
ESN	Esan in Nigeria
ASW	Americans of African Ancestry in SW USA
ACB	African Caribbeans in Barbados
MXL	Mexican Ancestry from Los Angeles USA
PUR	Puerto Ricans from Puerto Rico
CLM	Colombians from Medellin, Colombia
PEL	Peruvians from Lima, Peru

Principal Components

	PC1	PC2
0	0.00339	-0.00426
1	0.00263	0.0201
10	-0.00229	9.89E-05
100	-0.000349	-0.000163
1000	0.000201	0.0115
1001	0.00435	-0.000958
1002	0.00135	-0.000712
1003	-0.00498	-0.146
1004	0.0465	0.0717
1005	0.000131	-0.000262
1006	0.000132	-0.000177
1007	1.09E-34	4.77E-26
1008	-0.00711	-0.092
1009	-0.000315	-9.85E-05
101	0.00262	-0.00394
1010	0.0275	-0.0173
1011	-0.0219	-0.00247
1012	0.048	0.033
1013	0.0413	-0.0241
1014	-0.0609	0.0424
1015	-0.00233	-0.01
1016	0.00042	-0.000827
1017	0.000426	-0.000939
1018	-0.00412	0.000308
1019	0.0922	-0.131
102	0	1.03E-41

Genomic locations like this one are very varied, 430 individuals had a 0 in this position, 692 had a 1 in this position and 338 had a 2 in this position. These SNPs may be important in understanding population structure.

Examining genomic locations like this one often reveals invariant sites: SNPs that don't display any differences at all in the population. I tested this one in particular and found that it was 0 in every individual in my sample. PCA can eliminate these unnecessary variables.

Uses of SVD

The most famous use of SVD is in principal components analysis and its cousins.
However, SVD is useful for more prosaic problems:

- Computing rank: rank is the number of singular values above some small specified tolerance.
- Useful in computing orthonormal bases of Null A and Col A.
- least-squares: unlike QR decomposition, SVD can be used even when matrix A does not have linearly independent columns.

Least squares via SVD

Algorithm for finding minimizer of $\|\mathbf{b} - A\mathbf{x}\|$:

> Find compact singular value decomposition (U, Σ, V) of A
> return $V\Sigma^{-1}U^T\mathbf{b}$

Justification: Let $\hat{x}$ be the vector returned by the algorithm.

$$\begin{aligned} A\hat{\mathbf{x}} &= (U\Sigma V^T)(V\Sigma^{-1}U^T\mathbf{b}) \\ &= U\Sigma\Sigma^{-1}U^T b \\ &= UU^T\mathbf{b} \\ &= U(\text{coord. repr. of } \mathbf{b}^{||\text{Col } U} \text{ in terms of cols of } U) \\ &= \mathbf{b}^{||\text{Col } U} \end{aligned}$$

and Col U = Col A.

Claim: The choice of $\hat{x}$ is the one minimizing $\|\hat{x}\|$.

We tried out deblurring. Currently no slides for this part of the lecture.