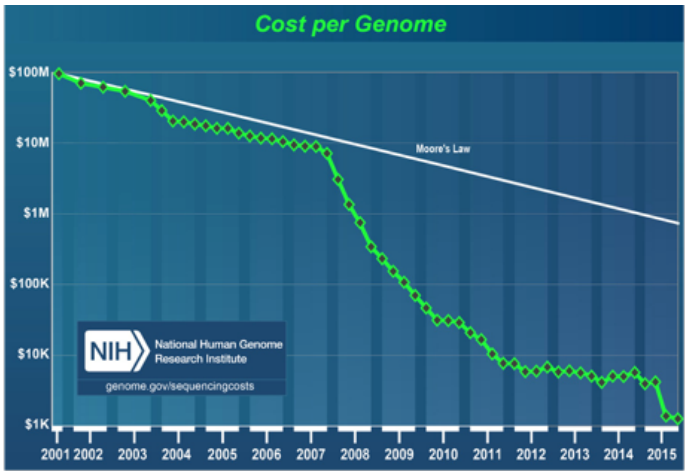
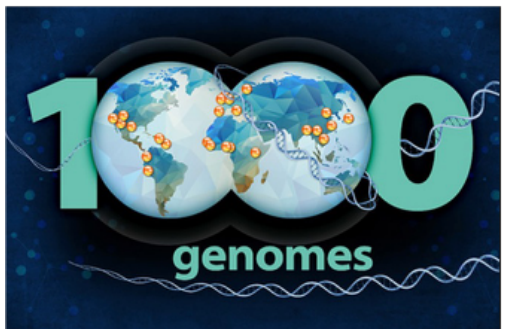
Scalable and flexible probabilistic PCA for large-scale genetic variation data

Sriram Sankararaman

Department of Computer Science
Department of Human Genetics
UCLA

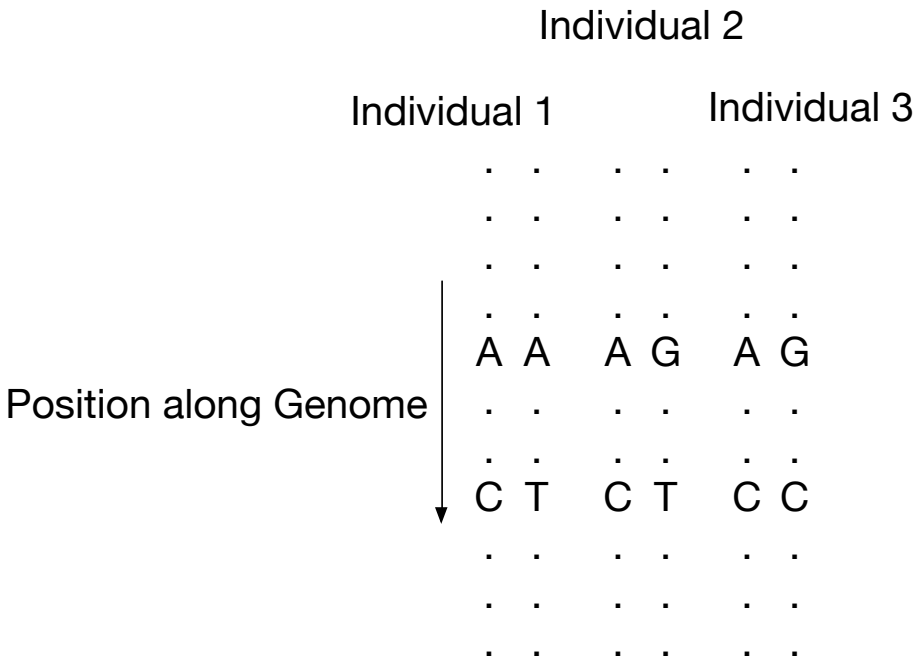
The genomic revolution

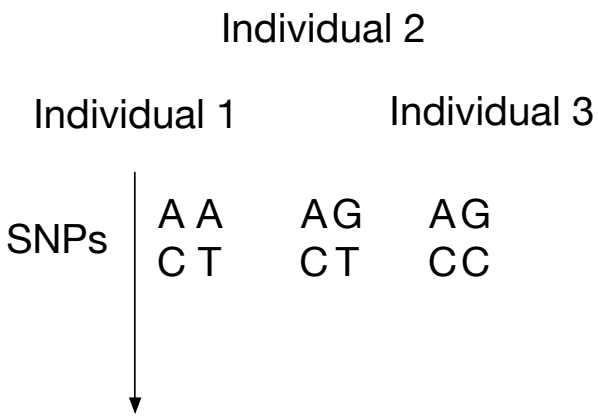


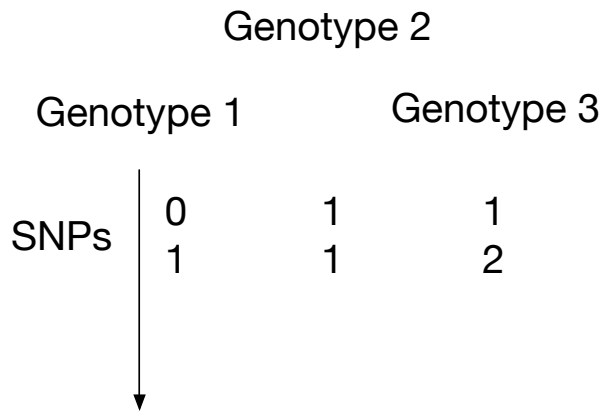


Challenge: scalable methods to analyze and visualize this data

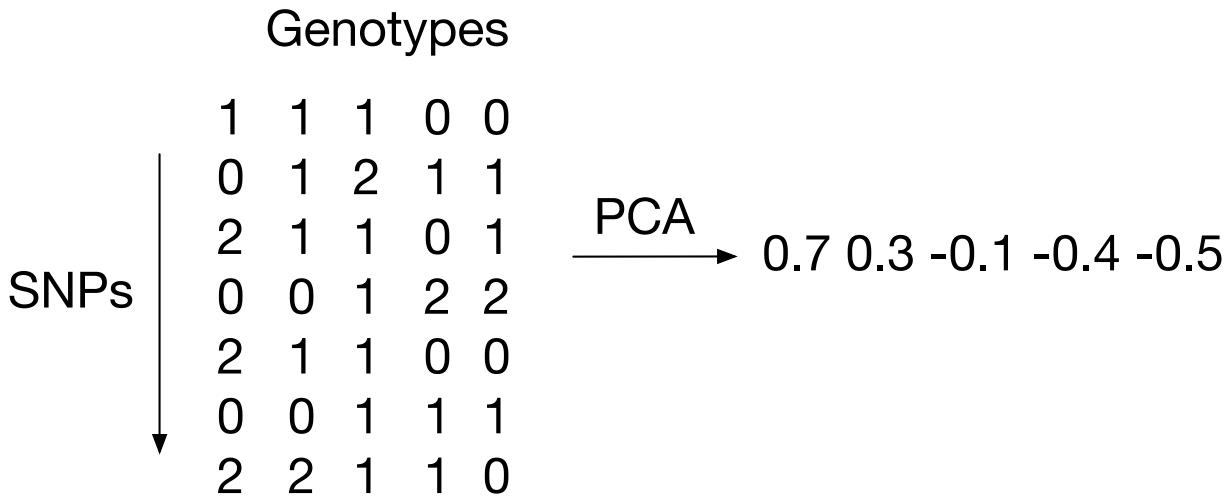
Genetic data





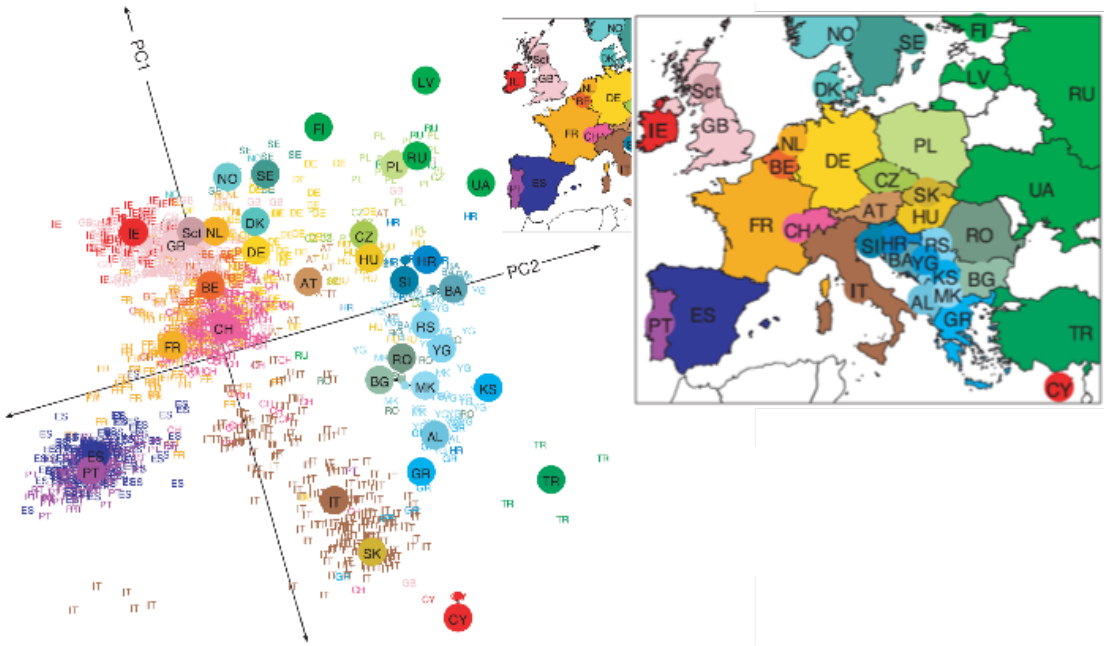


PCA on genetic data



PCA on genetic data

Visualize genetic structure



Novembre et al. Nature 2008

PCA

Given N genotype vectors over M SNPs $\boldsymbol{x}_n \in \mathbb{R}^M, n \in \{1, \ldots, N\}$ and $K \leq M$.

$$\begin{aligned}\boldsymbol{x}_n &\approx \boldsymbol{w}_1 z_{n,1} + \ldots + \boldsymbol{w}_K z_{n,K} \\ &= [\boldsymbol{w}_1 \ldots \boldsymbol{w}_K] \begin{bmatrix} z_{n,1} \\ \vdots \\ z_{n,K} \end{bmatrix} \\ &= \boldsymbol{W} \boldsymbol{z}_n\end{aligned}$$

- PCA Constraint: columns of $\boldsymbol{W}$ are orthonormal.
- The PCA solution $\widehat{\boldsymbol{W}} = \boldsymbol{U}_K$ where $\boldsymbol{U}_K$ contains the top K eigenvectors of the sample covariance matrix.

Challenges with PCA

Computational

- Compute all eigenvalue, eigenvectors.
  - Singular-Value Decomposition (SVD):

$$\mathcal{O}(MN \min(M, N)) \approx \mathcal{O}(MN^2)$$

  - Infeasible for genetic datasets (large number of SNPs M or individuals N).
- Recent Randomized approximation algorithms

Statistical

- Missing genotypes.
- Correlation among SNPs.

Halko et al. 2009, Galinskey et al. 2016

Probabilistic PCA

Model

$$\boldsymbol{z}_n \overset{iid}{\sim} \mathcal{N}(0, \boldsymbol{I}_K)$$

$$p(\boldsymbol{x}_n|\boldsymbol{z}_n, \boldsymbol{W}, \sigma^2) = \mathcal{N}(\boldsymbol{W}\boldsymbol{z}_n, \sigma^2 \boldsymbol{I}_M)$$

Log likelihood

$$\mathcal{LL}(\boldsymbol{W}, \sigma^2) \equiv \log P(\boldsymbol{X}|\boldsymbol{W}, \sigma^2) \quad = \log \prod_{n=1}^{N} p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{W}, \sigma^2)$$

The maximum likelihood estimator is equivalent to PCA.

Roweis 1999, Tipping and Bishop 1999, Engelhardt and Stephens 2010

Probabilistic PCA

EM algorithm

- E-step:

$$\boldsymbol{Z} = (\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W})^{-1}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}$$

- M-step:

$$\boldsymbol{W} = \boldsymbol{X}\boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}})^{-1}$$

- Assume: $\sigma^2 \to 0$

Probabilistic PCA

EM algorithm: computational complexity

- E-step:

$$\boldsymbol{Z} = \underbrace{(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W})}_{K\times K}^{-1} \underbrace{\boldsymbol{W}^{\mathrm{T}}}_{K\times M} \underbrace{\boldsymbol{X}}_{M\times N}$$

$\mathcal{O}(NMK)$

- M-step:

$$\boldsymbol{W} = \underbrace{\boldsymbol{X}}_{M\times N} \underbrace{\boldsymbol{Z}^{\mathrm{T}}}_{N\times K} \underbrace{(\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}})}_{K\times K}^{-1}$$

$\mathcal{O}(NMK)$

Probabilistic PCA

EM algorithm: computational complexity

- Run for I iterations with each iteration costing $\mathcal{O}(NMK)$.
- For small K, leads to a linear-time algorithm.

Probabilistic PCA

EM algorithm: computational complexity

- Run for I iterations with each iteration costing $\mathcal{O}(NMK)$.
- For small K, leads to a linear-time algorithm.
- Ignores the special structure of the genotype matrix $\boldsymbol{X}$.

Probabilistic PCA

EM algorithm: computational complexity

Each E and M step, perform the following operations K times:

$$\boldsymbol{c} = \boldsymbol{X}\boldsymbol{b}$$

- $\boldsymbol{X}$ is a fixed $M \times N$ matrix of genotypes.
- $\boldsymbol{b}$ is a real-valued vector that could potentially change each iteration.
- Naive multiplication takes $\mathcal{O}(NM)$.

Liberty and Zucker 2009

Probabilistic PCA

EM algorithm: computational complexity

Each E and M step, perform the following operations K times:

$$\boldsymbol{c} = \boldsymbol{X}\boldsymbol{b}$$

- $\boldsymbol{X}$ is a fixed $M \times N$ matrix of genotypes.
- $\boldsymbol{b}$ is a real-valued vector that could potentially change each iteration.
- Naive multiplication takes $\mathcal{O}(NM)$.
- For a genotype matrix, can we do some pre-processing so that $\boldsymbol{X}\boldsymbol{b}$ can be computed more efficiently ?

Liberty and Zucker 2009

Probabilistic PCA

EM algorithm: computational complexity

Each E and M step, perform the following operations K times:

$$\boldsymbol{c} = \boldsymbol{X}\boldsymbol{b}$$

- $\boldsymbol{X}$ is a fixed $M \times N$ matrix of genotypes.
- $\boldsymbol{b}$ is a real-valued vector that could potentially change each iteration.
- Naive multiplication takes $\mathcal{O}(NM)$.
- For a genotype matrix, can we do some pre-processing so that $\boldsymbol{X}\boldsymbol{b}$ can be computed more efficiently ?
- Yes! For a matrix with binary entries: $\mathcal{O}(\frac{MN}{\log_2(N)})$.

Liberty and Zucker 2009

EM with the Mailman algorithm

- Entries in genotype matrix take one of three values: $\{0, 1, 2\}$.
- Using the Mailman algorithm, per-iteration time complexity of EM for genotype matrix: $\mathcal{O}(\frac{MNK}{\log_3(N)})$.

Sub-linear time algorithm for computing PCA

Simulations

Accuracy

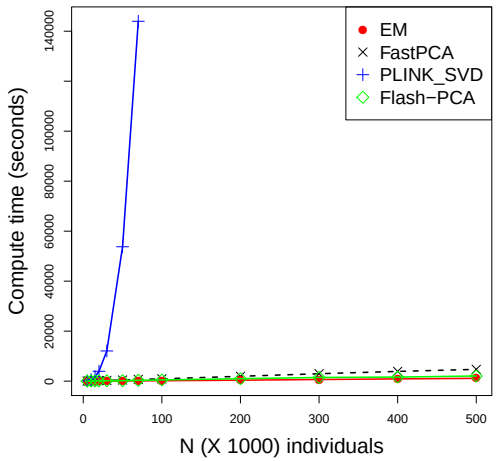
$50,000$ SNPs, $10,000$ individuals

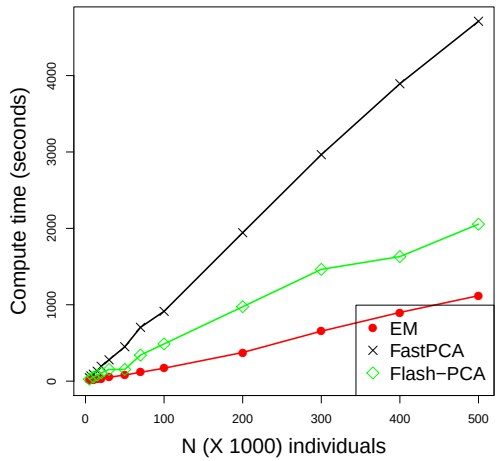
F_{st}	MEV	
	$K = 5$	$K = 10$
0.001	0.987	1.000
0.002	0.999	1.000
0.003	0.999	1.000
0.004	0.999	1.000
0.005	1.000	1.000
0.006	1.000	1.000
0.007	1.000	1.000
0.008	1.000	1.000
0.009	1.000	1.000
0.010	1.000	1.000

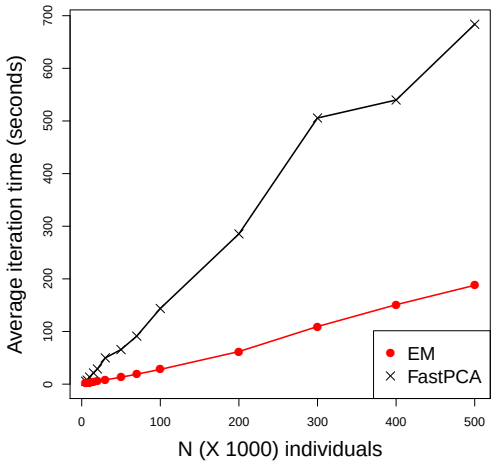
Simulations

Efficiency

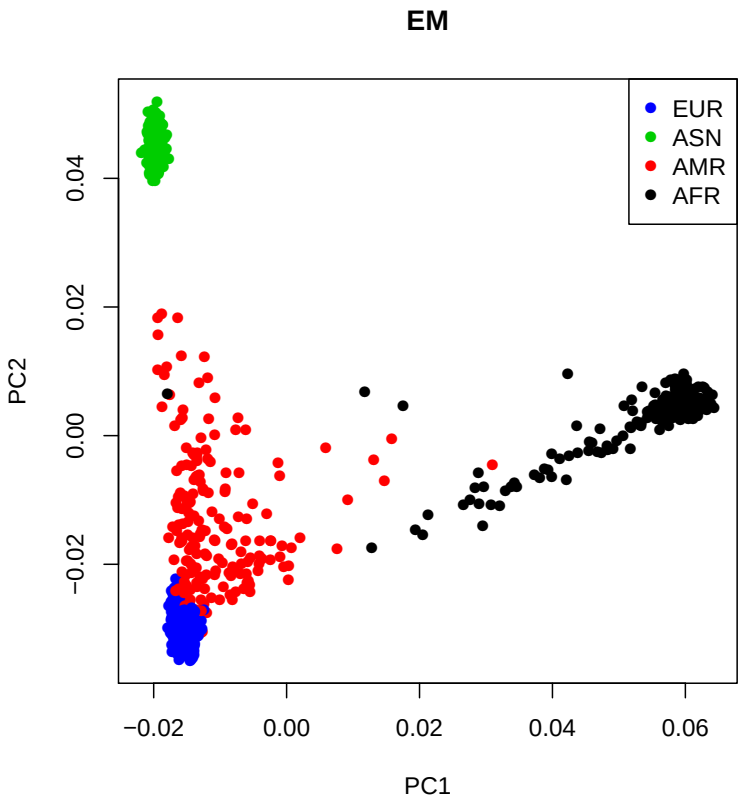
$$M = 100,000 \text{ SNPs},\ K = 5,\ F_{ST} = 0.01$$

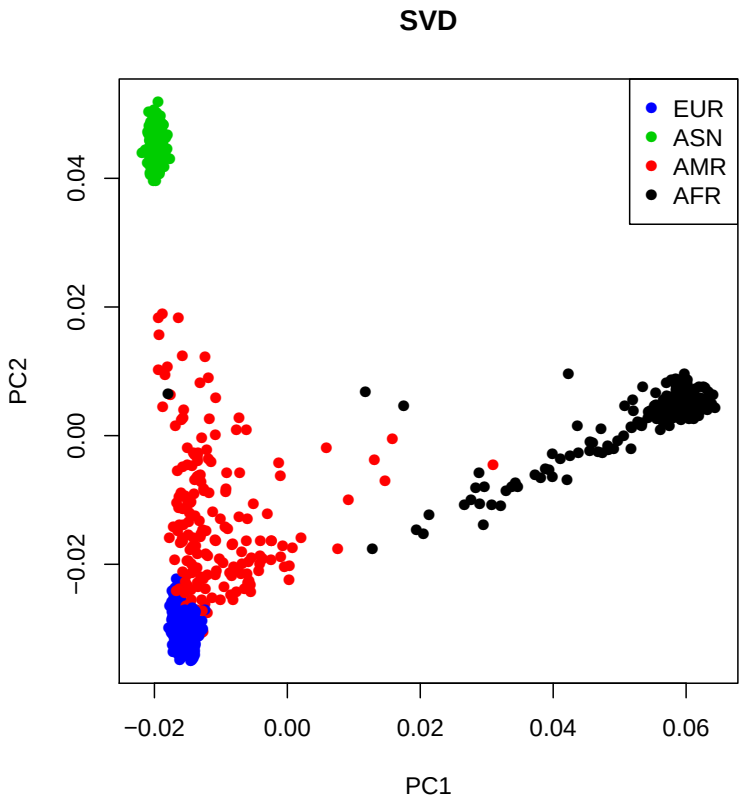






Application to 1000 Genomes data





Other advantages of Probabilistic PCA

Can naturally handle missing data

- E-step involves inferring hidden variables $\boldsymbol{Z}$ as well as hidden (missing observations).
- Can handle missing data efficiently.

Can use model selection to infer K.

- Choose K to maximize the marginal likelihood $P(\boldsymbol{X}|K)$.
- Use cross-validation and pick K that maximizes likelihood on held out data.

Open questions

- Modeling correlations .
- Beyond Gaussian outputs

Baran et al. 2013, Wen and Stephens 2012
Collins et al. 2002

Summary

- PCA can be interpreted as a latent variable model with continuous latent variable.
  - Probabilistic interpretation useful to generalize PCA.
  - Leads to efficient inference.
- Fast matrix-vector multiplication for genotype data more generally applicable and can lead to sub-linear time algorithms.

Acknowledgments

- Aman Agrawal
- Minh Le
- Eran Halperin

Email sriram@cs.ucla.edu