Efficient dimension reduction on massive data

MMDS 2010

Stoyan Georgiev[1], Sayan Mukherjee[2], Nick Patterson[3]

[1]Computational Biology and Bioinformatics Program
Institute for Genome Sciences & Policy, Duke University[1]
[2]Departments of Statistical Science, Computer Science,
and Mathematics, Institute for Genome Sciences & Policy
Duke University
[3] Broad Institute of MIT and Harvard

June 17, 2010

Overview

(1) Motivating genetics problem.

Overview

(1) Motivating genetics problem.

(2) Current approach (PCA).

Overview

(1) Motivating genetics problem.
(2) Current approach (PCA).
(3) Factor models.
    (a) A genetic example.
    (b) The Frisch problem.

Overview

(1) Motivating genetics problem.
(2) Current approach (PCA).
(3) Factor models.
    (a) A genetic example.
    (b) The Frisch problem.
(4) Supervised dimension reduction.

Overview

(1) Motivating genetics problem.
(2) Current approach (PCA).
(3) Factor models.
    (a) A genetic example.
    (b) The Frisch problem.
(4) Supervised dimension reduction.
(5) Efficient subspace inference.

Overview

(1) Motivating genetics problem.
(2) Current approach (PCA).
(3) Factor models.
    (a) A genetic example.
    (b) The Frisch problem.
(4) Supervised dimension reduction.
(5) Efficient subspace inference.
(6) Empirical results.
    (a) Wishart simulations.
    (b) Text example.

Inference of population structure

A classic problem in biology and genetics is to study population structure.

(1) Does genetic variation in populations follow geography ?

Inference of population structure

A classic problem in biology and genetics is to study population structure.

(1) Does genetic variation in populations follow geography ?

(2) Can we infer population histories from genetic variation ?

Inference of population structure

A classic problem in biology and genetics is to study population structure.

(1) Does genetic variation in populations follow geography ?

(2) Can we infer population histories from genetic variation ?

(3) When we associate genetic loci (locations) to disease we need to correct for population structure.

Genetic data

For each individual we have two letters from $\{A, C, T, G\}$ at each polymorphic (SNP) site which is coded as an integer $\{0, 1, 2\}$

$$C_i = \begin{pmatrix} AC \\ \vdots \\ GG \\ \vdots \\ TT \end{pmatrix} \Longrightarrow \begin{pmatrix} 1 \\ \vdots \\ 0 \\ \vdots \\ 2 \end{pmatrix} \in \mathbb{R}^{500{,}000},$$

Genetic data

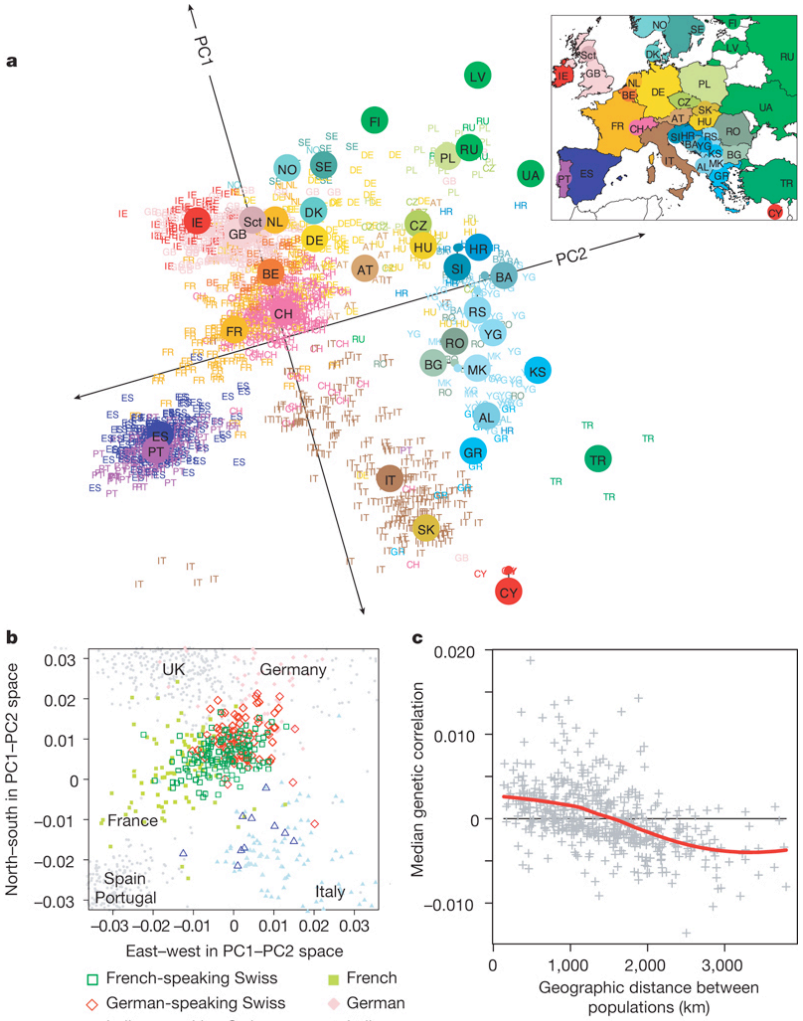
For each individual we have two letters from $\{A, C, T, G\}$ at each polymorphic (SNP) site which is coded as an integer $\{0, 1, 2\}$

$$C_i = \begin{pmatrix} AC \\ \vdots \\ GG \\ \vdots \\ TT \end{pmatrix} \Longrightarrow \begin{pmatrix} 1 \\ \vdots \\ 0 \\ \vdots \\ 2 \end{pmatrix} \in \mathbb{R}^{500{,}000},$$

$$C = [C_1, ..., C_m].$$

Genetic data encodes population history

From Novembre et al 2008 (Nature)



Dominant method for inference of population

Eigenstrat: Patterson et al 2006 (PLoS Genetics)
Combines principal components analysis and Tracy-Widom theory to infer population structure.

Dominant method for inference of population

Eigenstrat: Patterson et al 2006 (PLoS Genetics)
Combines principal components analysis and Tracy-Widom theory to infer population structure.

(1) $M_{ij} = \frac{C_{ij} - \hat{\mu}_j}{\sqrt{\frac{\hat{\mu}_j}{2}(1 - \frac{\hat{\mu}_j}{2})}} \quad \forall i, j.$

Dominant method for inference of population

Eigenstrat: Patterson et al 2006 (PLoS Genetics)
Combines principal components analysis and Tracy-Widom theory to infer population structure.

(1) $M_{ij} = \frac{C_{ij} - \hat{\mu}_j}{\sqrt{\frac{\hat{\mu}_j}{2}(1 - \frac{\hat{\mu}_j}{2})}} \quad \forall i, j.$

(2) $X = \frac{1}{n} MM'$

Dominant method for inference of population

Eigenstrat: Patterson et al 2006 (PLoS Genetics)
Combines principal components analysis and Tracy-Widom theory to infer population structure.

(1) $M_{ij} = \frac{C_{ij} - \hat{\mu}_j}{\sqrt{\frac{\hat{\mu}_j}{2}(1 - \frac{\hat{\mu}_j}{2})}} \quad \forall i, j.$

(2) $X = \frac{1}{n} MM'$

(3) Order $\lambda_1, ..., \lambda_m$ and test for significant eigenvalues using TW statistics

Dominant method for inference of population

Eigenstrat: Patterson et al 2006 (PLoS Genetics)
Combines principal components analysis and Tracy-Widom theory to infer population structure.

(1) $M_{ij} = \frac{C_{ij} - \hat{\mu}_j}{\sqrt{\frac{\hat{\mu}_j}{2}(1 - \frac{\hat{\mu}_j}{2})}} \quad \forall i, j.$

(2) $X = \frac{1}{n} MM'$

(3) Order $\lambda_1, ..., \lambda_m$ and test for significant eigenvalues using TW statistics

(4) Compute

$$n' = \frac{(m+1)\left(\sum_i \lambda_i\right)^2}{\left((m-1)\sum_i \lambda_i^2\right) - \left(\sum_i \lambda_i\right)^2}.$$

The challenge

We will be getting genetic data with
$n \geq 500,000$
$m \geq 30,000$.

The challenge

We will be getting genetic data with
$n \geq 500,000$
$m \geq 30,000$.

Can we extend Eigenstrat to this data to be run on a standard desktop ?

The challenge

We will be getting genetic data with
$n \geq 500,000$
$m \geq 30,000$.

Can we extend Eigenstrat to this data to be run on a standard desktop ?

Yes ! But....

Probabilistic view of PCA

$X \in \mathbb{R}^p$ is charterized by a multivariate normal

$$X \sim \mathrm{No}(\mu + A\nu, \Delta),$$
$$\nu \sim \mathrm{No}(0, \mathbf{I}_d)$$

$\mu \in \mathbb{R}^p$
$A \in \mathbb{R}^{p \times d}$
$\Delta \in \mathbb{R}^{p \times p}$
$\nu \in \mathbb{R}^d$.

Probabilistic view of PCA

$X \in \mathbb{R}^p$ is charterized by a multivariate normal

$$X \sim \mathrm{No}(\mu + A\nu, \Delta),$$
$$\nu \sim \mathrm{No}(0, \mathbf{I}_d)$$

$\mu \in \mathbb{R}^p$
$A \in \mathbb{R}^{p \times d}$
$\Delta \in \mathbb{R}^{p \times p}$
$\nu \in \mathbb{R}^d$.

ν is a latent variable, what is d.

A genetic example

We obtain genetic data from Yorba (African) and Japanese people.

(1) Run Eignestrat: obtain 4 pcs

A genetic example

We obtain genetic data from Yorba (African) and Japanese people.

(1) Run Eignestrat: obtain 4 pcs

(2) Run a factor model constrained to two factors. Observe the principal components are the mean allele frequencies of the two populations.

A genetic example

We obtain genetic data from Yorba (African) and Japanese people.

(1) Run Eignestrat: obtain 4 pcs

(2) Run a factor model constrained to two factors. Observe the principal components are the mean allele frequencies of the two populations.

What is right ?

Both

Let us decompose the covariance of the genetic variation Σ

(1) μ_1: mean allele frequency in Yorba

Both

Let us decompose the covariance of the genetic variation Σ

(1) μ_1: mean allele frequency in Yorba

(2) μ_2: mean allele frequency in Japanese

Both

Let us decompose the covariance of the genetic variation Σ

(1) μ_1: mean allele frequency in Yorba

(2) μ_2: mean allele frequency in Japanese

(3) $\Sigma = g(\mu_1\mu_1' + \mu_2\mu_2' + \mu_1\mu_2')$

Both

Let us decompose the covariance of the genetic variation Σ

(1) μ_1: mean allele frequency in Yorba

(2) μ_2: mean allele frequency in Japanese

(3) $\Sigma = g(\mu_1\mu_1' + \mu_2\mu_2' + \mu_1\mu_2')$

So the covariance is rank 4 even if two factors capture the allele structure in the two populations.

Frisch problem (1934)

Given m observations of n variables, what are the linear relations between the variables and how many linear relations are there ?

Frisch problem (1934)

Given m observations of n variables, what are the linear relations between the variables and how many linear relations are there ?

$$\begin{aligned} \text{minimize } & \operatorname{rank}(\Sigma - \Psi) \\ \text{subject to } & \Sigma - \Psi \succeq 0 \\ & \psi_j > 0, \end{aligned}$$

Possible way out

The important quantity we should worry about is the subspace we project onto.

Possible way out

The important quantity we should worry about is the subspace we project onto.

Infer $B = \text{span}(v_1, ..., v_d)$.

Supervised dimension reduction (SDR)

Given response variables $Y_1, ..., Y_m \in \mathbb{R}$ and explanatory variables or covariates $X_1, ..., X_m \in \mathbb{X} \subset \mathbb{R}^p$

$$Y_i = f(X_i) + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \text{No}(0, \sigma^2).$$

Supervised dimension reduction (SDR)

Given response variables $Y_1, ..., Y_m \in \mathbb{R}$ and explanatory variables or covariates $X_1, ..., X_m \in \mathbb{X} \subset \mathbb{R}^p$

$$Y_i = f(X_i) + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \text{No}(0, \sigma^2).$$

Is there a subspace $\mathcal{S} \equiv \mathcal{S}_{Y|X}$ such that $Y \perp\!\!\!\perp X \mid P_{\mathcal{S}}(X)$ with

$$P_{\mathcal{S}}(X) = B'X, \quad B = (b_1, ..., b_d).$$

Distribution theory for SDR

Sliced inverse regression: K.C. Li 1991, (JASA):

(1) Define the following quantities

$$\Omega \equiv \mathrm{cov}\left(\mathbb{E}[X \mid Y]\right), \quad \Sigma_x = \mathrm{cov}\left(X\right).$$

Distribution theory for SDR

Sliced inverse regression: K.C. Li 1991, (JASA):

(1) Define the following quantities

$$\Omega \equiv \mathrm{cov}\left(\mathbb{E}[X \mid Y]\right), \quad \Sigma_x = \mathrm{cov}\left(X\right).$$

(2) Solve the following generalized eigen-decomposition problem

$$\Omega b = \lambda \Sigma b.$$

Distribution theory for SDR

Sliced inverse regression: K.C. Li 1991, (JASA):

(1) Define the following quantities

$$\Omega \equiv \mathrm{cov}\left(\mathbb{E}[X \mid Y]\right), \quad \Sigma_x = \mathrm{cov}\left(X\right).$$

(2) Solve the following generalized eigen-decomposition problem

$$\Omega b = \lambda \Sigma b.$$

(3) $B = \mathrm{span}(b_1, ..., b_d)$ for all $i = 1, ..., d$ such that $\lambda_i \geq \epsilon$.

Distribution theory for SDR

Sliced inverse regression: K.C. Li 1991, (JASA):

(1) Define the following quantities

$$\Omega \equiv \mathrm{cov}\left(\mathbb{E}[X \mid Y]\right), \quad \Sigma_x = \mathrm{cov}\left(X\right).$$

(2) Solve the following generalized eigen-decomposition problem

$$\Omega b = \lambda \Sigma b.$$

(3) $B = \mathrm{span}(b_1, ..., b_d)$ for all $i = 1, ..., d$ such that $\lambda_i \geq \epsilon$.

(4) This idea works if $p(X \mid Y)$ is elliptical (unimodal).

An algorithm

The data is $\{x_1, ..., x_m\}$ and $\{y_1, ..., y_m\}$

An algorithm

The data is $\{x_1, ..., x_m\}$ and $\{y_1, ..., y_m\}$

(1) Compute sample covariance matrix $\hat{\Sigma}_X$

An algorithm

The data is $\{x_1, ..., x_m\}$ and $\{y_1, ..., y_m\}$

(1) Compute sample covariance matrix $\hat{\Sigma}_X$

(2) Bin the $\{y_i\}_{i=1}^m$ values into S bins.

An algorithm

The data is $\{x_1, ..., x_m\}$ and $\{y_1, ..., y_m\}$

(1) Compute sample covariance matrix $\hat{\Sigma}_X$

(2) Bin the $\{y_i\}_{i=1}^m$ values into S bins.

(3) For each bin $s = 1, ..., S$ compute the mean, $x_{i \in s}$

$$\hat{\mu}_s = \frac{1}{n_s} \sum_{i \in s} x_i.$$

An algorithm

The data is $\{x_1, ..., x_m\}$ and $\{y_1, ..., y_m\}$

(1) Compute sample covariance matrix $\hat{\Sigma}_X$

(2) Bin the $\{y_i\}_{i=1}^m$ values into S bins.

(3) For each bin $s = 1, ..., S$ compute the mean, $x_{i \in s}$

$$\hat{\mu}_s = \frac{1}{n_s} \sum_{i \in s} x_i.$$

(4) Compute $\hat{\Omega}$

$$\hat{\Omega} = \frac{1}{S} \sum_{s} \hat{\mu}_s \hat{\mu}_s'.$$

An algorithm

The data is $\{x_1, ..., x_m\}$ and $\{y_1, ..., y_m\}$

(1) Compute sample covariance matrix $\hat{\Sigma}_X$

(2) Bin the $\{y_i\}_{i=1}^m$ values into S bins.

(3) For each bin $s = 1, ..., S$ compute the mean, $x_{i \in s}$

$$\hat{\mu}_s = \frac{1}{n_s} \sum_{i \in s} x_i.$$

(4) Compute $\hat{\Omega}$

$$\hat{\Omega} = \frac{1}{S} \sum_s \hat{\mu}_s \hat{\mu}_s'.$$

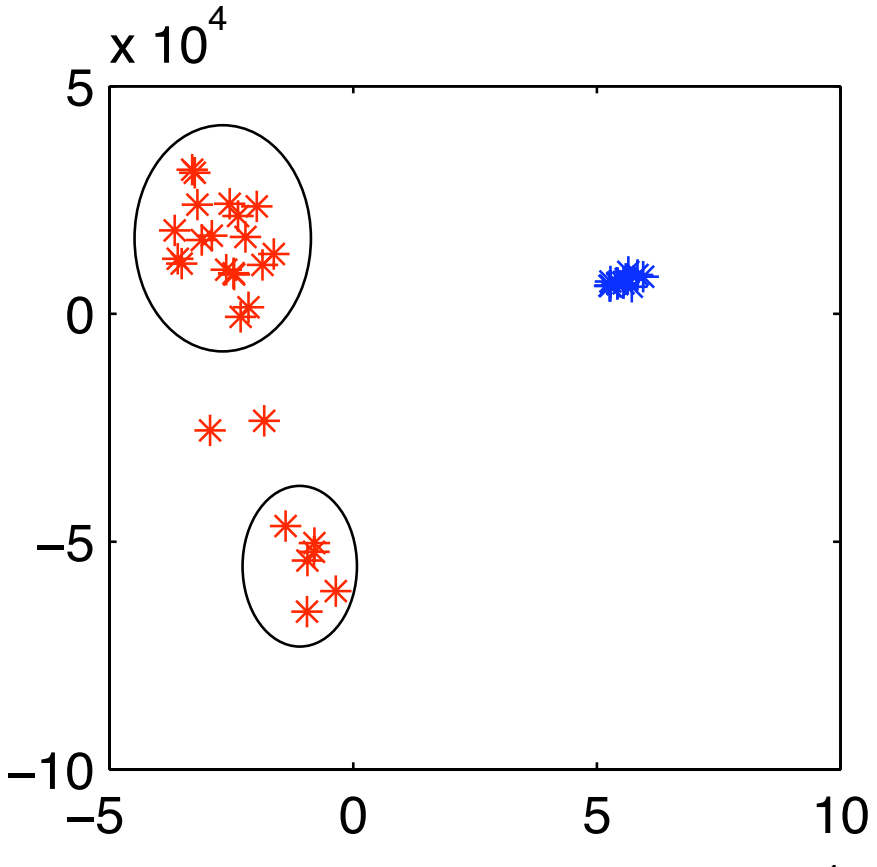
(4) Solve $\hat{\Omega} b = \lambda \hat{\Sigma} b$.

Subgroups or multimodal

$n = 7129$ dimensions, $m = 38$ samples,

19: Acute Myeloid Leukemia (AML)

19 are Acute Lymphoblastic Leukemia – B-cell and T-cell



Localization

Local sliced inverse regression: Wu et al 2010, (JCGS)

(1) Define the following quantities

$$\Omega_{\text{loc}} \equiv \text{cov}\left(\mathbb{E}[X_{\text{loc}} \mid Y]\right), \quad \Sigma_X = \text{cov}\left(X\right).$$

Localization

Local sliced inverse regression: Wu et al 2010, (JCGS)

(1) Define the following quantities

$$\Omega_{\text{loc}} \equiv \text{cov}\left(\mathbb{E}[X_{\text{loc}} \mid Y]\right), \quad \Sigma_X = \text{cov}\left(X\right).$$

(2) Solve the following generalized eigen-decomposition problem

$$\Omega_{\text{loc}} b = \lambda \Sigma b.$$

Localization

Local sliced inverse regression: Wu et al 2010, (JCGS)

(1) Define the following quantities

$$\Omega_{\text{loc}} \equiv \text{cov}\left(\mathbb{E}[X_{\text{loc}} \mid Y]\right), \quad \Sigma_X = \text{cov}\left(X\right).$$

(2) Solve the following generalized eigen-decomposition problem

$$\Omega_{\text{loc}} b = \lambda \Sigma b.$$

(3) $B = \text{span}(b_1, ..., b_d)$ for all $i = 1, ..., d$ such that $\lambda_i \geq \epsilon$.

Localization

Local sliced inverse regression: Wu et al 2010, (JCGS)

(1) Define the following quantities

$$\Omega_{\text{loc}} \equiv \text{cov}\left(\mathbb{E}[X_{\text{loc}} \mid Y]\right), \quad \Sigma_X = \text{cov}\left(X\right).$$

(2) Solve the following generalized eigen-decomposition problem

$$\Omega_{\text{loc}} b = \lambda \Sigma b.$$

(3) $B = \text{span}(b_1, ..., b_d)$ for all $i = 1, ..., d$ such that $\lambda_i \geq \epsilon$.

(4) This idea works if $p(X_{\text{loc}} \mid Y)$ is elliptical (unimodal).

Metrics for subspace estimates

Given two subspaces $\hat{B}$ and B we will look at two metrics to compute the similarity of $\hat{B}$ to B

(1) Qiang: Projection onto

$$\frac{1}{d}\sum_{i=1}^{d}||P_B\hat{b}_i||^2 = \frac{1}{d}\sum_{i=1}^{d}||(BB^T)\hat{b}_i||^2$$

Metrics for subspace estimates

Given two subspaces $\hat{B}$ and B we will look at two metrics to compute the similarity of $\hat{B}$ to B

(1) Qiang: Projection onto

$$\frac{1}{d}\sum_{i=1}^{d}||P_B\hat{b}_i||^2 = \frac{1}{d}\sum_{i=1}^{d}||(BB^T)\hat{b}_i||^2$$

(2) Golub: Angle between

$$dist(\hat{B}, B) = \sqrt{1-\cos(\theta_d)^2},$$

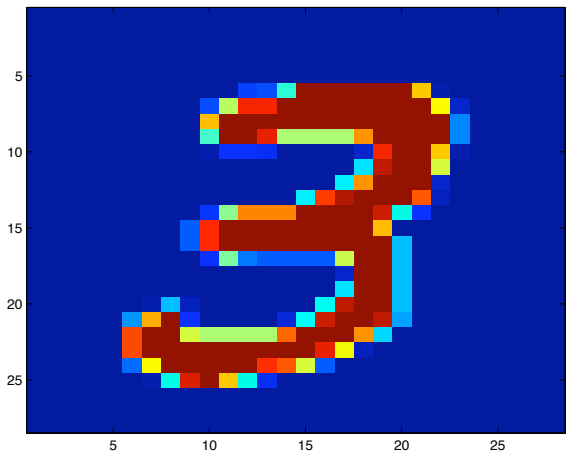
where the principle angles $\theta_1, ..., \theta_d$ are computed recursively

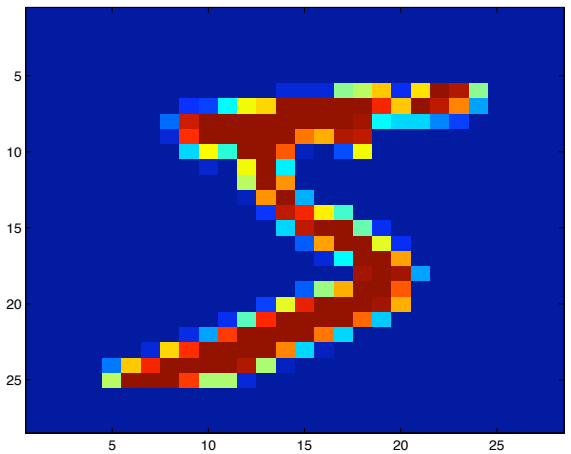
$$\cos(\theta_i) = \max_{u\in B}\max_{v\in\hat{B}} u'v = u_i'v_i$$

subject to

$$\|u\| = \|v\| = 1, \quad u \perp \{u_1, .., u_{i-1}\}, \quad v \perp \{v_1, .., v_{i-1}\}.$$

Digits





All ten digits

digit	Nonlinear	Linear
0	0.04($\pm$ 0.01)	0.05 ($\pm$ 0.01)
1	0.01($\pm$ 0.003)	0.03 ($\pm$ 0.01)
2	0.14($\pm$ 0.02)	0.19 ($\pm$ 0.02)
3	0.11($\pm$ 0.01)	0.17 ($\pm$ 0.03)
4	0.13($\pm$ 0.02)	0.13 ($\pm$ 0.03)
5	0.12($\pm$ 0.02)	0.21 ($\pm$ 0.03)
6	0.04($\pm$ 0.01)	0.0816 ($\pm$ 0.02)
7	0.11($\pm$ 0.01)	0.14 ($\pm$ 0.02)
8	0.14($\pm$ 0.02)	0.20 ($\pm$ 0.03)
9	0.11($\pm$ 0.02)	0.15 ($\pm$ 0.02)
average	0.09	0.14

Table: Average classification error rate and standard deviation on the digits data.

Randomized methods

By combining (block) Lanzcos and random projections Rhoklin et al 2009 (SIAM J Mat Anal Appl) came up with a fast, provable, randomized method.

Randomized methods

By combining (block) Lanzcos and random projections Rhoklin et al 2009 (SIAM J Mat Anal Appl) came up with a fast, provable, randomized method.
The matrix is $m \times n$ it is of rank k and t is the number of iterations in a power method. With high probability approximations of the top k eigenvalues and eigenvectors can be well approximated in time

$$\mathcal{O}(mnkt).$$

Randomized PCA

data: $A \in \mathbb{R}^{m \times n}$, number of eigenvalues: $2k \leq m$, number of iterations i

Randomized PCA

data: $A \in \mathbb{R}^{m \times n}$, number of eigenvalues: $2k \leq m$, number of iterations i

(A) Find orthonormal basis for the range of A

0.1 $G \sim U[-1, 1] \in \mathbb{R}^{m \times \ell}$

0.2 $R_0 = A^T G$

0.3 $\forall j = 1, ...i \; R_j = (A^T A) R_{j-1}$

0.4 $R = [R_0 \; . \; . \; . \; R_i]$

0.5 $R = QS$, Q orthonormal, S upper triangular

Randomized PCA

data: $A \in \mathbb{R}^{m \times n}$, number of eigenvalues: $2k \leq m$, number of iterations i

(A) Find orthonormal basis for the range of A

0.1 $G \sim U[-1, 1] \in \mathbb{R}^{m \times \ell}$
0.2 $R_0 = A^T G$
0.3 $\forall j = 1, ...i \; R_j = (A^T A) R_{j-1}$
0.4 $R = [R_0 \; . \; . \; . \; R_i]$
0.5 $R = QS$, Q orthonormal, S upper triangular

(B) Project data and do SVD

0.1 $B = AQ$
0.2 Factorize $B = U \Sigma W^T$ (using SVD)
0.3 Set $\hat{U} = U(:, 1 : k)$
Set $\hat{\Sigma} = \Sigma(1 : k)$
Set $\hat{V} = A^T \hat{U} \hat{\Sigma}^{-1}$

Our method

Iterate random PCA on the gram matrix $A = XX' \in \mathbb{R}^{n \times n}$ until subspace converge.

Our method

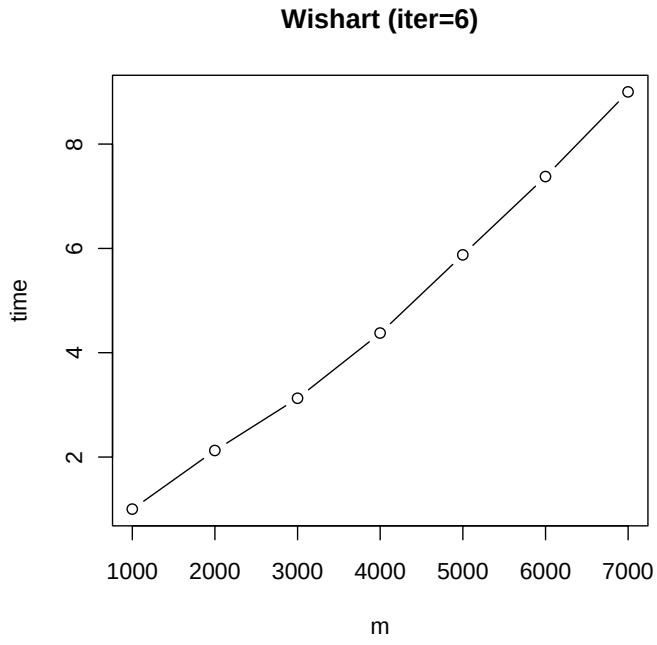
Iterate random PCA on the gram matrix $A = XX' \in \mathbb{R}^{n \times n}$ until subspace converge.

Main differences

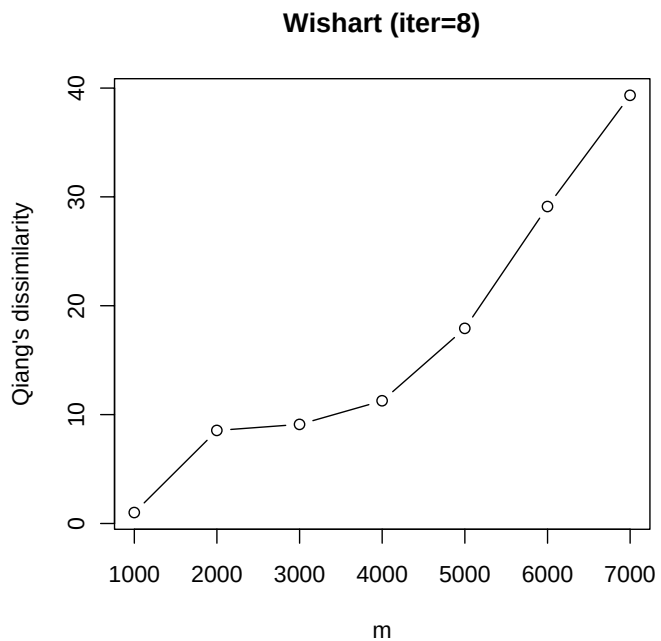
(1) Work with gram matrix to avoid storing in memory matrices the size of the data.

(2) Implemented packing/unpacking into bytes and 2-bit fields for SNP data.

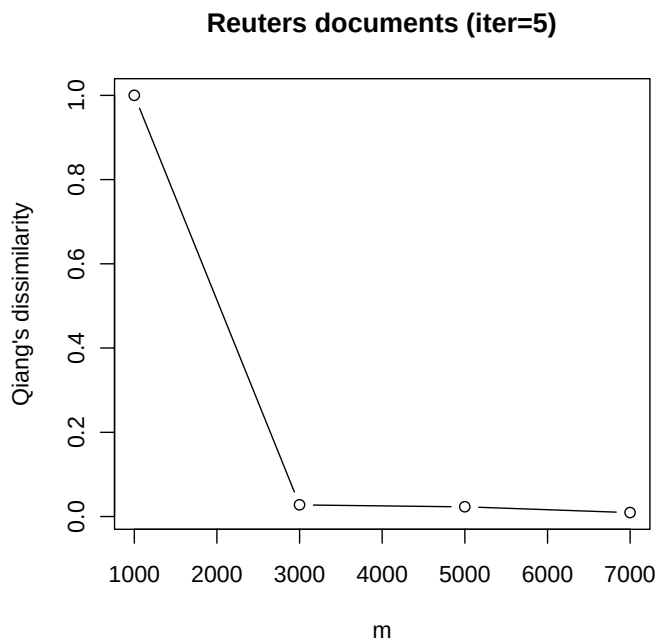
Timing



Error


Wishart (iter=8)
Qiang's dissimilarity
0
10
20
30
40
1000
2000
3000
4000
5000
6000
7000
m

Error


Reuters documents (iter=5)
Qiang's dissimilarity
0.0
0.2
0.4
0.6
0.8
1.0
1000
2000
3000
4000
5000
6000
7000
m

Conclusion

(1) Can inference of population structure in large data using eigen-decomposition.

Conclusion

(1) Can inference of population structure in large data using eigen-decomposition.

(2) Interpretation of subspace is easier than factors.

Conclusion

(1) Can inference of population structure in large data using eigen-decomposition.

(2) Interpretation of subspace is easier than factors.

(3) Method can be applied to generalzed eigen-decomposition.

Conclusion

(1) Can inference of population structure in large data using eigen-decomposition.

(2) Interpretation of subspace is easier than factors.

(3) Method can be applied to generalzed eigen-decomposition.

(4) Loss of numerical precision ?

Conclusion

(1) Can inference of population structure in large data using eigen-decomposition.

(2) Interpretation of subspace is easier than factors.

(3) Method can be applied to generalzed eigen-decomposition.

(4) Loss of numerical precision ?

(5) Fast computation of Tracy-Widom statistics using Fredholm determinants, Bourneman 2009, (ArchivX).

Acknowledgements

Funding:

- Center for Systems Biology at Duke
- NSF
- NIH