Principal Component Analysis

Announcements

- Exams still being graded
- HW 6 due next Thursday
- Final project
    - Begin forming groups. Group size is 2-4.
    - Start thinking about the project topic.
    - See project guidelines – linked from Canvas home page
    - As part of HW7, you will be asked to submit a project proposal.

Dimension(ality) Reduction

- The transformation of data to a lower-dimensional space such that meaningful information is retained
- Feature extraction vs. feature selection

new features are functions of old features

select a few of the existing features

- Motivations:

Feature selection: Find important original features

Reduce computational cost

Extract interpretable features

Visualization

Improve performance of a supervised learning algorithm.

Eliminate redundant/noisy features to avoid overfitting

PCA

- Feature extraction
- Unsupervised
- Linear
- Objective: squared reconstruction error

PCA

- Given $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, the idea behind PCA is to approximate

$$\boldsymbol{x}_i \approx \boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{\theta}_i$$

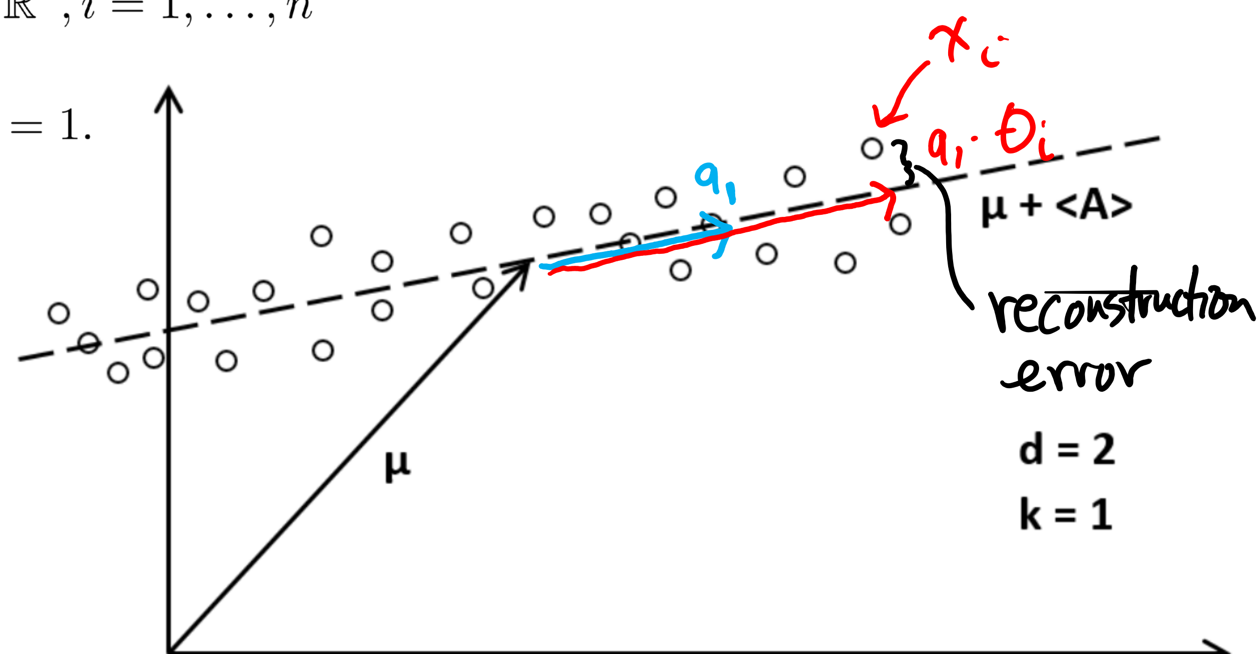
d×1 d×k k×1

where

$$\boldsymbol{\mu} \in \mathbb{R}^d$$
$$\boldsymbol{A} \in \mathcal{A}_k := \{\boldsymbol{A} \in \mathbb{R}^{d \times k} \mid \boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{I}_{k \times k}\}$$
$$\boldsymbol{\theta}_i \in \mathbb{R}^k, i = 1, \ldots, n$$

- **Example:** $d = 2, k = 1$.

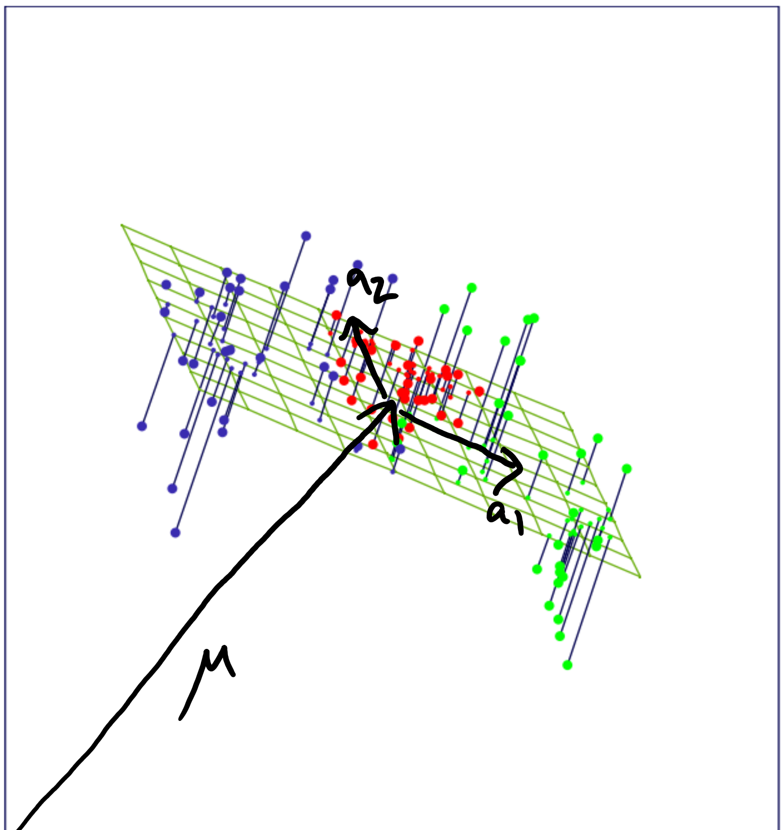
$A = [a_1] \in \mathbb{R}^2$



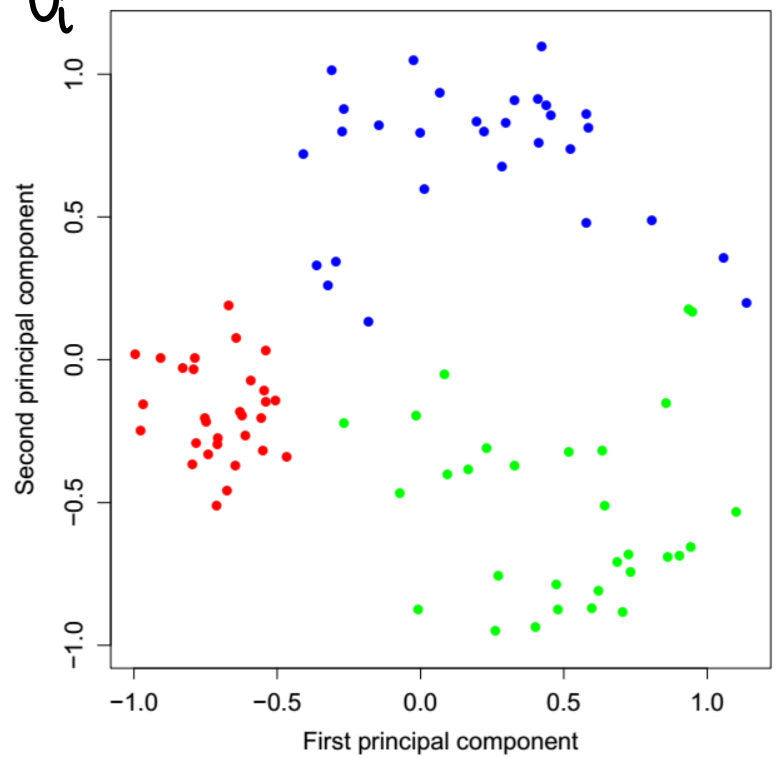
PCA

$$A = \begin{bmatrix} a_1 & a_2 \end{bmatrix}_{(3\times 2)}, \quad \theta_i = \begin{bmatrix} \theta_i^{(1)} \\ \theta_i^{(2)} \end{bmatrix}$$

- **Example:** $d = 3, k = 2$



$\theta_i^{(2)}$



$\theta_i^{(1)}$

$$x_i \approx \mu + A\theta_i = \mu + \theta_i^{(1)} \cdot a_1 + \theta_i^{(2)} a_2$$

PCA

- Mathematically, we define $\boldsymbol{\mu}, \boldsymbol{A}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ to be the solution of

$$\min_{\substack{\mu \in \mathbb{R}^d,\ \theta_1, \ldots, \theta_n \in \mathbb{R}^k \\ A \in \mathbb{R}^{d \times k}:\ A^T A = I}} \sum_{i=1}^{n} \| x_i - (\mu + A\theta_i) \|^2$$

- PCA gives the least squares rank-k linear approximation to the data set.

- The solution is given in terms of the spectral (or eigenvalue) decomposition of the sample covariance matrix

$$S = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T$$

$$\bar{x} = \frac{1}{n} \sum x_i$$

Poll

What do we know about the sample covariance matrix

$$\boldsymbol{S} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T?$$

Select the best answer.

(A) diagonal

(B) positive semi-definite

(C) positive definite

(D) B and C

$$z^T S z = \frac{1}{n}\sum \underbrace{z^T(x_i - \bar{x})}\underbrace{(x_i - \bar{x})^T z}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left[z^T(x_i - \bar{x})\right]^2$$

$$\geq 0$$

PCA Solution

- Denote the eigenvalue decomposition of $\boldsymbol{S}$

$$S = U \Lambda U^T \quad \text{where} \quad \Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_d \end{bmatrix}$$

$$U = \begin{bmatrix} u_1 & \cdots & u_d \end{bmatrix}, \quad U^T U = U U^T = I, \quad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0.$$

- A solution to PCA is

$$\mu = \bar{x}$$

$$A = \begin{bmatrix} u_1 & \cdots & u_k \end{bmatrix}, \quad \theta_i = A^T (x_i - \bar{x})$$

Terminology and Concepts

- Principal components

$$\theta_i^{(j)} = j^{th} \text{ principal component of } x_i$$
$$= u_j^T (x_i - \bar{x})$$

- Principal eigenvectors/directions

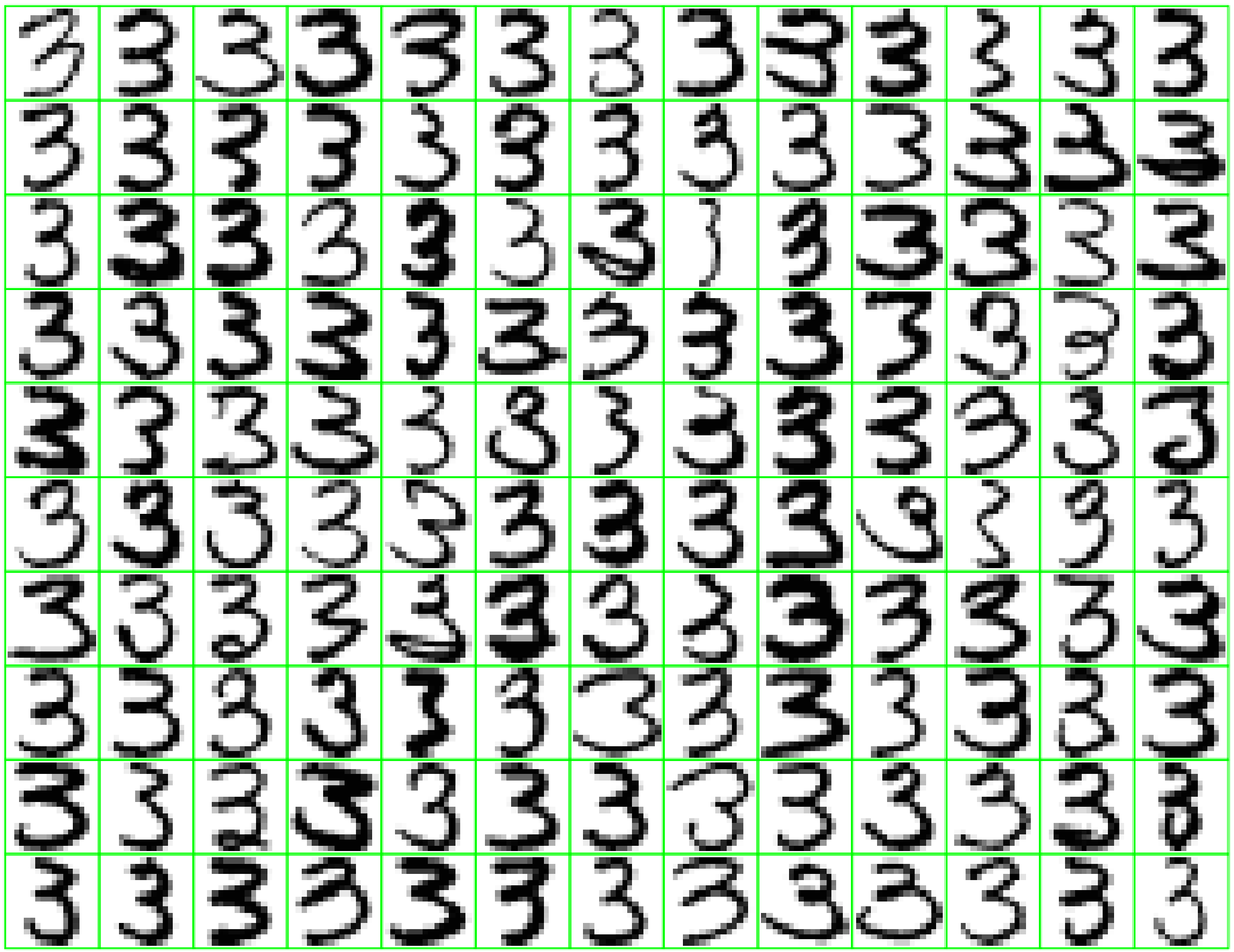
$$u_j = j^{th} \text{ principal eigenvector}$$

- Reconstruction of $\boldsymbol{x}_i$

$$\hat{x}_i = \mu + A\theta_i = \bar{x} + AA^T(x_i - \bar{x})$$

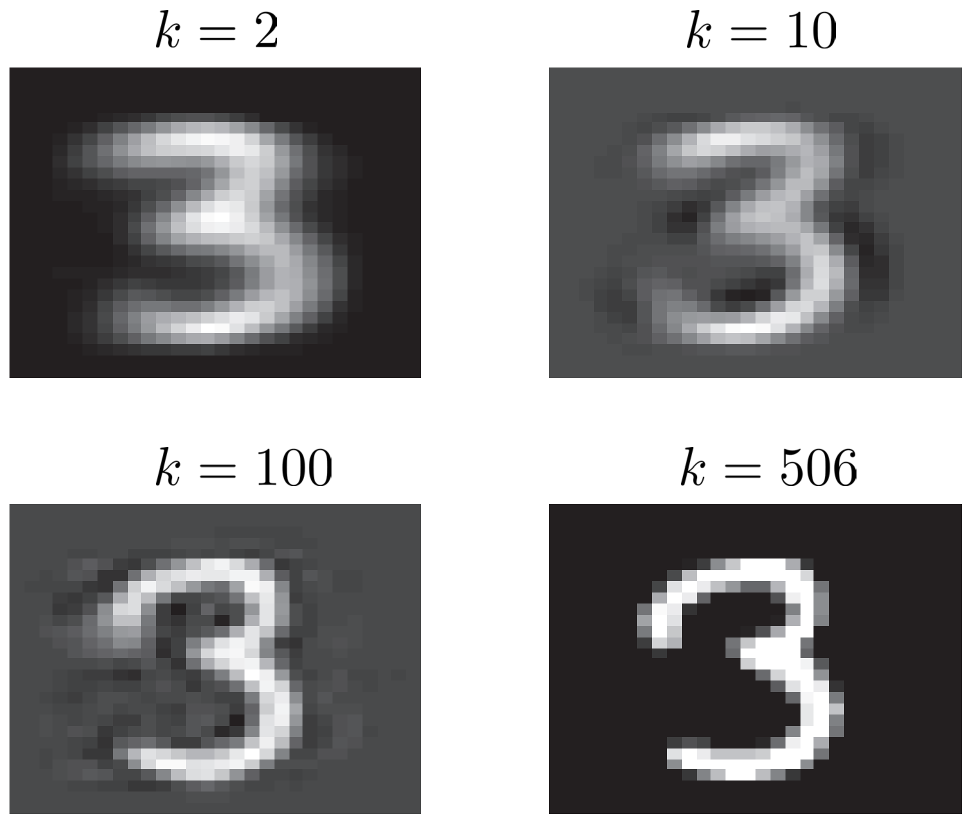
Digits Example

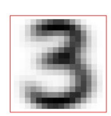
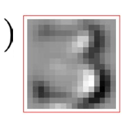
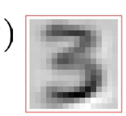
- Training data



Digits Example

- Reconstruction

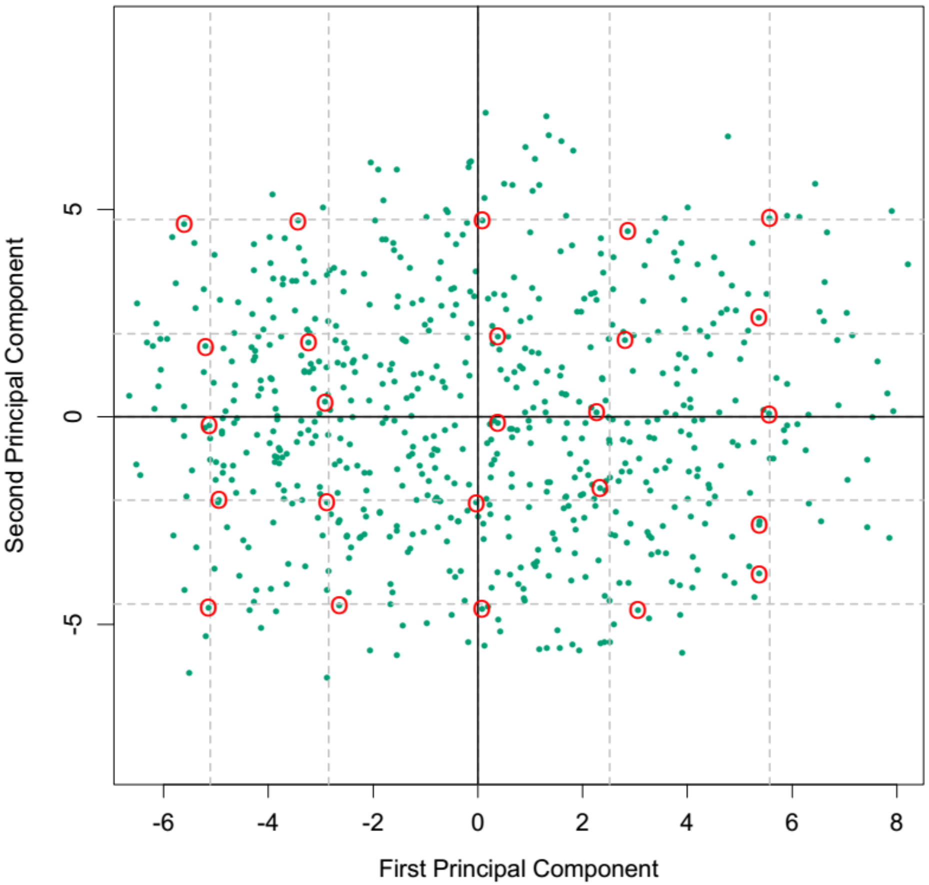


$\widehat{\boldsymbol{x}}_i = \boldsymbol{\mu} + \theta_i^{(1)} \boldsymbol{u}_1 + \theta_i^{(2)} \boldsymbol{u}_2 =$  $+\,\theta_i^{(1)}$  $+\,\theta_i^{(2)}$ 

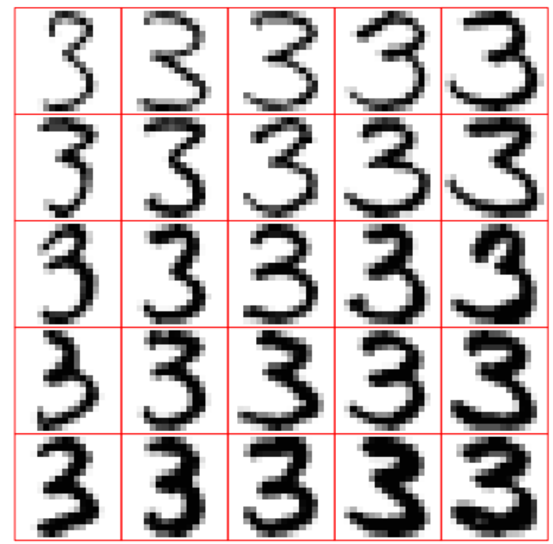
Digits Example

First two PCs

$\theta_i^{(2)}$



$\theta_i^{(1)}$



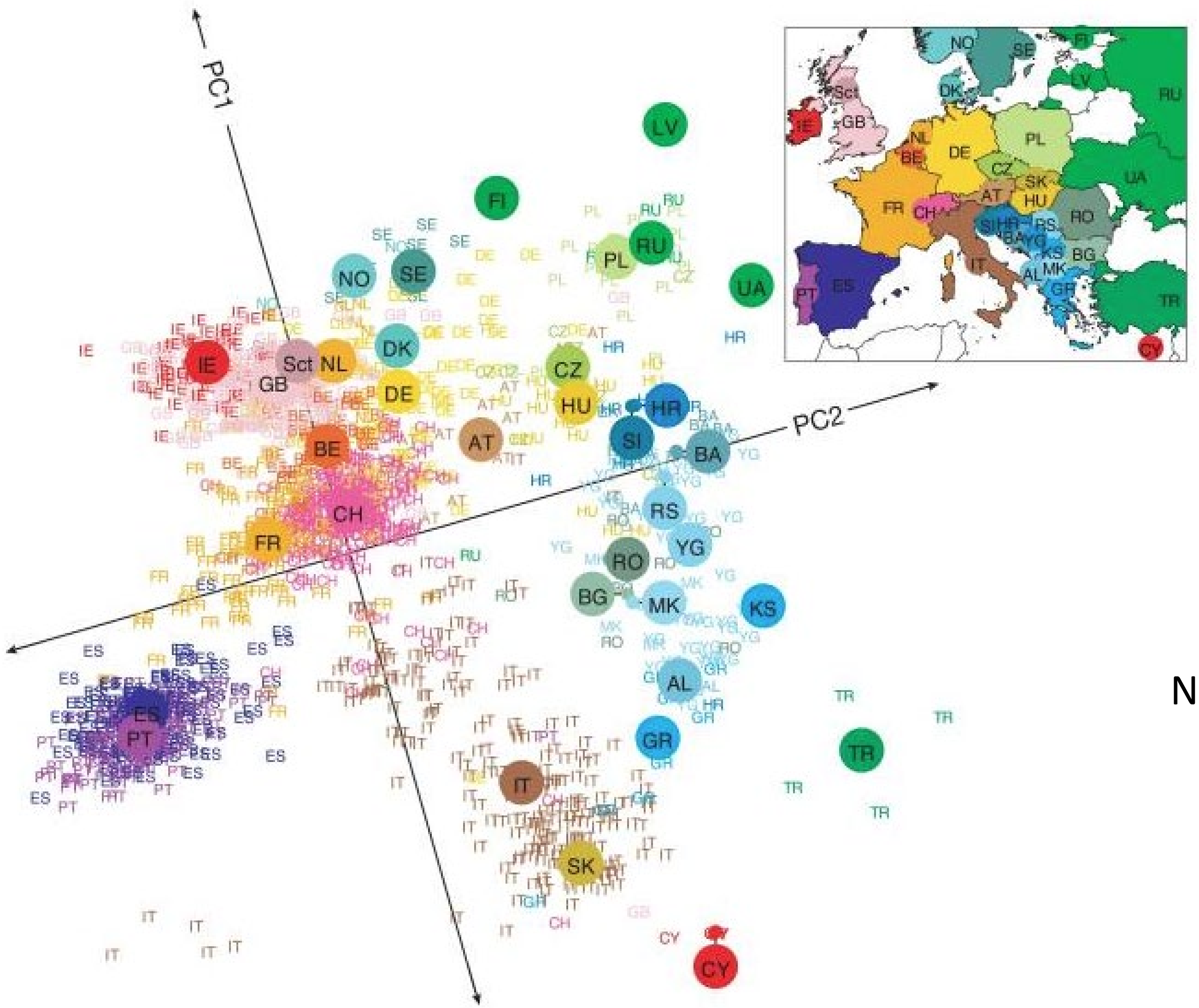
thickness of penstroke

width

and/or articulation of the 3 (how curved)

Interpretation?

Genotype Data



Novembre et al. (2008)

Poll

True or False: If we first center the data (that is, we subtract the mean $\bar{\boldsymbol{x}}$ from each $\boldsymbol{x}_i$) before applying PCA, the principal components and principal eigenvectors do not change.

(A) True

(B) False

(C) Not enough information

$$\tilde{x}_i = x_i - \bar{x}$$

$$\tilde{S} = \frac{1}{n} \sum_{i=1}^{n} (\tilde{x}_i - 0)(\tilde{x}_i - 0)^T$$

$$= \frac{1}{n} \sum (x_i - \bar{x})(x_i - \bar{x})^T$$

$$= S$$

$$\tilde{\theta}_i = \tilde{A}^T(\tilde{x}_i - 0) = A^T(x_i - \bar{x}) = \theta_i$$

Connection to Projections

- Suppose $\bar{\boldsymbol{x}} = \mathbf{0}$ (a common preprocessing step).
- Then the rank-k approximation to $\boldsymbol{x}_i$ is

$$\widehat{\boldsymbol{x}}_i = \mu + A\theta_i = \bar{x} + \underbrace{AA^T}_{=A(A^TA)^{-1}A^T}(x_i - \bar{x})$$

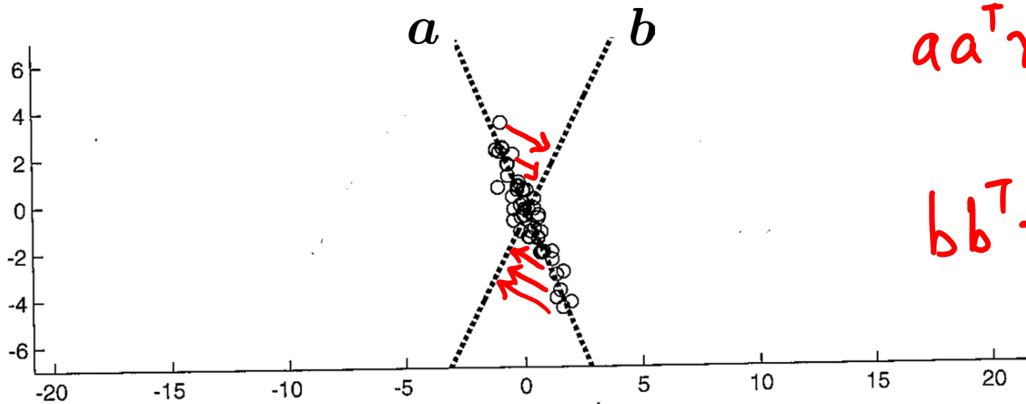
$$= A\theta_i = AA^Tx_i$$

$$= \sum_{j=1}^{k} \theta_i^{(j)} u_j$$

projection onto colspan(A) $=: \langle A \rangle$

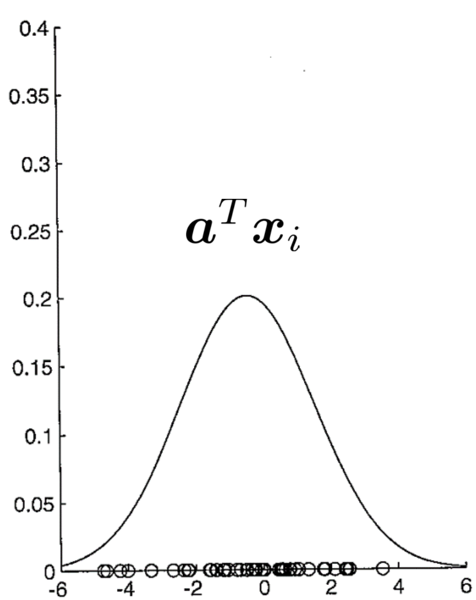
- Intuition:
  - $\widehat{\boldsymbol{x}}_i$ is the projection of $\boldsymbol{x}_i$ onto $\langle \boldsymbol{A} \rangle$.
  - Columns of $\boldsymbol{A}$ define a k dimensional coordinate system for $\langle \boldsymbol{A} \rangle$.
  - $\boldsymbol{\theta}_i = \boldsymbol{A}^T \boldsymbol{x}_i$ are the coordinates of $\widehat{\boldsymbol{x}}_i$ in the subspace

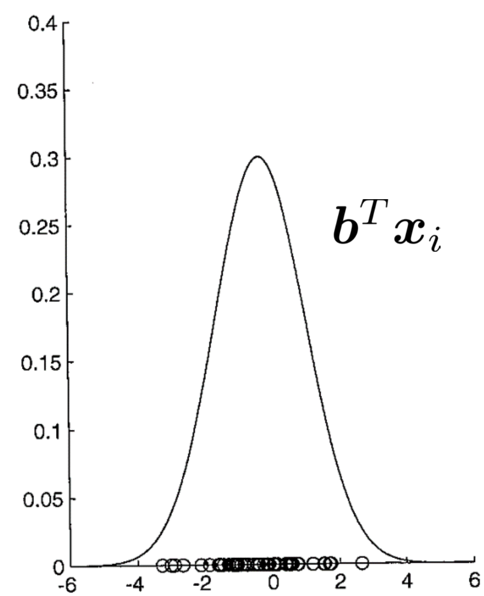
Maximum Variance Projections



$$a a^T x_i = a(\underbrace{a^T x_i}_{\theta_i})$$

$$b b^T x_i = b(\underbrace{b^T x_i}_{\Theta_i})$$





Maximum Variance Perspective

- Suppose $\bar{\boldsymbol{x}} = \mathbf{0}$.
- What is the unit vector $\boldsymbol{a}_1 \in \mathbb{R}^d$ ($\|\boldsymbol{a}_1\| = 1$) for which the sample variance of

$$\theta^{(1)} = \boldsymbol{a}_1^T X$$

  is maximized?

- The sample mean of $\boldsymbol{a}_1^T X$ is

$$\frac{1}{n} \sum a_1^T x_i = a_1^T \bar{x} = 0$$

- The sample variance of $\theta^{(1)}$ is therefore

$$\frac{1}{n} \sum_{i=1}^{n} (a_1^T x_i - 0)^2 = \frac{1}{n} \sum (a_1^T x_i)(x_i^T a_1)$$

$$= a_1^T \left(\frac{1}{n} \sum x_i x_i^T \right) a_1 = a_1^T S a_1$$

Maximum Variance Perspective

- We can express the sample variance of $\theta^{(1)}$ as

$$\mathrm{var}(\theta^{(1)}) = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{a}_1^T \boldsymbol{x}_i)^2 \quad = \quad a_1^T S a_1$$

- The solution of

$$\max_{\boldsymbol{a}_1 : \|\boldsymbol{a}_1\| = 1} \quad \boldsymbol{a}_1^T S \boldsymbol{a}_1$$

is

u_1 = the "largest eigenvector" of S

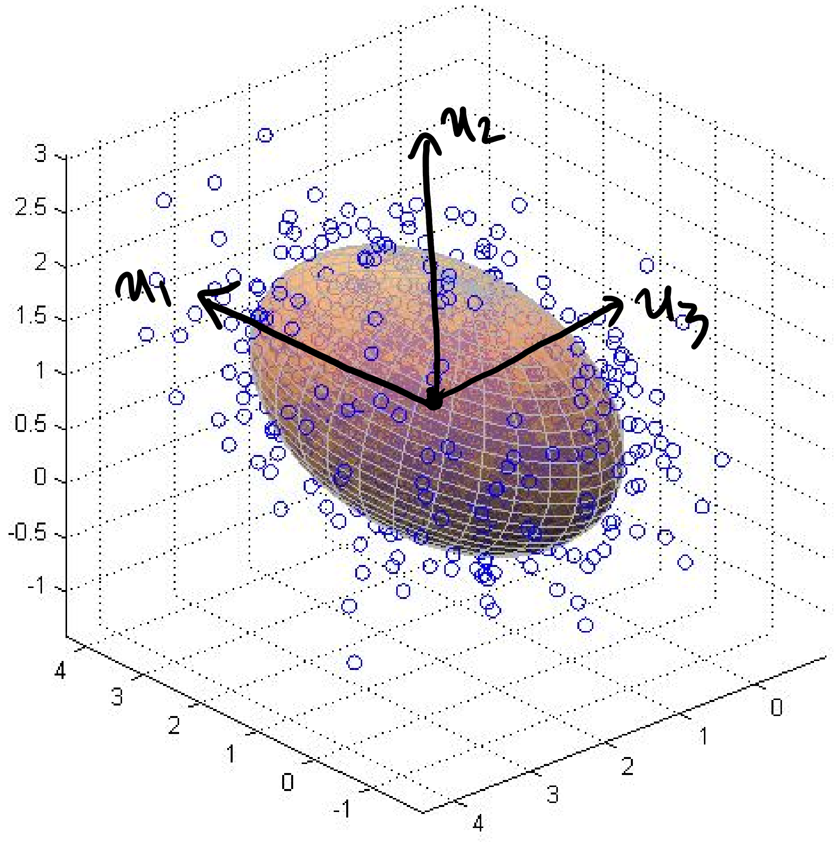
Maximum Variance Perspective

- More generally, we have the following result:
- **Theorem:** Let $\theta^{(k)} = \boldsymbol{a}_k^T X$ and $\text{var}(\theta^{(k)}) = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{a}_k^T \boldsymbol{x}_i)^2$. A vector $\boldsymbol{a}_k$ that maximizes $\text{var}(\theta^{(k)})$ subject to
    - $\|\boldsymbol{a}_k\| = 1$
    - $\boldsymbol{a}_k \perp \boldsymbol{u}_1, \ldots, \boldsymbol{u}_{k-1}$

  is $\boldsymbol{a}_k = \boldsymbol{u}_k$.
- What is the variance of $\theta^{(k)}$?

$$a_k^T S a_k = u_k^T (U \Lambda U^T) u_k$$

$$= e_k^T \Lambda e_k = \lambda_k$$



Selecting k

- It can be shown that the optimal objective function value is

$$\min_{\boldsymbol{\mu}, \boldsymbol{A}, \boldsymbol{\theta}_i} \sum \|\boldsymbol{x}_i - \boldsymbol{\mu} - \boldsymbol{A}\boldsymbol{\theta}_i\|^2 = n(\lambda_{k+1} + \cdots + \lambda_d)$$

- When $k = 0$, this specializes to

$$\min_{\boldsymbol{\mu}} \sum \|\boldsymbol{x}_i - \boldsymbol{\mu}\|^2 = n(\lambda_1 + \cdots + \lambda_d)$$

which we call the *total variation* of the data.

- One heuristic for choosing k is to select the smallest k such that

$$\frac{\lambda_1 + \cdots + \lambda_k}{\lambda_1 + \cdots + \lambda_d} \geq .95$$

% total variation explained by 1st k PCs.

Connection to SVD

- Assume $\bar{\boldsymbol{x}} = \mathbf{0}$
- Data matrix $(d \times n)$

$$\boldsymbol{X} = [\boldsymbol{x}_1 \cdots \boldsymbol{x}_n] \,.$$

- Let the singular value decomposition of $\boldsymbol{X}$ be

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$$

where $\boldsymbol{U}$ $(d \times d)$ and $\boldsymbol{V}$ $(n \times n)$ are orthogonal matrices, and $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_{\min\{d,n\}})$ is $d \times n$.

- Then

principal eigenvectors = left singular vectors

$$\lambda_j = \begin{cases} \frac{1}{n}\sigma_j^2, & \text{if } j \le \min(d,n) \\ 0, & \text{o/w.} \end{cases}$$

Extensions

- Kernel PCA
- Streaming PCA
- Supervised PCA
- Robust PCA
- Sparse PCA
- Probabilistic PCA
- Nonnegative matrix factorization
- . . .