Dimension reduction

COMS 4771 Fall 2023

Linear dimension reduction

Dimension reduction: map feature vectors from $\mathbb{R}^d$ to $\mathbb{R}^k$ with $k < d$

- Reduce storage requirements for dataset
- Improve understandability of individual data points
- Improve performance of learning algorithms on dataset
- ...

Many methods are linear: i.e., based on linear map $\varphi \colon \mathbb{R}^d \to \mathbb{R}^k$

This lecture: unsupervised methods for dimension reduction

Throughout this lecture, $X = (X_1, \ldots, X_d)$ is a random vector

e.g., $X =$ data point drawn uniformly at random from $\mathcal{S}$

Axis-aligned embeddings

Axis-aligned embeddings:

- Let $\varphi(x) \in \mathbb{R}^k$ keep a subset of k features x_i, throw away the rest

Question: Which features to keep?

- Simple heuristic: Choose the k most "informative" features

  Sort features by variance

$$\mathrm{var}(X_{(1)}) \geq \cdots \geq \mathrm{var}(X_{(d)})$$

  and choose $\varphi(x) = (x_{(1)}, \ldots, x_{(k)})$

Suppose only k features have non-negligible variance

$$\mathrm{var}(X_{(1)}) \geq \cdots \geq \mathrm{var}(X_{(k)}) \gg \mathrm{var}(X_{(k+1)}) \approx \cdots \approx \mathrm{var}(X_{(d)}) \approx 0$$

And $\varphi(x) = (x_{(1)}, \ldots, x_{(k)}) \in \mathbb{R}^k$

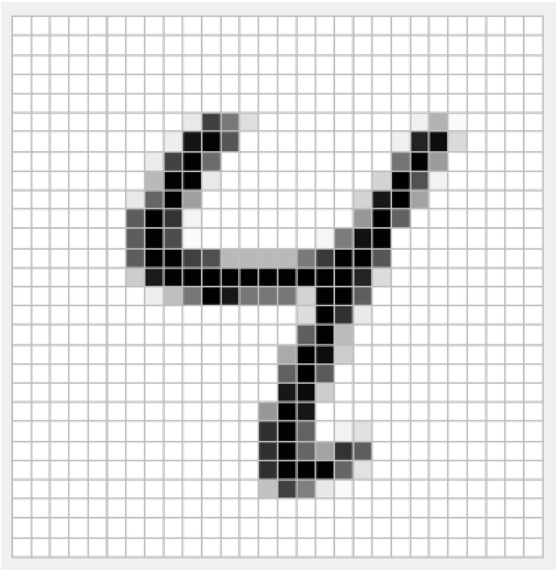
For affine function $w^\top x + b$, we have

$$w^\top X + b \approx$$

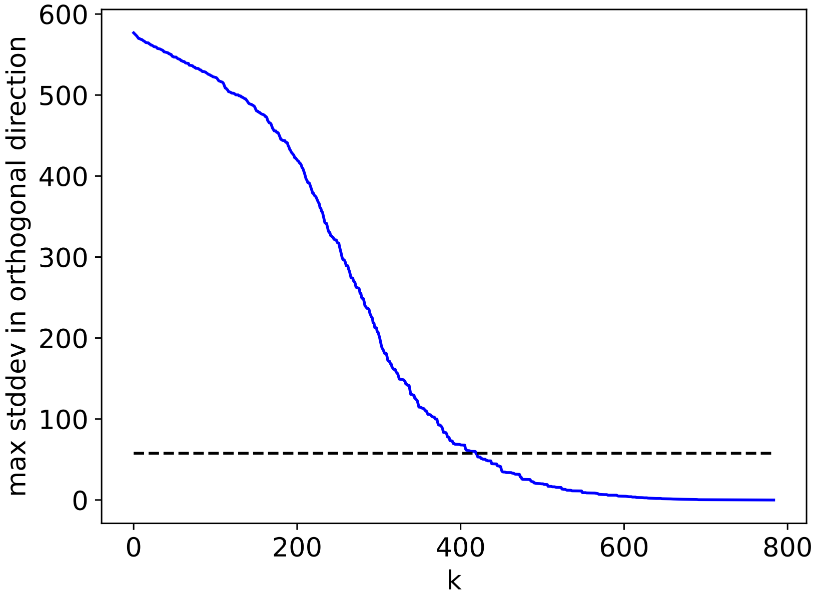
Therefore, this is close to $\tilde{w}^\top \varphi(X) + \tilde{b}$ for some $\tilde{w} \in \mathbb{R}^k$ and $\tilde{b} \in \mathbb{R}$

Example: MNIST dataset of handwritten digit images

- 784 features corresponding to pixel intensity values (from $\{0, 1, \ldots, 255\}$)



Vertical axis: $$\max_{\beta \in \mathbb{R}^{d-k}} \frac{\mathrm{stddev}(\beta_{k+1} X_{(k+1)} + \cdots + \beta_d X_{(d)})}{\|\beta\|}$$



Can we do better than "axis-aligned embeddings"?

- Maybe there is a better way to choose which variables to keep?
- Retained features could contain a lot of redundancy!
- Can possibly reduce dimension even further by accounting for covariance between features

Covariance matrices

Covariance matrix $\operatorname{cov}(X)$ of a random vector $X = (X_1, \ldots, X_d)$:

- $d \times d$ matrix whose (i, j)-th entry is $\operatorname{cov}(X_i, X_j)$
- Matrix notation:

$$\operatorname{cov}(X) = \mathbb{E}[(X - \mathbb{E}(X))(X - \mathbb{E}(X))^{\mathsf{T}}]$$

- $\operatorname{cov}(X)$ "encodes" covariance between all linear functions of X

Consider linear function $f(x) = \alpha^{\mathsf{T}} x$, given by some $\alpha \in \mathbb{R}^d$

- If α is a unit vector (i.e., $\|\alpha\| = 1$), then $\alpha^{\mathsf{T}} x$ is the "coordinate" of the orthogonal projection of x to the line spanned by α
- The "coordinate" $\alpha^{\mathsf{T}} x$ is often referred to as the "projection of x in direction α", even though this is not technically correct

- What is the mean of $\alpha^{\mathsf{T}} X$?

- What is the variance of $\alpha^{\mathsf{T}} X$?

- What is the covariance between $\alpha^{\mathsf{T}} X$ and $\beta^{\mathsf{T}} X$?

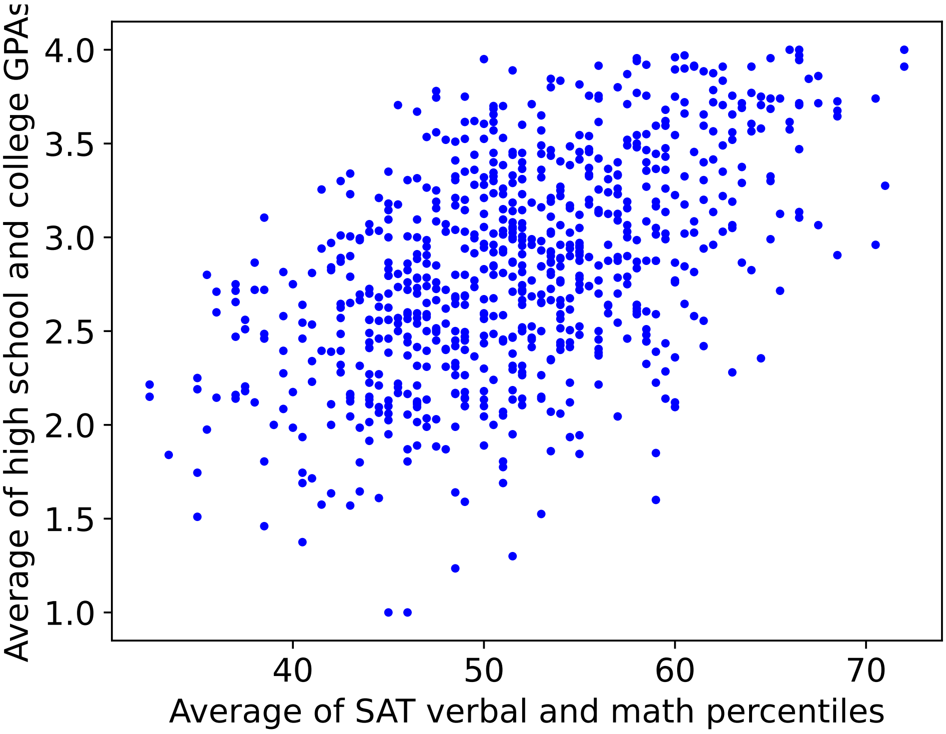
Example: Dartmouth student data

- x_1 = SAT verbal percentile, x_2 = SAT math percentile, x_3 = high school GPA, x_4 = (first year) college GPA
- X = data point drawn uniformly at random from dataset

$$\operatorname{cov}(X) = \begin{bmatrix} 69.8 & 33.8 & 1.74 & 2.71 \\ 33.8 & 72.3 & 1.76 & 2.43 \\ 1.74 & 1.76 & 0.29 & 0.22 \\ 2.71 & 2.43 & 0.22 & 0.56 \end{bmatrix}$$

- Define random variables Y and Z:

$$Y = \frac{1}{2}(\text{SAT verbal} + \text{SAT math})$$
$$Z = \frac{1}{2}(\text{high school GPA} + \text{college GPA})$$



Using $\mathrm{cov}(X)$, can compute $\mathrm{cor}(Y, Z)$:

$$\mathrm{var}(Y) = \alpha^\mathsf{T} \mathrm{cov}(X)\alpha = 52.4$$
$$\mathrm{var}(Z) = \beta^\mathsf{T} \mathrm{cov}(X)\beta = 0.32$$
$$\mathrm{cov}(Y, Z) = \alpha^\mathsf{T} \mathrm{cov}(X)\beta = 2.16$$
$$\mathrm{cor}(Y, Z) = \frac{\mathrm{cov}(Y, Z)}{\sqrt{\mathrm{var}(Y)\,\mathrm{var}(Z)}} = 0.52$$

where

$$\alpha = \underline{\hspace{3cm}}$$

$$\beta = \underline{\hspace{3cm}}$$

Review of eigenvalues and eigenvectors

- Every symmetric $d \times d$ matrix M has d real eigenvalues, conventionally numbered in non-increasing order

$$\lambda_1 \geq \cdots \geq \lambda_d$$

- Because M is symmetric, it is always possible to find d corresponding eigenvectors that form an orthonormal basis for $\mathbb{R}^d$:

$$v_1, \ldots, v_d \in \mathbb{R}^d$$

such that

$$M v_i = \lambda_i v_i$$

and

$$v_i^\mathsf{T} v_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

Eigendecomposition of M

$$M = \sum_{i=1}^{d} \lambda_i \, v_i v_i^{\mathsf{T}}$$

For rest of lecture, let $\mathrm{cov}(X)$ have eigendecomposition

$$\mathrm{cov}(X) = \sum_{i=1}^{d} \lambda_i \, v_i v_i^{\mathsf{T}}$$

with $\lambda_1 \geq \cdots \geq \lambda_d$ and $v_1, \ldots, v_d$ orthonormal

Variance maximizing direction

"Variance of X in direction α":

$$\operatorname{var}\left(\frac{1}{\|\alpha\|}\alpha^\intercal X\right) = \frac{\alpha^\intercal \operatorname{cov}(X)\alpha}{\|\alpha\|^2}$$

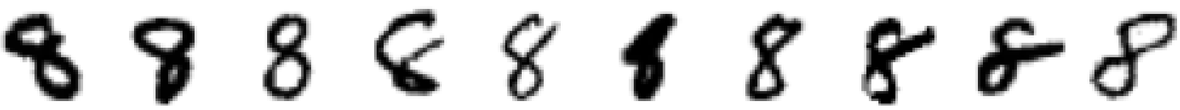
Question: In which direction α does X have the highest variance?

$$\max_{\alpha\in\mathbb{R}^d\setminus\{0\}} \frac{\alpha^\intercal \operatorname{cov}(X)\alpha}{\|\alpha\|^2}$$

Answer: $\alpha = v_1$—i.e., eigenvector of $\mathrm{cov}(X)$ corresponding to largest eigenvalue (a.k.a. top eigenvector)

Upshot: If you want to reduce to dimension $k = 1$, use direction of the top eigenvector of $\mathrm{cov}(X)$

Example: MNIST (just the 8's); 10 images sorted by "coordinate" along v_1



Principal components analysis

What we want: minimize variance of X in directions that are "thrown away"

For $k = 1$, goal is captured by following problem:

$$\min_{\alpha \in \mathbb{R}^d} \max_{\substack{\beta \in \mathbb{R}^d \setminus \{0\}, \\ \beta \perp \alpha}} \frac{\beta^\mathsf{T} \operatorname{cov}(X) \beta}{\|\beta\|^2}$$

Solution also is given by $\alpha = v_1$

This fact is a special case of the "Courant min-max principle"

- For $\alpha = v_1$,

$$\max_{\substack{\beta \in \mathbb{R}^d \setminus \{0\}, \\ \beta \perp \alpha}} \frac{\beta^\mathsf{T} \operatorname{cov}(X) \beta}{\|\beta\|^2} = \underline{\hspace{2cm}}$$

- For any other α:

Courant min-max principle says

$$\min_{\substack{\mathcal{W} \subseteq \mathbb{R}^d, \\ \dim(\mathcal{W}) = k}} \max_{\substack{\beta \in \mathbb{R}^d \setminus \{0\}, \\ \beta \perp \mathcal{W}}} \frac{\beta^\mathsf{T} \operatorname{cov}(X) \beta}{\|\beta\|^2} = \underline{\hspace{3cm}}$$

and this is achieved by the subspace $\mathcal{W} = \operatorname{span}\{v_1, \ldots, v_k\}$ spanned by top-k eigenvectors of $\operatorname{cov}(X)$

Principal components analysis (PCA): dimension reduction method that, for target dimension k, uses the linear map

$$\varphi(x) = (v_1^\mathsf{T} x, \ldots, v_k^\mathsf{T} x)$$

based on the top-k eigenvectors of $\operatorname{cov}(X)$

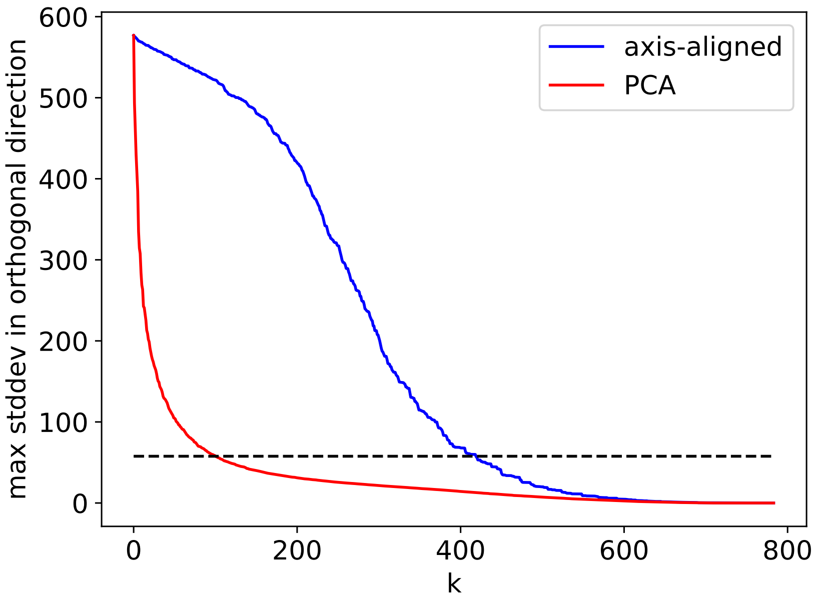
- $\varphi(x)$ gives the "coordinates" of the orthogonal projection of x to span of $v_1, \ldots, v_k$, a.k.a. the dimension-k PCA projection
- Also

$$\operatorname{cov}(\varphi(X)_i, \varphi(X)_j) = \underline{\hspace{3cm}}$$

So new "variables" in $\varphi(X)$ are uncorrelated

MNIST: What subspace dimension k is needed so worst standard deviation in an orthogonal direction is at most $0.1 \times \lambda_1$?

- Axis-aligned embeddings: $k = 419$; PCA embeddings: $k = 101$



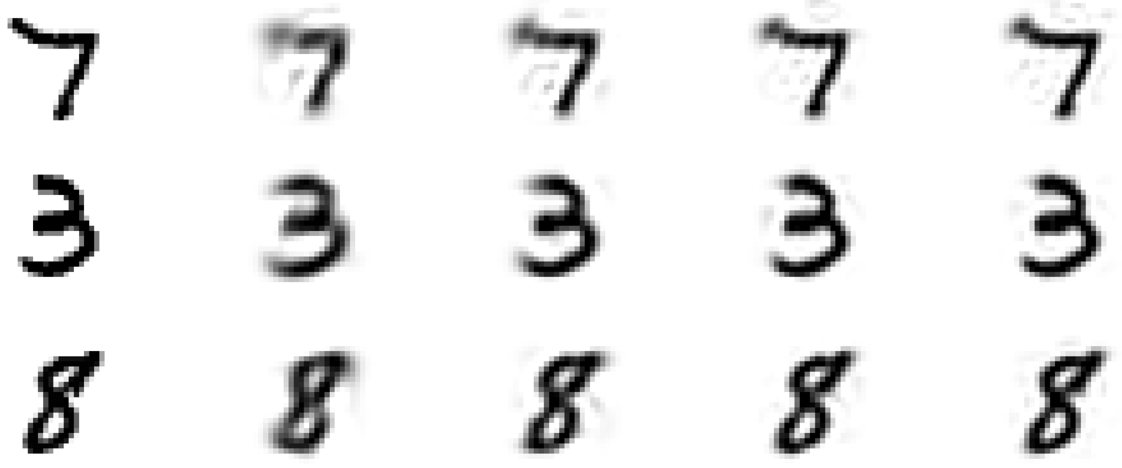
Given $\varphi(x) \in \mathbb{R}^k$ (from PCA), along with $v_1, \ldots, v_k$, can obtain d-dimensional "reconstruction" of x:

$$\sum_{i=1}^{k} \varphi(x)_i \, v_i$$

(orthogonal projection of x to the subspace spanned by $v_1, \ldots, v_k$)

MNIST

original	$k = 25$	$k = 50$	$k = 75$	$k = 100$



Matrix approximation

PCA (on finite dataset) is related to singular value decomposition of $n \times d$ matrix

$$A = \begin{bmatrix} \longleftarrow & (x^{(1)})^\mathsf{T} & \longrightarrow \\ & \vdots & \\ \longleftarrow & (x^{(n)})^\mathsf{T} & \longrightarrow \end{bmatrix}$$

Every matrix A has a singular value decomposition (SVD): decomposition of A into the sum of r rank-1 matrices

$$A = \sum_{i=1}^{r} s_i u^{(i)} (v^{(i)})^\intercal$$

where

- $r = \operatorname{rank}(A)$
- $s_1 \geq \cdots \geq s_r > 0$ as positive real numbers (singular values of A)
- $u^{(1)}, \ldots, u^{(r)}$ is ONB for CS(A) (left singular vectors of A)
- $v^{(1)}, \ldots, v^{(r)}$ is ONB for CS$(A^\intercal)$ (right singular vectors of A)

Matrix form of SVD:

$$A = \underbrace{\begin{bmatrix} \uparrow & & \uparrow \\ u^{(1)} & \cdots & u^{(r)} \\ \downarrow & & \downarrow \end{bmatrix}}_{U} \underbrace{\begin{bmatrix} s_1 & & \\ & \ddots & \\ & & s_r \end{bmatrix}}_{S} \underbrace{\begin{bmatrix} \longleftarrow & (v^{(1)})^\intercal & \longrightarrow \\ & \vdots & \\ \longleftarrow & (v^{(r)})^\intercal & \longrightarrow \end{bmatrix}}_{V^\intercal}$$

Computation: `numpy.linalg.svd`

Rank-k (truncated) SVD: keep only the first $k \le r$ components of the SVD

$$A^{(k)} = \sum_{i=1}^{k} s_i u^{(i)} (v^{(i)})^{\mathsf{T}}$$

In matrix form:

$$A^{(k)} = \underbrace{\begin{bmatrix} \uparrow & & \uparrow \\ u^{(1)} & \cdots & u^{(k)} \\ \downarrow & & \downarrow \end{bmatrix}}_{U^{(k)}} \underbrace{\begin{bmatrix} s_1 & & \\ & \ddots & \\ & & s_k \end{bmatrix}}_{S^{(k)}} \underbrace{\begin{bmatrix} \longleftarrow & (v^{(1)})^{\mathsf{T}} & \longrightarrow \\ & \vdots & \\ \longleftarrow & (v^{(k)})^{\mathsf{T}} & \longrightarrow \end{bmatrix}}_{(V^{(k)})^{\mathsf{T}}}$$

Eckart-Young Theorem: If $k \le \mathrm{rank}(A)$, then $A^{(k)} = \sum_{i=1}^{k} s_i u^{(i)} (v^{(i)})^{\mathsf{T}}$ from rank-k SVD has smallest sum-of-squared errors

$$\sum_{i=1}^{n} \sum_{j=1}^{d} (A_{i,j} - \tilde{A}_{i,j})^2$$

among all $n \times d$ matrices $\tilde{A}$ of rank k

Connection to PCA: Let X be random vector with uniform distribution over $\{x^{(1)}, \dots, x^{(n)}\}$ (and assume A is row-centered, so $\frac{1}{n}\sum_{i=1}^{n} x^{(i)} = 0$)

- Then $\mathrm{cov}(X) =$ ________
- Moreover,

$$A^{\mathsf{T}} A = \_\_\_\_\_\_\_\_\_$$

- Non-zero eigenvalues of $\mathrm{cov}(X)$ are ________________
- Corresponding eigenvectors of $\mathrm{cov}(X)$ are ______________

Statistical model: A is $n \times d$ matrix of independent random variables, with

$$A_{i,j} \sim \mathrm{N}(H_{i,j}, \sigma^2)$$

where H is $n \times d$ matrix with rank $\leq k$ (the "parameter" of this model)

Maximum likelihood estimator of H: __________________________________

J Novembre et al. Nature 000, 1-4 (2008) doi:10.1038/nature07331