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 $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}^k$

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 $S$
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

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

and choose $\varphi(x) = (x_{(1)}, \ldots, x_{(k)})$
Suppose only $k$ features have non-negligible variance

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

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

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

$$w^T X + b \approx \tilde{w}^T \varphi(X) + \tilde{b}$$

Therefore, this is close to $\tilde{w}^T \varphi(X) + \tilde{b}$ 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: \[
\max_{\beta \in \mathbb{R}^{d-k}} \frac{\text{stddev}(\beta_{k+1}X_{(k+1)} + \cdots + \beta_dX_{(d)})}{\|\beta\|}
\]
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** $\text{cov}(X)$ of a random vector $X = (X_1, \ldots, X_d)$:

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

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

- $\text{cov}(X)$ “encodes” covariance between all linear functions of $X$
Consider linear function $f(x) = \alpha^T x$, given by some $\alpha \in \mathbb{R}^d$

- If $\alpha$ is a unit vector (i.e., $\|\alpha\| = 1$), then $\alpha^T x$ is the “coordinate” of the orthogonal projection of $x$ to the line spanned by $\alpha$

- The “coordinate” $\alpha^T x$ is often referred to as the “projection of $x$ in direction $\alpha$”, even though this is not technically correct
What is the mean of $\alpha^\top X$?

What is the variance of $\alpha^\top X$?
What is the covariance between $\alpha^T X$ and $\beta^T X$?
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

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

- 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 $\text{cov}(X)$, can compute $\text{cor}(Y, Z)$:

$$\text{var}(Y) = \alpha^\top \text{cov}(X) \alpha = 52.4$$

$$\text{var}(Z) = \beta^\top \text{cov}(X) \beta = 0.32$$

$$\text{cov}(Y, Z) = \alpha^\top \text{cov}(X) \beta = 2.16$$

$$\text{cor}(Y, Z) = \frac{\text{cov}(Y, Z)}{\sqrt{\text{var}(Y) \text{var}(Z)}} = 0.52$$

where

$$\alpha = \text{__________}$$

$$\beta = \text{__________}$$
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

$$Mv_i = \lambda_i v_i$$

and

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

$$M = \sum_{i=1}^{d} \lambda_i v_i v_i^T$$
For rest of lecture, let $\text{cov}(X)$ have eigendecomposition

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

with $\lambda_1 \geq \cdots \geq \lambda_d$ and $v_1, \ldots, v_d$ orthonormal.
Variance maximizing direction
“Variance of $X$ in direction $\alpha$”:

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

Question: In which direction $\alpha$ does $X$ have the highest variance?

$$\max_{\alpha \in \mathbb{R}^d \setminus \{0\}} \frac{\alpha^T \text{cov}(X) \alpha}{\|\alpha\|^2}$$
Answer: $\alpha = v_1$—i.e., eigenvector of $\text{cov}(X)$ 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 $\text{cov}(X)$

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:

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

Solution also is given by $\alpha = v_1$

This fact is a special case of the “Courant min-max principle”
For $\alpha = v_1$, 

$$
\max_{\beta \in \mathbb{R}^d \setminus \{0\}, \beta \perp \alpha} \frac{\beta^T \text{cov}(X) \beta}{\|\beta\|^2} = \quad \text{(expression to be filled)}
$$

For any other $\alpha$: 

...
Courant min-max principle says

$$\min_{\mathcal{W} \subseteq \mathbb{R}^d, \dim(\mathcal{W}) = k} \max_{\beta \in \mathbb{R}^d \setminus \{0\}, \beta \perp \mathcal{W}} \frac{\beta^T \text{cov}(X) \beta}{\|\beta\|^2} = \ldots$$

and this is achieved by the subspace $\mathcal{W} = \text{span}\{v_1, \ldots, v_k\}$ spanned by top-$k$ eigenvectors of $\text{cov}(X)$.
**Principal components analysis (PCA):** dimension reduction method that, for target dimension $k$, uses the linear map

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

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

- $\varphi(x)$ 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

$$\text{cov}(\varphi(X)_i, \varphi(X)_j) =$$

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 $v_1, \ldots, v_k$)
<table>
<thead>
<tr>
<th></th>
<th>original</th>
<th>( k = 25 )</th>
<th>( k = 50 )</th>
<th>( k = 75 )</th>
<th>( k = 100 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
</tbody>
</table>
Matrix approximation
PCA (on finite dataset) is related to singular value decomposition of $n \times d$ matrix

$$A = \begin{bmatrix}
\leftarrow & (x^{(1)})^T & \rightarrow \\
\vdots \\
\leftarrow & (x^{(n)})^T & \rightarrow 
\end{bmatrix}$$
Every matrix $A$ has a **singular value decomposition (SVD)**: decomposition of $A$ into the sum of $r$ rank-1 matrices

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

where

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


Matrix form of SVD:

\[ A = \begin{bmatrix} u^{(1)} & \cdots & u^{(r)} \end{bmatrix} \begin{bmatrix} s_1 & \cdots & \vdots \\ \vdots & \ddots & \vdots \\ \vdots & \cdots & s_r \end{bmatrix} \begin{bmatrix} (v^{(1)})^T \\ \vdots \\ (v^{(r)})^T \end{bmatrix} \]

Computation: `numpy.linalg.svd`
**Rank-\(k\) (truncated) SVD**: keep only the first \(k \leq r\) components of the SVD

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

In matrix form:

\[
A^{(k)} = 
\begin{bmatrix}
u^{(1)} & \ldots & u^{(k)}
\end{bmatrix} 
\begin{bmatrix}
s_1 & \ldots & 0 \\
0 & \ldots & 0
\end{bmatrix} 
\begin{bmatrix}
(v^{(1)})^T & \ldots & (v^{(k)})^T
\end{bmatrix}
\]

\[
\begin{bmatrix}
U^{(k)} \\
S^{(k)} \\
V^{(k)}
\end{bmatrix}
\]
**Eckart-Young Theorem**: If $k \leq \text{rank}(A)$, then $A^{(k)} = \sum_{i=1}^{k} s_i u^{(i)} (v^{(i)})^\top$ 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 \{${x^{(1)}, \ldots, x^{(n)}}$\} (and assume $A$ is row-centered, so $\frac{1}{n} \sum_{i=1}^{n} x^{(i)} = 0$)

▶ Then $\text{cov}(X) =$

▶ Moreover,

$$A^\top A =$$

▶ Non-zero eigenvalues of $\text{cov}(X)$ are

▶ Corresponding eigenvectors of $\text{cov}(X)$ are
Statistical model: $A$ is $n \times d$ matrix of independent random variables, with

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

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

Maximum likelihood estimator of $H$: ____________________________