

COMPSCI 514: ALGORITHMS FOR DATA SCIENCE

Cameron Musco

University of Massachusetts Amherst. Fall 2020.

Lecture 15

- Problem Set 3 is due next Friday 10/23, 8pm.
- Problem set grades seem to be strongly correlated with whether people are working in groups. So if you don't have a group, I encourage you to join one. There are multiple people looking so post on Piazza to find some.
- This week's quiz due Monday at 8pm.

Last Class: Low-Rank Approximation

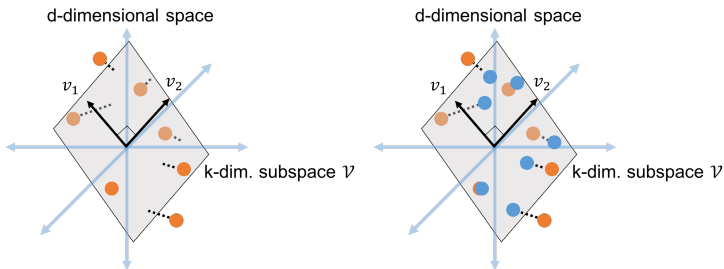
- When data lies in a k -dimensional subspace \mathcal{V} , we can perfectly embed into k dimensions using an orthonormal span $\mathbf{V} \in \mathbb{R}^{d \times k}$.
- When data lies **close** to \mathcal{V} , the optimal embedding in that space is given by projecting onto that space.

$$\mathbf{XV}^T = \underset{\mathbf{B} \text{ with rows in } \mathcal{V}}{\operatorname{arg\,min}} \|\mathbf{X} - \mathbf{B}\|_F^2.$$

This Class: Finding \mathcal{V} via eigendecomposition.

- How do we find the best low-dimensional subspace to approximate \mathbf{X} ?
- PCA and its connection to eigendecomposition.

Reminder of Set Up: Assume that $\vec{x}_1, \dots, \vec{x}_n$ lie close to any k -dimensional subspace \mathcal{V} of \mathbb{R}^d . Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be the data matrix.

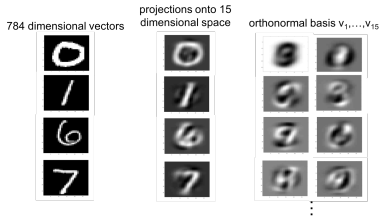


Let $\vec{v}_1, \dots, \vec{v}_k$ be an orthonormal basis for \mathcal{V} and $\mathbf{V} \in \mathbb{R}^{d \times k}$ be the matrix with these vectors as its columns.

- $\mathbf{W}^T \in \mathbb{R}^{d \times d}$ is the **projection matrix** onto \mathcal{V} .
- $\mathbf{X} \approx \mathbf{X}(\mathbf{W}^T)$. Gives the closest approximation to \mathbf{X} with rows in \mathcal{V} .

$\vec{x}_1, \dots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \dots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for subspace \mathcal{V} . $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \dots, \vec{v}_k$.

DUAL VIEW OF LOW-RANK APPROXIMATION



Row (data point) compression

Column (feature) compression

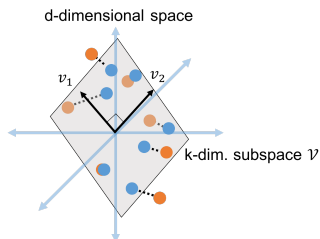
$10000 * \text{bathrooms} + 10 * (\text{sq. ft.}) \approx \text{list price}$

	bedrooms	bathrooms	sq.ft.	floors	list price	sale price
home 1	2	2	1800	2	200,000	195,000
home 2	4	2.5	2700	1	300,000	310,000
.
.
.
home n	5	3.5	3600	3	450,000	450,000

If $\vec{x}_1, \dots, \vec{x}_n$ are close to a k -dimensional subspace \mathcal{V} with orthonormal basis $\mathbf{V} \in \mathbb{R}^{d \times k}$, the data matrix can be approximated as $\mathbf{XV}\mathbf{V}^T$. \mathbf{XV} gives optimal embedding of \mathbf{X} in \mathcal{V} .

How do we find \mathcal{V} (equivalently \mathbf{V})?

$$\arg \min_{\text{orthonormal } \mathbf{V} \in \mathbb{R}^{d \times k}} \|\mathbf{X} - \mathbf{XV}\mathbf{V}^T\|_F^2 = \arg \max_{\text{orthonormal } \mathbf{V} \in \mathbb{R}^{d \times k}} \|\mathbf{XV}\|_F^2.$$



SOLUTION VIA EIGENDECOMPOSITION

\mathbf{V} minimizing $\|\mathbf{X} - \mathbf{X}\mathbf{V}\mathbf{V}^T\|_F^2$ is given by:

$$\arg \max_{\text{orthonormal } \mathbf{V} \in \mathbb{R}^{d \times k}} \|\mathbf{X}\mathbf{V}\|_F^2 = \sum_{i=1}^n \|\mathbf{V}^T \vec{x}_i\|_2^2 = \sum_{j=1}^k \|\mathbf{X}\vec{v}_j\|_2^2$$

Surprisingly, can find the columns of \mathbf{V} , $\vec{v}_1, \dots, \vec{v}_k$ **greedily**.

$$\vec{v}_1 = \arg \max_{\vec{v} \text{ with } \|\vec{v}\|_2=1} \|\mathbf{X}\vec{v}\|_2^2 \vec{v}^T \mathbf{X}^T \mathbf{X} \vec{v}.$$

$$\vec{v}_2 = \arg \max_{\vec{v} \text{ with } \|\vec{v}\|_2=1, \langle \vec{v}, \vec{v}_1 \rangle = 0} \vec{v}^T \mathbf{X}^T \mathbf{X} \vec{v}.$$

...

$$\vec{v}_k = \arg \max_{\vec{v} \text{ with } \|\vec{v}\|_2=1, \langle \vec{v}, \vec{v}_j \rangle = 0 \ \forall j < k} \vec{v}^T \mathbf{X}^T \mathbf{X} \vec{v}.$$

These are exactly the top k eigenvectors of $\mathbf{X}^T \mathbf{X}$.

$\vec{x}_1, \dots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \dots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for subspace \mathcal{V} . $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \dots, \vec{v}_k$.

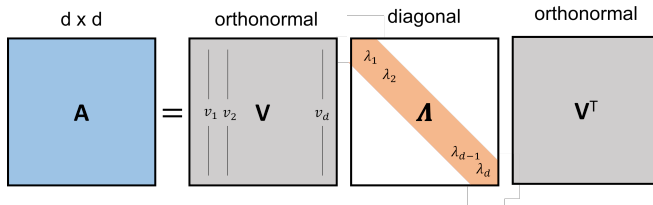
Eigenvector: $\vec{x} \in \mathbb{R}^d$ is an eigenvector of a matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ if $\mathbf{A}\vec{x} = \lambda\vec{x}$ for some scalar λ (the eigenvalue corresponding to \vec{x}).

- That is, \mathbf{A} just ‘stretches’ x .
- If \mathbf{A} is **symmetric**, can find d orthonormal eigenvectors $\vec{v}_1, \dots, \vec{v}_d$. Let $\mathbf{V} \in \mathbb{R}^{d \times d}$ have these vectors as columns.

$$\mathbf{AV} = \begin{bmatrix} | & | & | & | \\ \mathbf{A}\vec{v}_1 & \mathbf{A}\vec{v}_2 & \cdots & \mathbf{A}\vec{v}_d \\ | & | & | & | \end{bmatrix} = \begin{bmatrix} | & | & | & | \\ \lambda_1\vec{v}_1 & \lambda_2\vec{v}_2 & \cdots & \lambda\vec{v}_d \\ | & | & | & | \end{bmatrix} = \mathbf{V}\mathbf{\Lambda}$$

Yields eigendecomposition: $\mathbf{AVV}^T = \mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$.

REVIEW OF EIGENVECTORS AND EIGENDECOMPOSITION



Typically order the eigenvectors in decreasing order:

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d.$$

Courant-Fischer Principal: For symmetric \mathbf{A} , the eigenvectors are given via the greedy optimization:

$$\vec{v}_1 = \arg \max_{\vec{v} \text{ with } \|\vec{v}\|_2=1} \vec{v}^T \mathbf{A} \vec{v}.$$

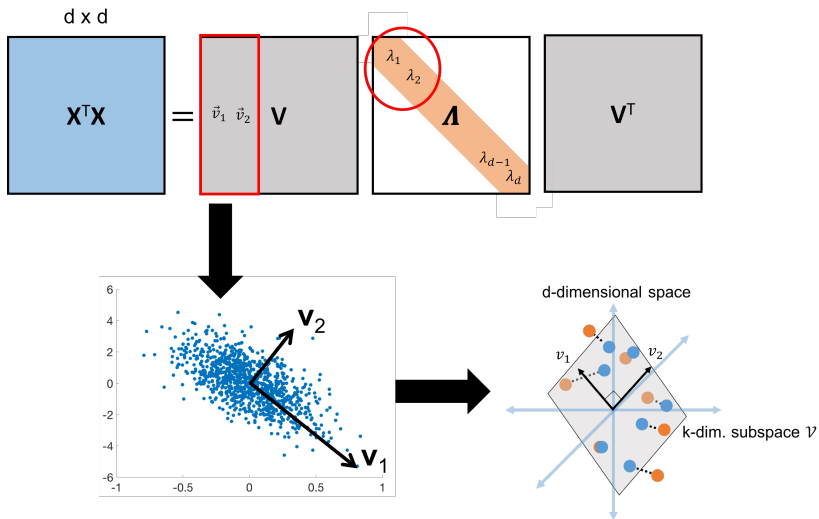
$$\vec{v}_2 = \arg \max_{\vec{v} \text{ with } \|\vec{v}\|_2=1, \langle \vec{v}, \vec{v}_1 \rangle = 0} \vec{v}^T \mathbf{A} \vec{v}.$$

...

$$\vec{v}_d = \arg \max_{\vec{v} \text{ with } \|\vec{v}\|_2=1, \langle \vec{v}, \vec{v}_j \rangle = 0 \ \forall j < d} \vec{v}^T \mathbf{A} \vec{v}.$$

- $\vec{v}_j^T \mathbf{A} \vec{v}_j = \lambda_j \cdot \vec{v}_j^T \vec{v}_j = \lambda_j$, the j^{th} largest eigenvalue.
- The first k eigenvectors of $\mathbf{X}^T \mathbf{X}$ (corresponding to the largest k eigenvalues) are exactly the directions of greatest variance in \mathbf{X} that we use for low-rank approximation.

LOW-RANK APPROXIMATION VIA EIGENDECOMPOSITION



Upshot: Letting \mathbf{V}_k have columns $\vec{v}_1, \dots, \vec{v}_k$ corresponding to the top k eigenvectors of the covariance matrix $\mathbf{X}^T\mathbf{X}$, \mathbf{V}_k is the orthogonal basis minimizing

$$\|\mathbf{X} - \mathbf{X}\mathbf{V}_k\mathbf{V}_k^T\|_F^2,$$

This is principal component analysis (PCA).

How accurate is this low-rank approximation? Can understand using eigenvalues of $\mathbf{X}^T\mathbf{X}$.

$\vec{x}_1, \dots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \dots, \vec{v}_k \in \mathbb{R}^d$: top eigenvectors of $\mathbf{X}^T\mathbf{X}$, $\mathbf{V}_k \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \dots, \vec{v}_k$.

Let $\vec{v}_1, \dots, \vec{v}_k$ be the top k eigenvectors of $\mathbf{X}^T\mathbf{X}$ (the top k principal components). Approximation error is:

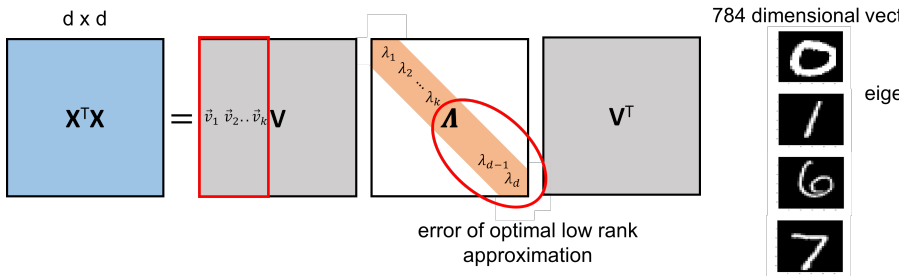
$$\begin{aligned} \|\mathbf{X} - \mathbf{X}\mathbf{V}_k\mathbf{V}_k^T\|_F^2 &= \|\mathbf{X}\|_F^2 \operatorname{tr}(\mathbf{X}^T\mathbf{X}) - \|\mathbf{X}\mathbf{V}_k\mathbf{V}_k^T\|_F^2 \operatorname{tr}(\mathbf{V}_k^T\mathbf{X}^T\mathbf{X}\mathbf{V}_k) \\ &= \sum_{i=1}^d \lambda_i(\mathbf{X}^T\mathbf{X}) - \sum_{i=1}^k \vec{v}_i^T \mathbf{X}^T\mathbf{X} \vec{v}_i \\ &= \sum_{i=1}^d \lambda_i(\mathbf{X}^T\mathbf{X}) - \sum_{i=1}^k \lambda_i(\mathbf{X}^T\mathbf{X}) = \sum_{i=k+1}^d \lambda_i(\mathbf{X}^T\mathbf{X}) \end{aligned}$$

- For any matrix \mathbf{A} , $\|\mathbf{A}\|_F^2 = \sum_{i=1}^d \|\vec{a}_i\|_2^2 = \operatorname{tr}(\mathbf{A}^T\mathbf{A})$ (sum of diagonal entries = sum eigenvalues).

$\vec{x}_1, \dots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \dots, \vec{v}_k \in \mathbb{R}^d$: top eigenvectors of $\mathbf{X}^T\mathbf{X}$, $\mathbf{V}_k \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \dots, \vec{v}_k$.

Claim: The error in approximating \mathbf{X} with the best rank k approximation (projecting onto the top k eigenvectors of $\mathbf{X}^T\mathbf{X}$ is:

$$\|\mathbf{X} - \mathbf{X}\mathbf{V}_k\mathbf{V}_k^T\|_F^2 = \sum_{i=k+1}^d \lambda_i(\mathbf{X}^T\mathbf{X})$$

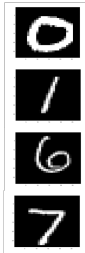


$\vec{x}_1, \dots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \dots, \vec{v}_k \in \mathbb{R}^d$: top eigenvectors of $\mathbf{X}^T\mathbf{X}$, $\mathbf{V}_k \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \dots, \vec{v}_k$.

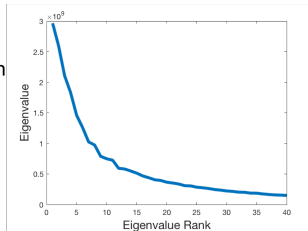
SPECTRUM ANALYSIS

Plotting the **spectrum** of the covariance matrix $\mathbf{X}^T\mathbf{X}$ (its eigenvalues) shows how compressible \mathbf{X} is using low-rank approximation (i.e., how close $\vec{x}_1, \dots, \vec{x}_n$ are to a low-dimensional subspace).

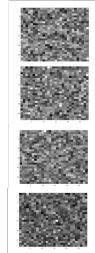
784 dimensional vectors



eigendecomposition



784 dimensional vectors



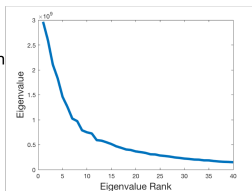
eigende

$\vec{x}_1, \dots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \dots, \vec{v}_k \in \mathbb{R}^d$: top eigenvectors of $\mathbf{X}^T\mathbf{X}$, $\mathbf{V}_k \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \dots, \vec{v}_k$.

784 dimensional vectors



eigendecomposition



Exercises:

1. Show that the eigenvalues of $\mathbf{X}^T\mathbf{X}$ are always positive. **Hint:** Use that $\lambda_j = \vec{v}_j^T \mathbf{X}^T \mathbf{X} \vec{v}_j$.
2. Show that for symmetric \mathbf{A} , the trace is the sum of eigenvalues: $\text{tr}(\mathbf{A}) = \sum_{i=1}^n \lambda_i(\mathbf{A})$.

- Many (most) datasets can be approximated via projection onto a low-dimensional subspace.
- Find this subspace via a maximization problem:

$$\max_{\text{orthonormal } \mathbf{V}} \|\mathbf{XV}\|_F^2.$$

- Greedy solution via eigendecomposition of $\mathbf{X}^T\mathbf{X}$.
- Columns of \mathbf{V} are the top eigenvectors of $\mathbf{X}^T\mathbf{X}$.
- Error of best low-rank approximation (compressibility of data) is determined by the tail of $\mathbf{X}^T\mathbf{X}$'s eigenvalue spectrum.

INTERPRETATION IN TERMS OF CORRELATION

Recall: Low-rank approximation is possible when our data features are correlated.

10000* bathrooms+ 10* (sq. ft.) \approx list price

	bedrooms	bathrooms	sq.ft.	floors	list price	sale price
home 1	2	2	1800	2	200,000	195,000
home 2	4	2.5	2700	1	300,000	310,000
.
.
.
home n	5	3.5	3600	3	450,000	450,000

Our compressed dataset is $\mathbf{C} = \mathbf{X}\mathbf{V}_k$ where the columns of \mathbf{V}_k are the top k eigenvectors of $\mathbf{X}^T\mathbf{X}$.

What is the covariance of \mathbf{C} ? $\mathbf{C}^T\mathbf{C} = \mathbf{V}_k^T\mathbf{X}^T\mathbf{X}\mathbf{V}_k = \mathbf{V}_k^T\mathbf{V}\mathbf{\Lambda}\mathbf{V}^T\mathbf{V}_k = \mathbf{\Lambda}_k$

Covariance becomes diagonal. I.e., all correlations have been removed. Maximal compression.

$\vec{x}_1, \dots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \dots, \vec{v}_k \in \mathbb{R}^d$: top eigenvectors of $\mathbf{X}^T\mathbf{X}$, $\mathbf{V}_k \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \dots, \vec{v}_k$.