COMPSCI 514: Algorithms for Data Science

Cameron Musco

University of Massachusetts Amherst. Fall 2024.

Lecture 16

Logistics

- Problem Set 3 due next Friday 11/8 at 11:59pm.
- There is no class next Tuesday due to election day. But I will hold my regular office hours from 2:30-3:30pm. Location TBD.
- Get any midterm regrade requests in by tomorrow.

Summary

Last Class:

- Finding an optimal orthogonal basis $\mathbf{V} \in \mathbb{R}^{d \times k}$ to minimize $\|\mathbf{X} - \mathbf{X}\mathbf{V}\mathbf{V}^T\|_F^2$ when the data does not exactly lie in a low-dimensional subspace.
- Solution by taking the top k eigenvectors of $\mathbf{X}^T\mathbf{X}$ (this is PCA/optimal low-rank approximation)
- Greedy optimization problem and connection to Courant-Fischer principal.

This Class:

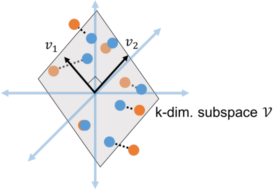
- Wrap up optimal low-rank approximation.
- Measuring the error of the low-rank approximation via covariance matrix eigenvalues.
- General linear algebra review.

Best Fit Subspace

If $\vec{x}_1, \ldots, \vec{x}_n$ 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 $\mathbf{X}\mathbf{V}\mathbf{V}^T$. $\mathbf{X}\mathbf{V}$ gives optimal embedding of $\mathbf{X}$ in $\mathcal{V}$.

We can find $\mathbf{V}$ by solving the optimization problem:

$$\underset{\text{orthonormal } \mathbf{V} \in \mathbb{R}^{d \times k}}{\arg\min} \|\mathbf{X} - \mathbf{X}\mathbf{V}\mathbf{V}^T\|_F^2 = \underset{\text{orthonormal } \mathbf{V} \in \mathbb{R}^{d \times k}}{\arg\max} \|\mathbf{X}\mathbf{V}\|_F^2 = \sum_{i=1}^{k} \|\mathbf{X}\vec{v}_i\|_2^2$$



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

Solution via Eigendecomposition

We can find the columns of $\mathbf{V}$, $\vec{v}_1, \ldots, \vec{v}_k$ greedily.

$$\vec{v}_1 = \underset{\vec{v} \text{ with } \|v\|_2 = 1}{\arg\max} \|\mathbf{X}\vec{v}\|_2^2$$

$$\vec{v}_2 = \underset{\vec{v} \text{ with } \|v\|_2 = 1,\ \langle \vec{v}, \vec{v}_1 \rangle = 0}{\arg\max} \|\mathbf{X}\vec{v}\|_2^2$$

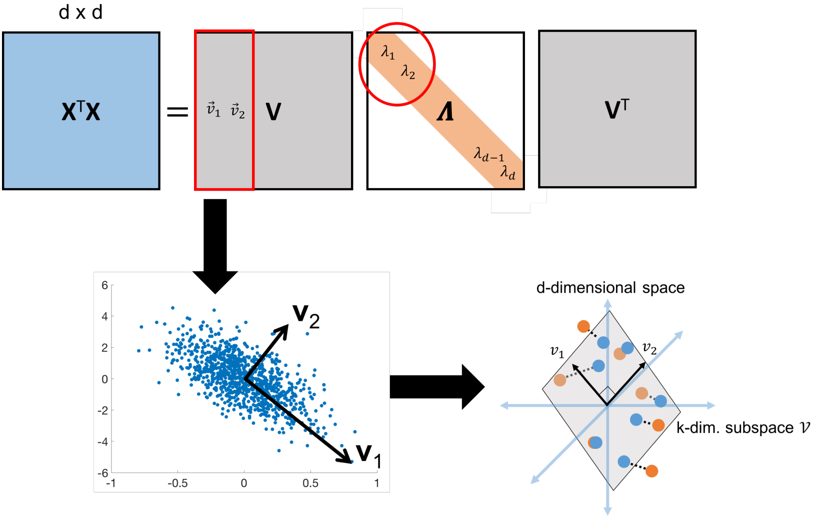
$$\cdots$$

$$\vec{v}_k = \underset{\vec{v} \text{ with } \|v\|_2 = 1,\ \langle \vec{v}, \vec{v}_j \rangle = 0\ \forall j < k}{\arg\max} \|\mathbf{X}\vec{v}\|_2^2.$$

$\vec{v}_1, \ldots, \vec{v}_k$ are the top k eigenvectors of $\mathbf{X}^T\mathbf{X}$ by the *Courant-Fischer Principle.*

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

Low-Rank Approximation via Eigendecomposition



Low-Rank Approximation via Eigendecomposition

Upshot: Letting $\mathbf{V}_k$ have columns $\vec{v}_1, \ldots, \vec{v}_k$ 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 $\mathbf{X}^T\mathbf{X}$.

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

Spectrum Analysis

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

$$\begin{aligned} \|X - XV_kV_k^T\|_F^2 = \|X\|_F^2 \operatorname{tr}(X^TX) - \|XV_kV_k^T\|_F^2 \operatorname{tr}(V_k^TX^TXV_k) \\ = \sum_{i=1}^{d} \lambda_i(X^TX) - \sum_{i=1}^{k} \vec{v}_i^T X^T X \vec{v}_i \\ = \sum_{i=1}^{d} \lambda_i(X^TX) - \sum_{i=1}^{k} \lambda_i(X^TX) = \sum_{i=k+1}^{d} \lambda_i(X^TX) \end{aligned}$$

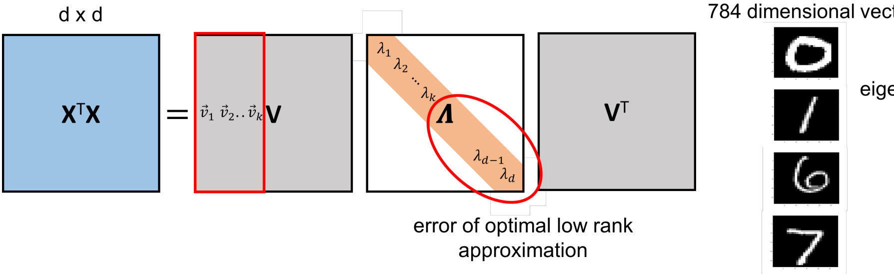
- **Exercise:** For any matrix A, $\|A\|_F^2 = \sum_{i=1}^d \|\vec{a}_i\|_2^2 = \operatorname{tr}(A^TA)$ (sum of diagonal entries = sum eigenvalues).

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

Spectrum Analysis

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:

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

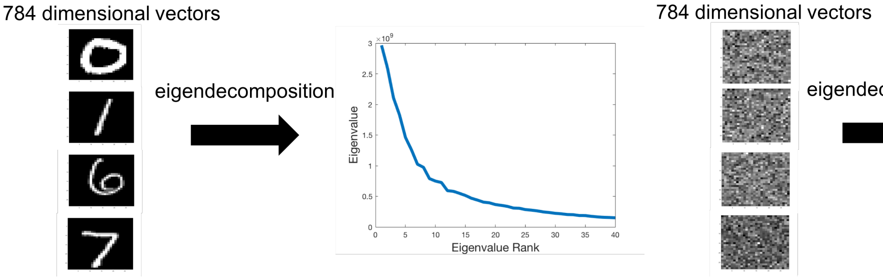


- Choose k to balance accuracy/compression – often at an ‘elbow’.

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

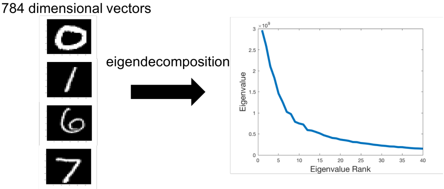
Spectrum Analysis

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



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

Spectrum Analysis



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: $\mathbf{tr}(\mathbf{A}) = \sum_{i=1}^{n} \lambda_i(\mathbf{A})$. **Hint:** First prove the cyclic property of trace, that for any $\mathbf{MN}$, $\mathbf{tr}(\mathbf{MN}) = \mathbf{tr}(\mathbf{NM})$ and then apply this to $\mathbf{A}$'s eigendecomposition

Summary

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

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

- 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.