COMPSCI 514: ALGORITHMS FOR DATA SCIENCE

Cameron Musco

University of Massachusetts Amherst. Fall 2020.

Lecture 13

- Midterm is this Thursday - Friday.
- Office Hours: I'll hold extra office hours, tomorrow from 2-3pm. The TAs will also hold their regular hours (see course page). Recordings of my office hours will be posted on Piazza.
- Logistics: Sometime on Thursday/Friday, you will download the exam in Gradescope, and should upload a pdf either of typed or handwritten answers 2 hours later. There will be a 15 minute buffer to upload in. Must submit by 11:59pm on Friday.
- Questions: Via private Piazza message. We'll try to answer frequently between 8am-10pm. If you don't get an answer, state any assumptions/interpretations you make clearly and move forward.
- Academic Honestly: You may not discuss the exam with any other students. Any cheating on the exam will result in failing the class. Please don't do this! It is much easier to catch than you might think, and the consequences seriously outweigh the benefits.

- Midterm is this Thursday - Friday.
- You must show you work/derive any answers to get full credit. Even on multiple choice questions.
- The exam is open notes. If you use outside resources (this should not be necessary) make sure to cite them.
- Very important to do some practice problems and to try them first with no resources, to simulate the exam.
- Make sure you can recognize when to apply the fundamentals: union bound, linearity of expectation and variance, Markov's inequality, Chebyshev's inequality, indicator random variables.
- Understand the goal of each algorithm/data structure. I.e., what problem it solves with what guarantees. No need to memorize proofs.

Last Few Classes:

The Johnson-Lindenstrauss Lemma

- Reduce n data points in any dimension d to $O\left(\frac{\log n/\delta}{\epsilon^2}\right)$ dimensions and preserve (with probability $\geq 1-\delta$) all pairwise distances up to $1 \pm \epsilon$.
- Compression is linear via multiplication with a random, data oblivious, matrix (linear compression)

High-Dimensional Geometry

- Why high-dimensional space is so different than low-dimensional space.
- How the JL Lemma can still work.

Next Few Classes: Low-rank approximation, the SVD, and principal component analysis (PCA).

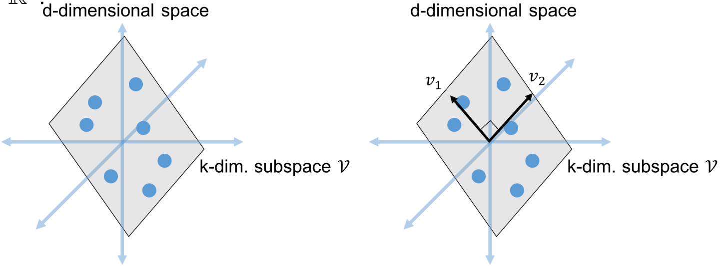
- Reduce d-dimesional data points to a smaller dimension m.
- Like JL, compression is linear – by applying a matrix.
- Chose this matrix carefully, taking into account structure of the dataset.
- Can give better compression than random projection.

Will be using a fair amount of linear algebra: orthogonal basis, column/row span, eigenvectors, etc,

RANDOMIZED ALGORITHMS UNIT TAKEAWAYS

- Randomization is an important tool in working with large datasets.
- Lets us solve 'easy' problems that get really difficult on massive datasets. Fast/space efficient look up (hash tables and bloom filters), distinct items counting, frequent items counting, near neighbor search (LSH), etc.
- The analysis of randomized algorithms leads to complex output distributions, which we can't compute exactly.
- We've covered many of the key ideas used through a small number of example applications/algorithms.
- We use concentration inequalities to bound these distributions and behaviors like accuracy, space usage, and runtime.
- Concentration inequalities and probability tools used in randomized algorithms are also fundamental in statistics, machine learning theory, probabilistic modeling of complex systems, etc.

Assume that data points $\vec{x}_1, \ldots, \vec{x}_n$ lie in any k-dimensional subspace $\mathcal{V}$ of $\mathbb{R}^d$.



Claim: Let $\vec{v}_1, \ldots, \vec{v}_k$ 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. For all $\vec{x}_i, \vec{x}_j$:

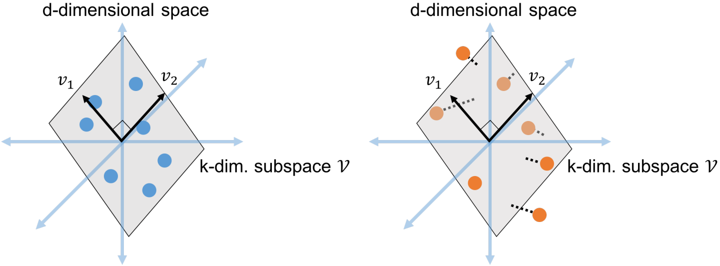
$$\|\mathbf{V}^T\vec{x}_i - \mathbf{V}^T\vec{x}_j\|_2 = \|\vec{x}_i - \vec{x}_j\|_2.$$

- $\mathbf{V}^T \in \mathbb{R}^{k \times d}$ is a linear embedding of $\vec{x}_1, \ldots, \vec{x}_n$ into k dimensions with no distortion.

Claim: Let $\vec{v}_1, \ldots, \vec{v}_k$ 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. For all $\vec{x}_i, \vec{x}_j \in \mathcal{V}$:

$$\|\mathbf{V}^T\vec{x}_i - \mathbf{V}^T\vec{x}_j\|_2 = \|\vec{x}_i - \vec{x}_j\|_2.$$

Main Focus of Upcoming Classes: Assume that data points $\vec{x}_1, \ldots, \vec{x}_n$ lie close to any k-dimensional subspace $\mathcal{V}$ of $\mathbb{R}^d$.



Letting $\vec{v}_1, \ldots, \vec{v}_k$ 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, $\mathbf{V}^T\vec{x}_i \in \mathbb{R}^k$ is still a good embedding for $x_i \in \mathbb{R}^d$. The key idea behind low-rank approximation and principal component analysis (PCA).

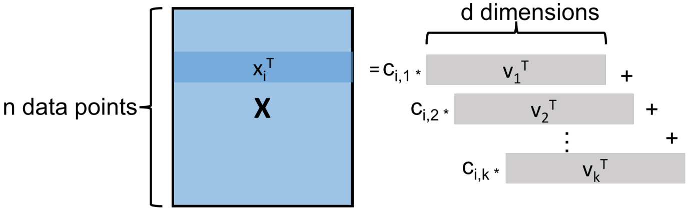
- How do we find $\mathcal{V}$ and $\mathbf{V}$?
- How good is the embedding?

Claim: $\vec{x}_1, \ldots, \vec{x}_n$ lie in a k-dimensional subspace $\mathcal{V} \Leftrightarrow$ the data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ has rank $\leq k$.

- Letting $\vec{v}_1, \ldots, \vec{v}_k$ be an orthonormal basis for $\mathcal{V}$, can write any $\vec{x}_i$ as:

$$\vec{x}_i = \mathbf{V}\vec{c}_i = c_{i,1} \cdot \vec{v}_1 + c_{i,2} \cdot \vec{v}_2 + \ldots + c_{i,k} \cdot \vec{v}_k.$$

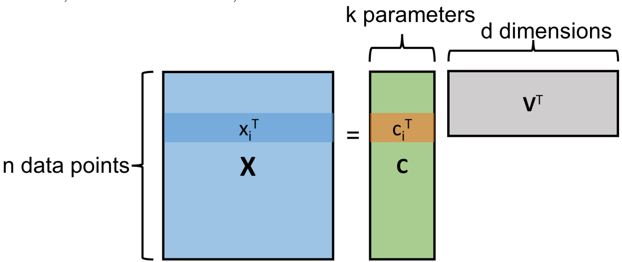
- So $\vec{v}_1, \ldots, \vec{v}_k$ span the rows of $\mathbf{X}$ and thus $\text{rank}(\mathbf{X}) \leq k$.



$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: 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 $\vec{v}_1, \ldots, \vec{v}_k$.

Claim: $\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$ lie in a k-dimensional subspace $\mathcal{V} \Leftrightarrow$ the data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ has rank $\leq k$.

- Every data point $\vec{x}_i$ (row of $\mathbf{X}$) can be written as $\vec{x}_i = \mathbf{V}\vec{c}_i = c_{i,1} \cdot \vec{v}_1 + \ldots + c_{i,k} \cdot \vec{v}_k$.

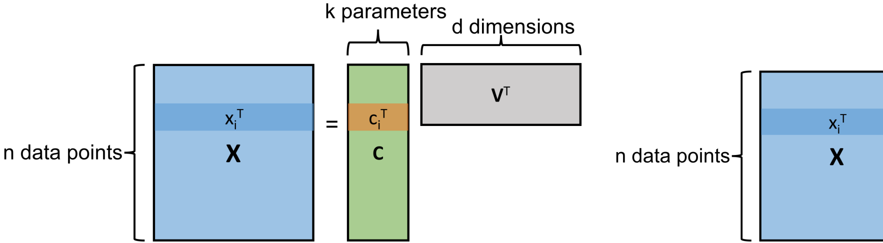


- $\mathbf{X}$ can be represented by $(n + d) \cdot k$ parameters vs. $n \cdot d$.
- The rows of $\mathbf{X}$ are spanned by k vectors: the columns of $\mathbf{V} \implies$ the columns of $\mathbf{X}$ are spanned by k vectors: the columns of $\mathbf{C}$.

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

LOW-RANK FACTORIZATION

Claim: If $\vec{x}_1, \ldots, \vec{x}_n$ lie in a k-dimensional subspace with orthonormal basis $\mathbf{V} \in \mathbb{R}^{d \times k}$, the data matrix can be written as $\mathbf{X} = \mathbf{C}\mathbf{V}^T$.



Exercise: What is this coefficient matrix $\mathbf{C}$? **Hint:** Use that $\mathbf{V}^T\mathbf{V} = \mathbf{I}$.

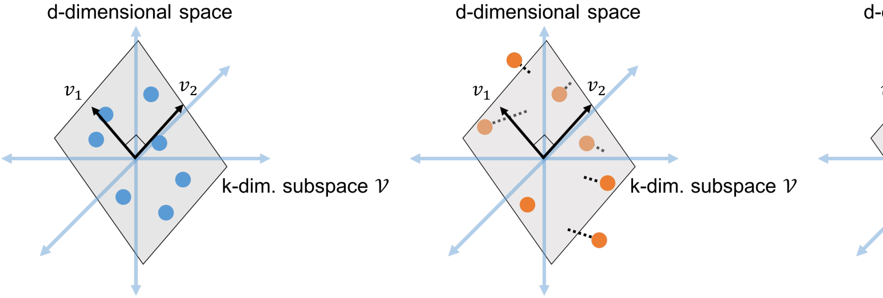
- $\mathbf{X} = \mathbf{C}\mathbf{V}^T \implies \mathbf{X}\mathbf{V} = \mathbf{C}\mathbf{V}^T\mathbf{V} \implies \mathbf{X}\mathbf{V} = \mathbf{C}$

$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: 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 $\vec{v}_1, \ldots, \vec{v}_k$.

Claim: If $\vec{x}_1, \ldots, \vec{x}_n$ lie in a k-dimensional subspace $\mathcal{V}$ with orthonormal basis $\mathbf{V} \in \mathbb{R}^{d \times k}$, the data matrix can be written as

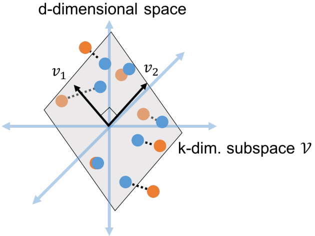
$$\mathbf{X} = \mathbf{C}\mathbf{V}^T\mathbf{X}\mathbf{V}\mathbf{V}^T.$$

- $\mathbf{V}\mathbf{V}^T$ is a projection matrix, which projects the rows of $\mathbf{X}$ (the data points $\vec{x}_1, \ldots, \vec{x}_n$ onto the subspace $\mathcal{V}$.



Claim: If $\vec{x}_1, \ldots, \vec{x}_n$ lie 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:

$$\mathbf{X} \approx \mathbf{X}\mathbf{V}\mathbf{V}^T$$



Note: $\mathbf{X}\mathbf{V}\mathbf{V}^T$ has rank k. It is a low-rank approximation of $\mathbf{X}$.

$$\mathbf{X}\mathbf{V}\mathbf{V}^{\mathrm{T}} = \underset{\mathbf{B} \text{ with rows in } \mathcal{V}}{\arg\min} \|\mathbf{X} - \mathbf{B}\|_F^2 = \sum_{i,j} (\mathbf{X}_{i,j} - \mathbf{B}_{i,j})^2.$$

$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: 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 $\vec{v}_1, \ldots, \vec{v}_k$.

So Far: If $\vec{x}_1, \ldots, \vec{x}_n$ lie 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:

$$\mathbf{X} \approx \mathbf{X}\mathbf{V}\mathbf{V}^T.$$

This is the closest approximation to $\mathbf{X}$ with rows in $\mathcal{V}$ (i.e., in the column span of $\mathbf{V}$).

- Letting $(\mathbf{X}\mathbf{V}\mathbf{V}^T)_i, (\mathbf{X}\mathbf{V}\mathbf{V}^T)_j$ be the i^{th} and j^{th} projected data points,

$$\|(\mathbf{X}\mathbf{V}\mathbf{V}^T)_i - (\mathbf{X}\mathbf{V}\mathbf{V}^T)_j\|_2 = \|[(\mathbf{X}\mathbf{V})_i - (\mathbf{X}\mathbf{V})_j]\mathbf{V}^T\|_2 = \|[(\mathbf{X}\mathbf{V})_i - (\mathbf{X}\mathbf{V})_j]\|_2.$$

- Can use $\mathbf{X}\mathbf{V} \in \mathbb{R}^{n \times k}$ as a compressed approximate data set.

Key question is how to find the subspace $\mathcal{V}$ and correspondingly $\mathbf{V}$.

$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: 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 $\vec{v}_1, \ldots, \vec{v}_k$.

Quick Exercise: Show that $\mathbf{V}\mathbf{V}^T$ is idempotent. I.e., $(\mathbf{V}\mathbf{V}^T)(\mathbf{V}\mathbf{V}^T)\vec{y} = (\mathbf{V}\mathbf{V}^T)\vec{y}$ for any $\vec{y} \in \mathbb{R}^d$.

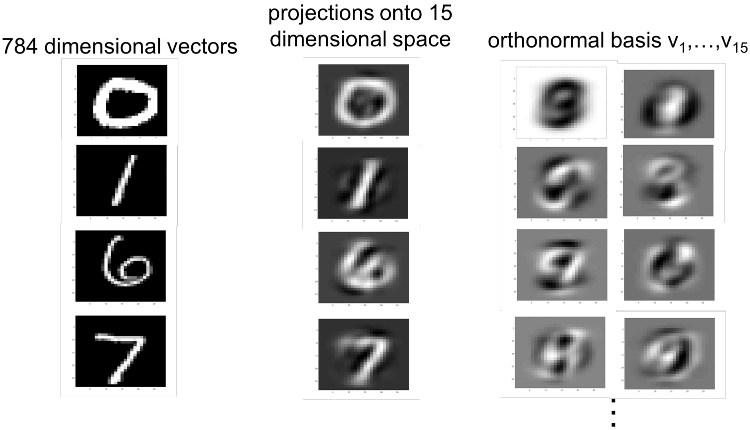
Why does this make sense intuitively?

Less Quick Exercise: (Pythagorean Theorem) Show that:

$$\|\vec{y}\|_2^2 = \|(\mathbf{V}\mathbf{V}^T)\vec{y}\|_2^2 + \|\vec{y} - (\mathbf{V}\mathbf{V}^T)\vec{y}\|_2^2.$$

Question: Why might we expect $\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$ to lie close to a k-dimensional subspace?

- The rows of X can be approximately reconstructed from a basis of k vectors.



DUAL VIEW OF LOW-RANK APPROXIMATION

Question: Why might we expect $\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$ to lie close to a k-dimensional subspace?

- Equivalently, the columns of $\mathbf{X}$ are approx. spanned by k vectors.

Linearly Dependent Variables:

	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

	bedrooms
home 1	2
home 2	4
⋮	⋮
home n	5