COMPSCI 514: Algorithms for Data Science

Cameron Musco

University of Massachusetts Amherst. Fall 2026.

Lecture 16

In-Class Quiz 6

Question 1: Let $\vec{g} \in \mathbb{R}^d$ have i.i.d. $\mathcal{N}(0,1)$ entries. Let $\vec{v} \in \mathbb{R}^d$ be a fixed vector. What is $\mathsf{Var}[\langle \vec{g}, \vec{v} \rangle]$?

Question 2: Let $\vec{g} \in \mathbb{R}^d$ have i.i.d. $\mathcal{N}(0,1)$ entries. Let $\vec{v} \in \mathbb{R}^d$ be a fixed vector. What is $\mathsf{Var}[\mathrm{sign}(\langle \vec{g}, \vec{v} \rangle)]$?

Logistics

- Problem Set 3 due this Friday at 11:59pm.
- Midterm 2 is next Wednesday **in class**.
- Study guide and practice exams have been posted.
- No class on Wednesday due to Veteran's day.

Summary

Last Class:

- Analysis of quantized JL transform for vector quantization.

This Class:

- Start on data dependent dimensionality reduction.
- Approximation via projection onto a low-dimensional subspace. Connection to low-rank matrix factorization.
- Finding an optimal subspace via eigendecomposition.

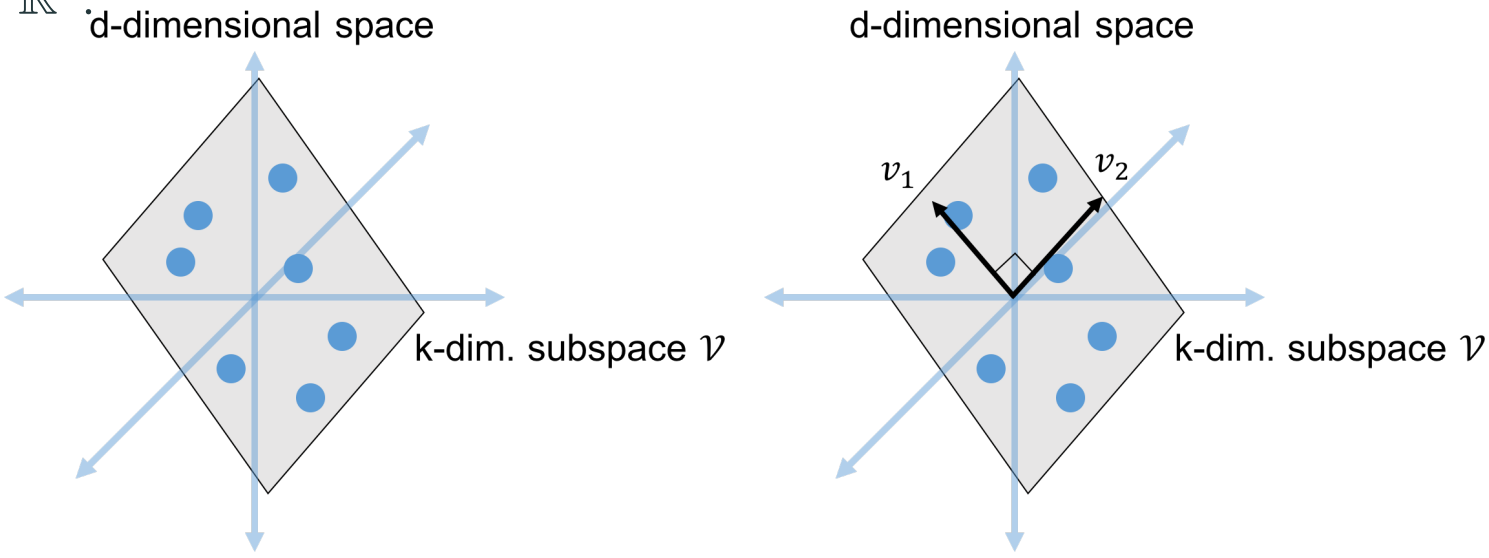
Data Dependent Dimensionality Reduction

Data Dependent Dimensionality Reduction

- The JL Lemma is very general – it applies to any input data set.
- The compression Π is linear and data-oblivious. This is very useful for computational efficiency, and leads to applications in streaming and distributed settings.
- However, often we can exploit properties about the underlying input data to get more accurate compressions than what JL gives.
- Leads to a variety of data dependent dimensionality reduction methods.

Embedding with Assumptions

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.

Very Simple Example

Another Example

Dot Product Transformation

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 $\vec{x}_i, \vec{x}_j \in \mathcal{V}$:

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

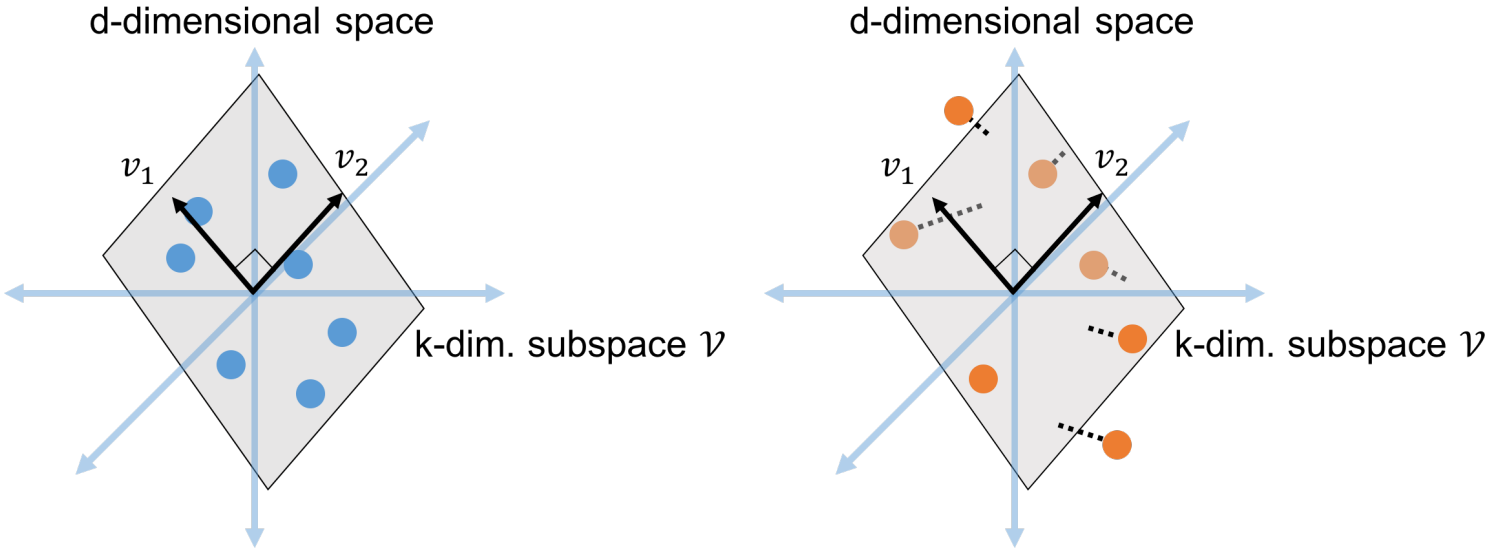
Dot Product Transformation

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 $\vec{x}_i, \vec{x}_j \in \mathcal{V}$:

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

Embedding with Assumptions

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?

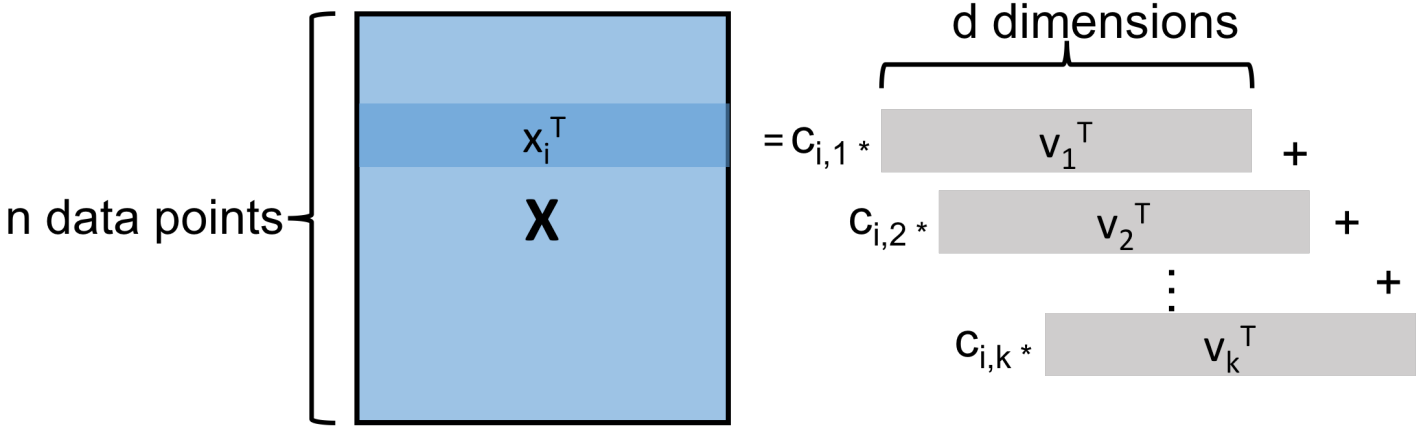
Low-Rank Factorization

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 $\vec{x}_i$ as:

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

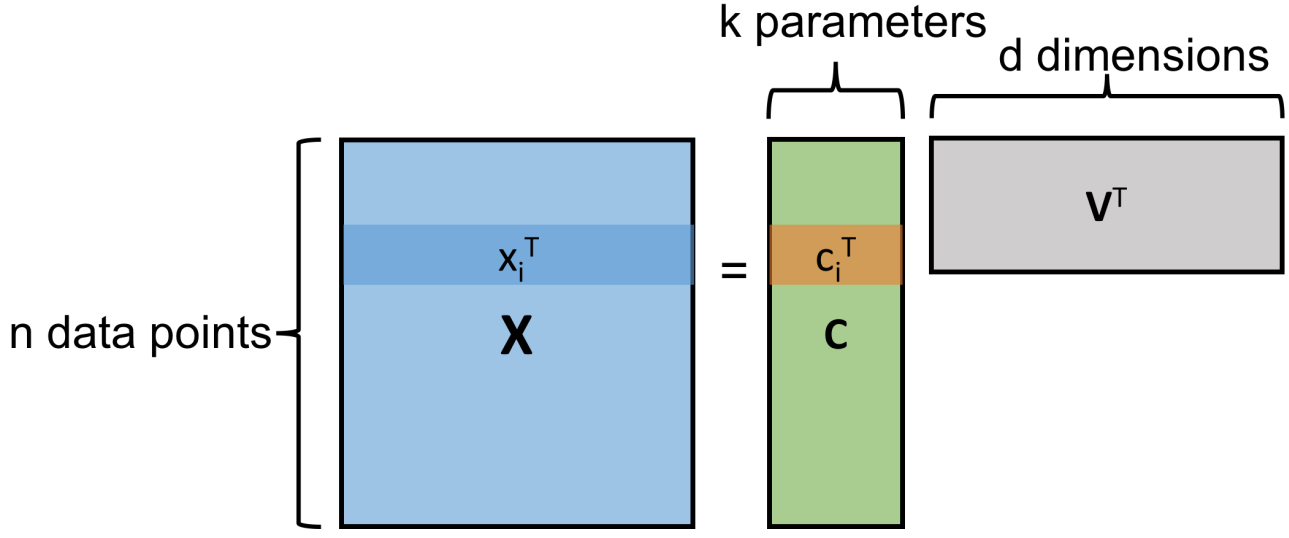
- So $\vec{v}_1, \ldots, \vec{v}_k$ span the rows of $\mathbf{X}$ and thus $\mathbf{rank}(\mathbf{X}) \leq k$.



$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: 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 $\vec{x}_i = \mathbf{V}\vec{c}_i = c_{i,1} \cdot \vec{v}_1 + \ldots + c_{i,k} \cdot \vec{v}_k$.

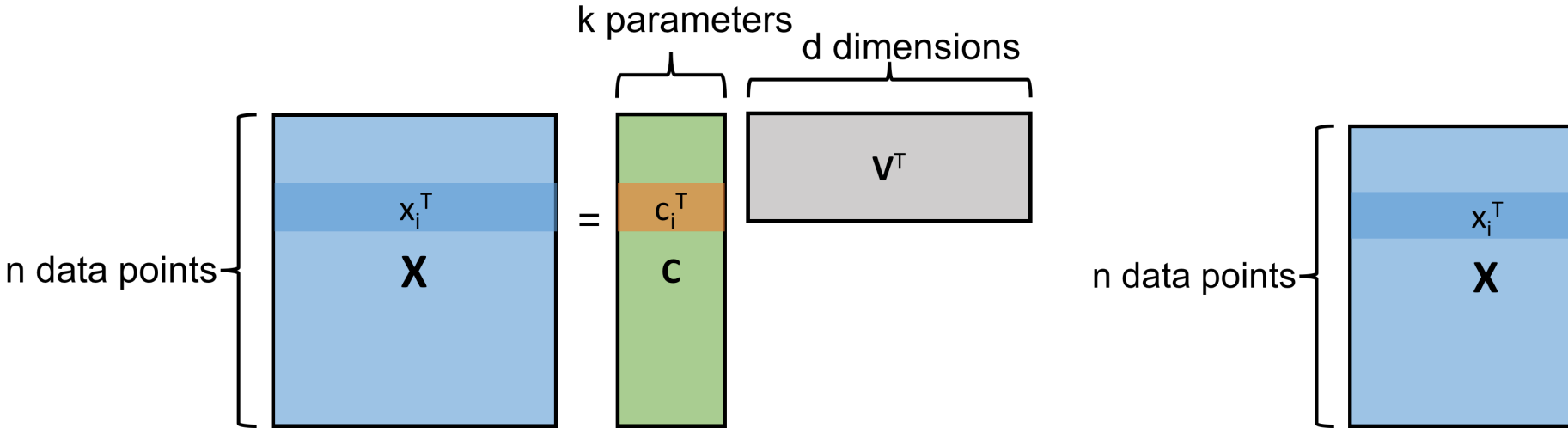


- $\mathbf{X}$ can be represented by $(n + d) \cdot k$ parameters vs. $n \cdot d$.
- The rows of $\mathbf{X}$ are spanned by k vectors: the columns of $\mathbf{V} \implies$ the columns of $\mathbf{X}$ are spanned by k vectors: the columns of $\mathbf{C}$.

$\vec{x}_1, \ldots, \vec{x}_n$: 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 $\mathbf{V}^T\mathbf{V} = \mathbf{I}$.

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

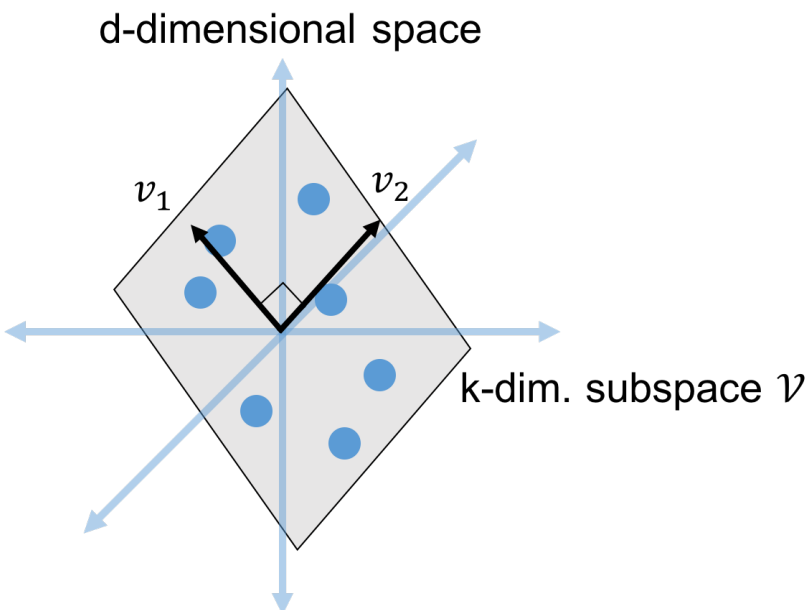
$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for subspace $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

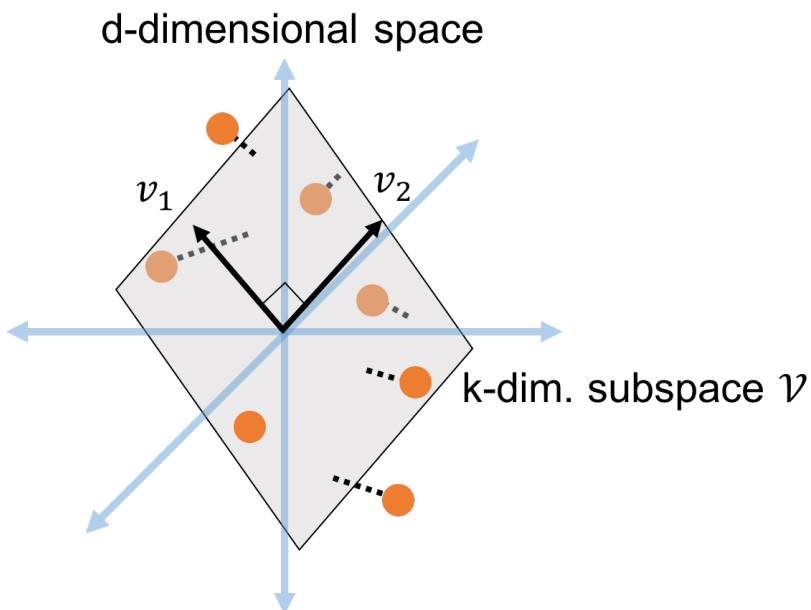
Projection View

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 vectors onto the subspace $\mathcal{V}$.

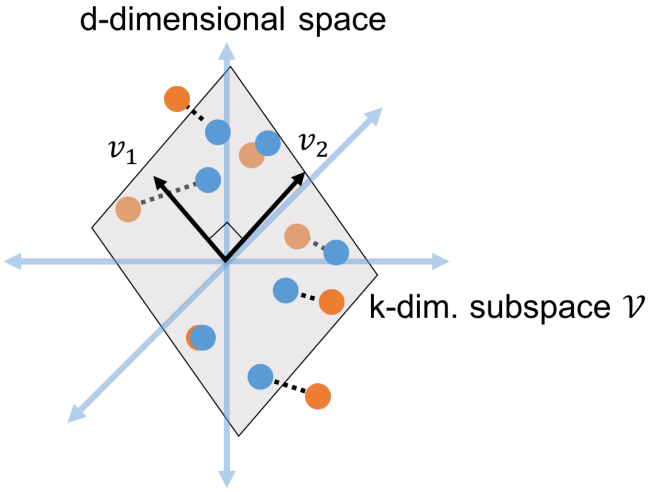




Low-Rank Approximation

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 $\mathbf{X}$.

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

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

Low-Rank Approximation

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 $\vec{v}_1, \ldots, \vec{v}_k$.

Properties of Projection Matrices

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:

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

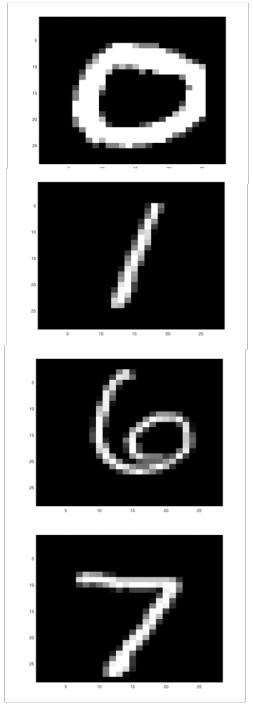
A Step Back

Why Low-Rank Approximation?

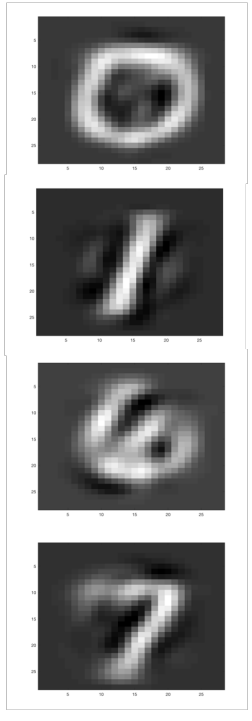
Question: Why might we expect $\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$ to lie close to a k-dimensional subspace?

- The rows of $\mathbf{X}$ can be approximately reconstructed from a basis of k vectors.

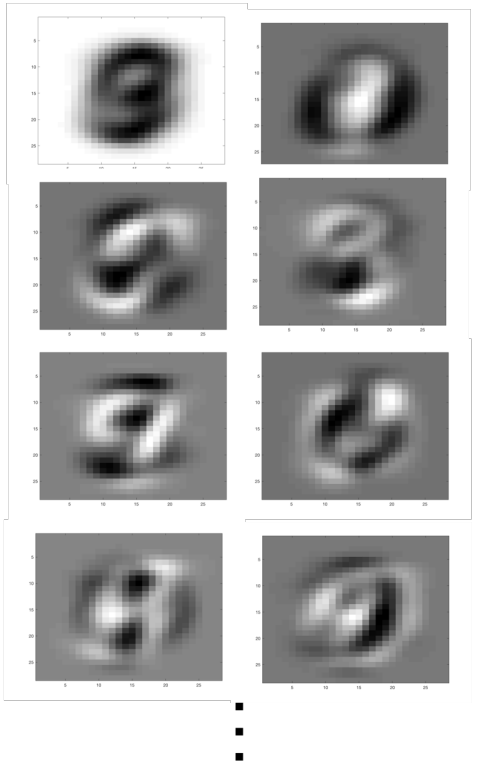
784 dimensional vectors



projections onto 15 dimensional space



orthonormal basis $v_1, \ldots, v_{15}$



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