COMPSCI 514: ALGORITHMS FOR DATA SCIENCE

Cameron Musco

University of Massachusetts Amherst. Fall 2019.

Lecture 11

- Problem Set 2 is due this Friday 10/11. Will allow submissions until Sunday 10/13 at midnight with no penalty.
- Midterm next Thursday 10/17.

Problem Set 2:

- Mean was a $32.74/40 = 81\%$.
- Mostly seem to have mastered Markov's, Chebyshev, etc.
- Some difficulties with exponential tail bounds (Chernoff and Bernstein). Will give some review exercises before midterm.

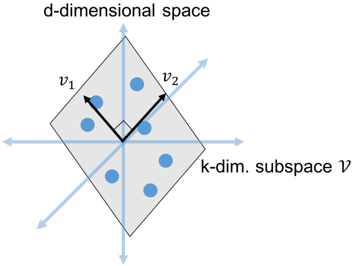
Last Two Classes: Randomized Dimensionality Reduction

- 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)

Next Two Classes: Low-rank approximation, the SVD, and principal component analysis.

- Compression is still linear – by applying a matrix.
- Chose this matrix carefully, taking into account structure of the dataset.
- Can give better compression than random projection.

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



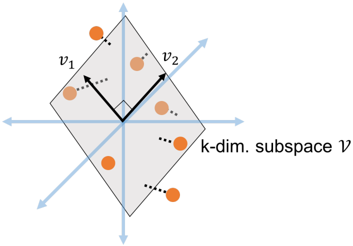
Recall: Let $\vec{v}_1, \ldots, \vec{v}_k$ be an orthonormal basis for $\mathcal{V}$ and $\mathbf{V} \in \mathbb{R}^{d \times k}$ be the matrix with these vectors as its columns. For all $\vec{x}_i, \vec{x}_j$:

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

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

EMBEDDING WITH ASSUMPTIONS

Main Focus of Today: 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).

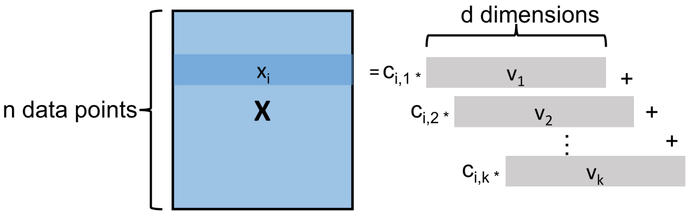
- How do we find $\mathcal{V}$ and $\mathbf{V}$?
- How good is the embedding?

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:

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

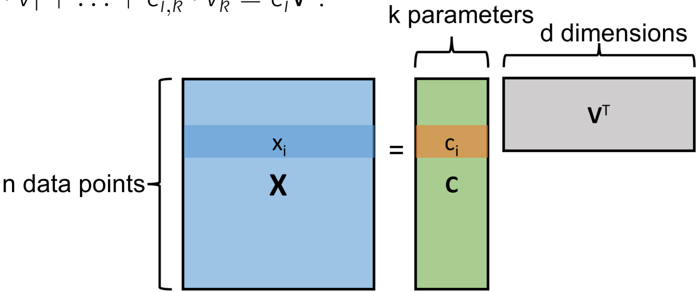
- So $\vec{v}_1, \ldots, \vec{v}_k$ span the rows of $\mathbf{X}$ and thus $\text{rank}(\mathbf{X}) \leq k$.



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

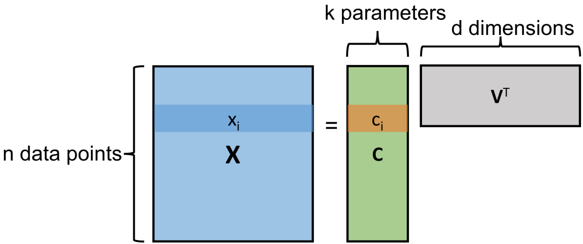
- Every data point $\vec{x}_i$ (row of $\mathbf{X}$) can be written as $c_{i,1} \cdot \vec{v}_1 + \ldots + c_{i,k} \cdot \vec{v}_k = \vec{c}_i \mathbf{V}^T$.



- $\mathbf{X}$ can be represented by $(n + d) \cdot k$ parameters vs. $n \cdot d$.
- 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 $\vec{v}_1, \ldots, \vec{v}_k$.

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

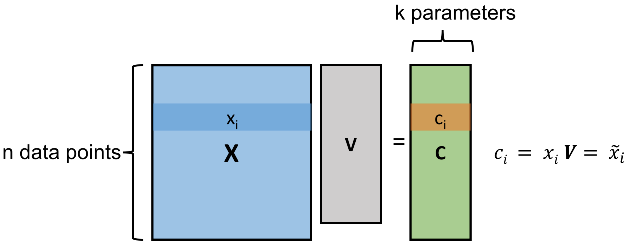


What is this coefficient matrix $\mathbf{C}$?

- $\mathbf{X} = \mathbf{C}\mathbf{V}^T \implies \mathbf{X}\mathbf{V} = \mathbf{C}\mathbf{V}^T\mathbf{V}$
- $\mathbf{V}^T\mathbf{V} = \mathbf{I}$, the identity (since $\mathbf{V}$ is orthonormal) $\implies \mathbf{X}\mathbf{V} = \mathbf{C}$.

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

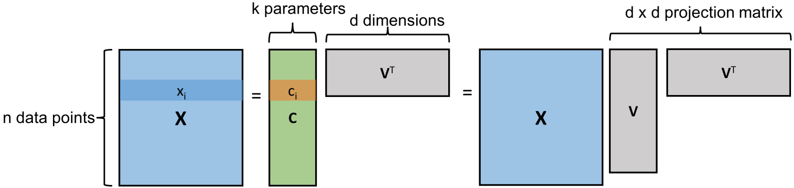


What is this coefficient matrix $\mathbf{C}$?

- $\mathbf{X} = \mathbf{C}\mathbf{V}^T \implies \mathbf{X}\mathbf{V} = \mathbf{C}\mathbf{V}^T\mathbf{V}$
- $\mathbf{V}^T\mathbf{V} = \mathbf{I}$, the identity (since $\mathbf{V}$ is orthonormal) $\implies \mathbf{X}\mathbf{V} = \mathbf{C}$.

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



What is this coefficient matrix $\mathbf{C}$?

- $\mathbf{X} = \mathbf{C}\mathbf{V}^T \implies \mathbf{X}\mathbf{V} = \mathbf{C}\mathbf{V}^T\mathbf{V}$
- $\mathbf{V}^T\mathbf{V} = \mathbf{I}$, the identity (since $\mathbf{V}$ is orthonormal) $\implies \mathbf{X}\mathbf{V} = \mathbf{C}$.

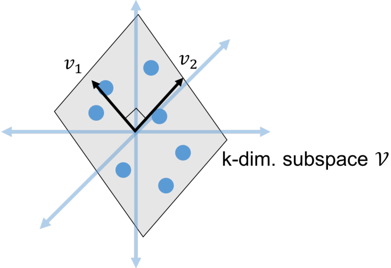
$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for subspace $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

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

$$\mathbf{X} = \mathbf{X}(\mathbf{V}\mathbf{V}^T).$$

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



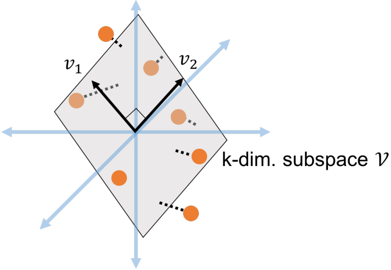
$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for subspace $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

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

$$\mathbf{X} = \mathbf{X}(\mathbf{V}\mathbf{V}^T).$$

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



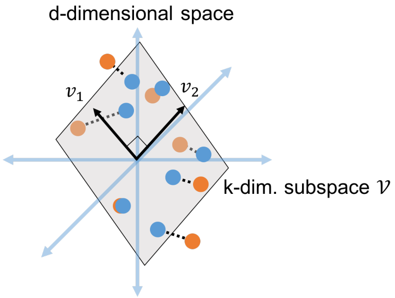
$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for subspace $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

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

$$\mathbf{X} = \mathbf{X}(\mathbf{V}\mathbf{V}^T).$$

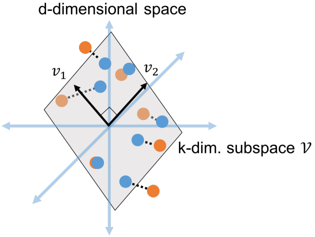
- $\mathbf{V}\mathbf{V}^T$ is a projection matrix, which projects the rows of $\mathbf{X}$ (the data points $\vec{x}_1, \ldots, \vec{x}_n$ onto the subspace $\mathcal{V}$.



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

$$\mathbf{X} \approx \mathbf{X}(\mathbf{V}\mathbf{V}^T) = \mathbf{X}\mathbf{P}_{\mathcal{V}}$$



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

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 $\mathbf{V}$).

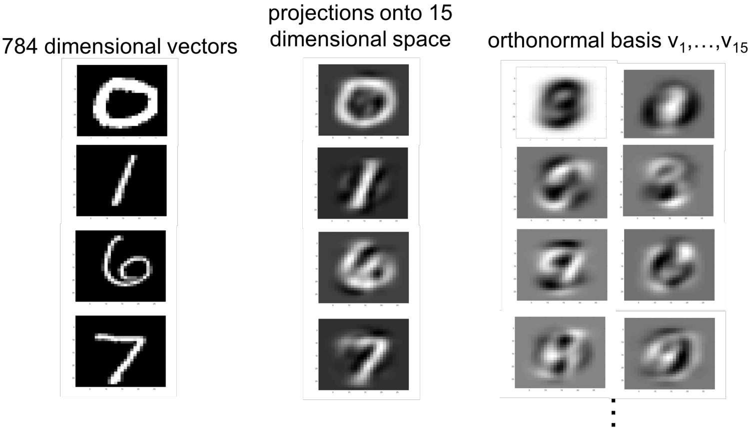
- Letting $(\mathbf{X}\mathbf{V}\mathbf{V}^T)_i, (\mathbf{X}\mathbf{V}\mathbf{V}^T)_j$ be the i^{th} and j^{th} projected data points,
$$\|(\mathbf{X}\mathbf{V}\mathbf{V}^T)_i - (\mathbf{X}\mathbf{V}\mathbf{V}^T)_j\|_2 = \|[(\mathbf{X}\mathbf{V})_i - (\mathbf{X}\mathbf{V})_j]\mathbf{V}^T\|_2 = \|[(\mathbf{X}\mathbf{V})_i - (\mathbf{X}\mathbf{V})_j]\|_2.$$
- 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 $\vec{v}_1, \ldots, \vec{v}_k$.

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

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



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

- Equivalently, the columns of **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

Question: Why might we expect $\vec{x}_1, \ldots, \vec{x}_n$ 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

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

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

Linearly Dependent Variables:

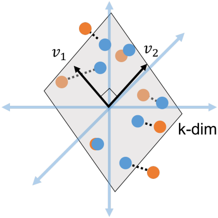
10000* bathrooms+ 10* (sq. ft.) ≈ list price

	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 $\vec{x}_1, \ldots, \vec{x}_n$ 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 $\mathcal{V}$.

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

$$\underset{\text{orthonormal } \mathbf{V} \in \mathbb{R}^{d \times k}}{\arg\min} \|\mathbf{X} - \mathbf{X}\mathbf{V}\mathbf{V}^T\|_F^2 = \sum_{i,j} (\mathbf{X}_{i,j} - (\mathbf{X}\mathbf{V}\mathbf{V}^T)_{i,j})^2 = \sum_{i=1}^{n} \|\vec{x}_i - \mathbf{V}\mathbf{V}^T\vec{x}_i\|_2^2$$



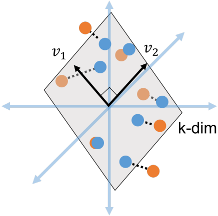
$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for subspace $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

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 $\mathcal{V}$.

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

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

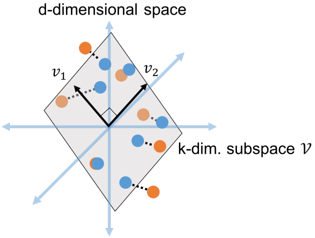


$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for subspace $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

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 $\mathcal{V}$.

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

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

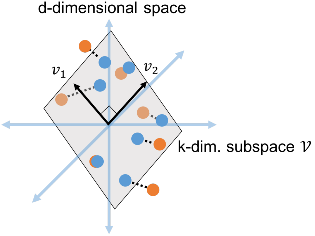


$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for subspace $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

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 $\mathcal{V}$.

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

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



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