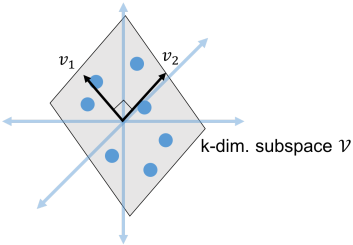
# COMPSCI 514: ALGORITHMS FOR DATA SCIENCE

Andrew McGregor

Lecture 15

# LAST CLASS: EMBEDDING WITH ASSUMPTIONS

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




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.

$$\|\mathbf{V}^T \vec{x}_i - \mathbf{V}^T \vec{x}_j\|_2^2 = \|\vec{x}_i - \vec{x}_j\|_2^2.$$

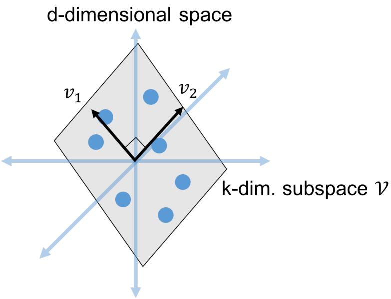
Letting $\tilde{x}_i = \mathbf{V}^T \vec{x}_i$, we have a perfect embedding from $\mathcal{V}$ into $\mathbb{R}^k$.

# LAST CLASS: 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

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

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



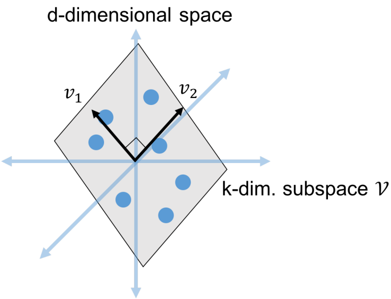

$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for subspace $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

# LAST CLASS: 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

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

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



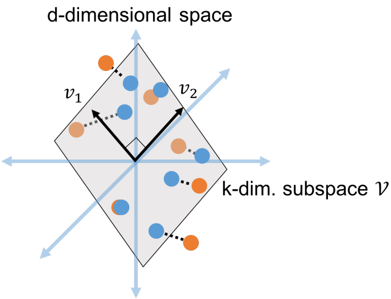

$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for subspace $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

# LAST CLASS: 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

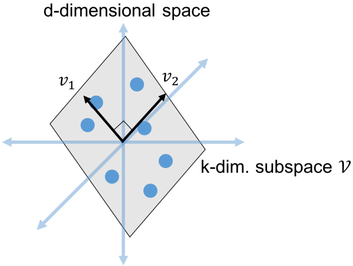
$$\mathbf{X} = \mathbf{X}\mathbf{V}\mathbf{V}^T = \mathbf{C}\mathbf{V}^T$$

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



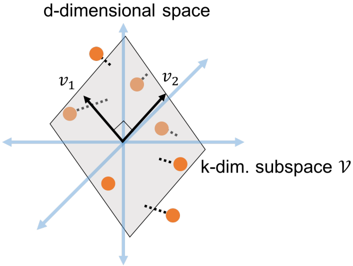

$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for subspace $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

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



# EMBEDDING WITH ASSUMPTIONS

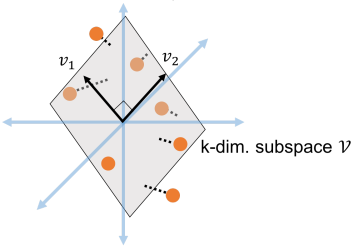
Assume that data points $\vec{x}_1, \ldots, \vec{x}_n$ lie close to any $k$-dimensional subspace $\mathcal{V}$ of $\mathbb{R}^d$.




Will show above in homework. Today's focus: How do we find $\mathcal{V}$ and $\mathbf{V}$?

# EMBEDDING WITH ASSUMPTIONS

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 $x_i \in \mathbb{R}^d$ and $\mathbf{XVV}^T$ is still a good approximation for $\mathbf{X}$:

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

Will show above in homework. Today's focus: How do we find $\mathcal{V}$ and $\mathbf{V}$?

# 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?

# A STEP BACK: WHY LOW-RANK APPROXIMATION?

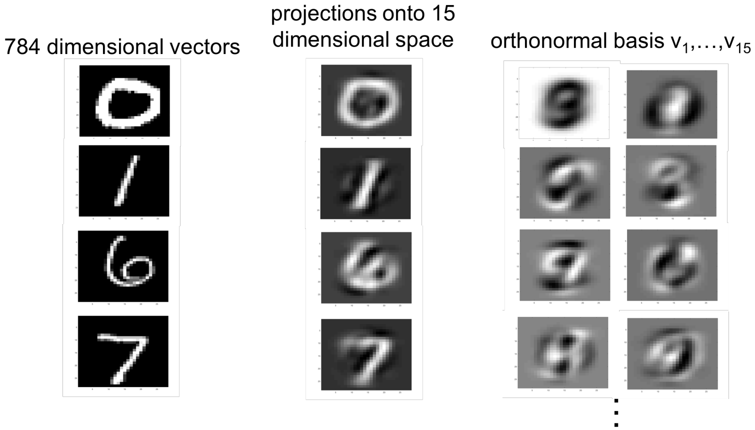
**Question:** Why might we expect $\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$ to lie close to a $k$-dimensional subspace?

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

# 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.



# 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?

# 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.

**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 |

# 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 |

# 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 |

**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:**

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 |

**Quick Exercise 1 :** 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 $\vec{y} \in \mathbb{R}^d$.

**Quick Exercise 1 :** 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 $\vec{y} \in \mathbb{R}^d$.

**Quick Exercise 2:** The projection is orthogonal to its complement: For any $\vec{y} \in \mathbb{R}^d$, $\langle \mathbf{V}\mathbf{V}^T\vec{y}, (\mathbf{I} - \mathbf{V}\mathbf{V}^T)\vec{y} \rangle = 0$

**Quick Exercise 1 :** 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 $\vec{y} \in \mathbb{R}^d$.

**Quick Exercise 2:** The projection is orthogonal to its complement: For any $\vec{y} \in \mathbb{R}^d$, $\langle \mathbf{V}\mathbf{V}^T\vec{y}, (\mathbf{I} - \mathbf{V}\mathbf{V}^T)\vec{y} \rangle = 0$

**Implies the Pythagorean Theorem:** Show that for any $\vec{y} \in \mathbb{R}^d$,

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

# 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 $\mathbf{X}\mathbf{V}\mathbf{V}^T$ and $\mathbf{X}\mathbf{V}$ gives optimal embedding of $\mathbf{X}$ in $\mathcal{V}$. How do we find $\mathcal{V}$ (equivalently $\mathbf{V}$)?

# 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 $\mathbf{XVV}^T$ and $\mathbf{XV}$ gives optimal embedding of $\mathbf{X}$ in $\mathcal{V}$. How do we find $\mathcal{V}$ (equivalently $\mathbf{V}$)?

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

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 $\mathbf{X}\mathbf{V}\mathbf{V}^T$ and $\mathbf{X}\mathbf{V}$ gives optimal embedding of $\mathbf{X}$ in $\mathcal{V}$. How do we find $\mathcal{V}$ (equivalently $\mathbf{V}$)?

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

So want to maximize $\sum_i \|\mathbf{V}\mathbf{V}^T \vec{x}_i\|_2^2 = \sum_i \vec{x}_i^T \mathbf{V}\mathbf{V}^T\mathbf{V}\mathbf{V}^T \vec{x}_i = \sum_i \|\mathbf{V}^T \vec{x}_i\|_2^2$

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

# SOLUTION VIA EIGENDECOMPOSITION

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

$$\underset{\text{orthonormal } \mathbf{V} \in \mathbb{R}^{d \times k}}{\arg\max} \sum_{i=1}^{n} \|\mathbf{V}^T \vec{x}_i\|_2^2 = \sum_{j=1}^{k} \sum_{i=1}^{n} \langle \vec{v}_j, \vec{x}_i \rangle^2$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

# SOLUTION VIA EIGENDECOMPOSITION

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

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

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

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

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

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

# SOLUTION VIA EIGENDECOMPOSITION

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

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

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

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

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

$$\cdots$$

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

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

# SOLUTION VIA EIGENDECOMPOSITION

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

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

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

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

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

$$\cdots$$

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

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

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