# COMPSCI 514: ALGORITHMS FOR DATA SCIENCE

Andrew McGregor

Lecture 14

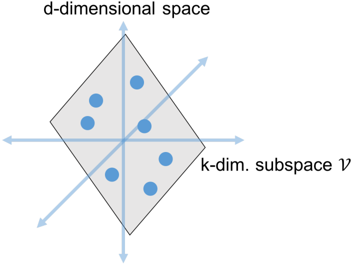
**Next Few Classes: Low-rank approximation, the SVD, and principal component analysis (PCA).**

- Reduce $d$-dimensional data points to a smaller dimension $m$.
- Like JL, compression is linear, i.e., by applying a matrix.
- Chose matrix taking into account structure of dataset.
- Can give better compression than random projection.

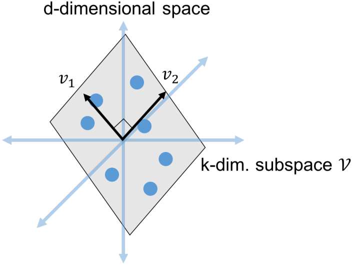
**Next Few Classes: Low-rank approximation, the SVD, and principal component analysis (PCA).**

- Reduce $d$-dimensional data points to a smaller dimension $m$.
- Like JL, compression is linear, i.e., by applying a matrix.
- Chose matrix taking into account structure of dataset.
- Can give better compression than random projection.

Will be using a fair amount of linear algebra: orthogonal basis, column/row span, eigenvectors, etc, Assume that data points $\vec{x}_1, \ldots, \vec{x}_n$ lie in any $k$-dimensional subspace $\mathcal{V}$ of $\mathbb{R}^d$.



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

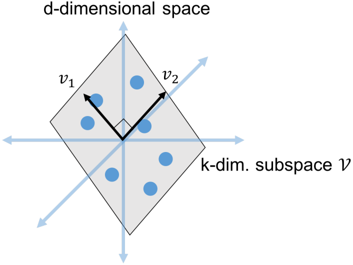



**Claim:** Let $\vec{v}_1, \ldots, \vec{v}_k$ be an orthonormal basis for $\mathcal{V}$ and $\mathbf{V} \in \mathbb{R}^{d \times k}$ be the matrix with these vectors as its columns. For all $\vec{x}_i, \vec{x}_j$:

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

# 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.

**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 $\vec{y} \in \mathcal{V}$:

$$\|\mathbf{V}^T y\|_2 = \|y\|_2.$$

Proof:

# 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 $\vec{y} \in \mathcal{V}$:

$$\|\mathbf{V}^T y\|_2 = \|y\|_2.$$

Proof:

- If $\vec{y} = \sum_i c_i \vec{v}_i$ then $\vec{y} = \mathbf{V}\vec{c}$ where $\vec{c}^T = (c_1, \ldots, c_k)$

**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 $\vec{y} \in \mathcal{V}$:

$$\|\mathbf{V}^T y\|_2 = \|y\|_2.$$

Proof:

- If $\vec{y} = \sum_i c_i \vec{v}_i$ then $\vec{y} = \mathbf{V}\vec{c}$ where $\vec{c}^T = (c_1, \ldots, c_k)$
- $\|\vec{y}\|_2^2 = \vec{y}^T\vec{y}$

**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 $\vec{y} \in \mathcal{V}$:

$$\|\mathbf{V}^T y\|_2 = \|y\|_2.$$

Proof:

- If $\vec{y} = \sum_i c_i \vec{v}_i$ then $\vec{y} = \mathbf{V}\vec{c}$ where $\vec{c}^T = (c_1, \ldots, c_k)$
- $\|\vec{y}\|_2^2 = \vec{y}^T \vec{y} = (\mathbf{V}\vec{c})^T (\mathbf{V}\vec{c})$

**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 $\vec{y} \in \mathcal{V}$:

$$\|\mathbf{V}^T y\|_2 = \|y\|_2.$$

Proof:

- If $\vec{y} = \sum_i c_i \vec{v}_i$ then $\vec{y} = \mathbf{V}\vec{c}$ where $\vec{c}^T = (c_1, \ldots, c_k)$
- $\|\vec{y}\|_2^2 = \vec{y}^T \vec{y} = (\mathbf{V}\vec{c})^T(\mathbf{V}\vec{c}) = \vec{c}^T \mathbf{V}^T \mathbf{V}\vec{c}$

**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 $\vec{y} \in \mathcal{V}$:

$$\|\mathbf{V}^T y\|_2 = \|y\|_2.$$

Proof:

- If $\vec{y} = \sum_i c_i \vec{v}_i$ then $\vec{y} = \mathbf{V}\vec{c}$ where $\vec{c}^T = (c_1, \ldots, c_k)$
- $\|\vec{y}\|_2^2 = \vec{y}^T \vec{y} = (\mathbf{V}\vec{c})^T (\mathbf{V}\vec{c}) = \vec{c}^T \mathbf{V}^T \mathbf{V} \vec{c}$
- $\|\mathbf{V}^T \vec{y}\|_2^2 = (\mathbf{V}^T \vec{y})^T (\mathbf{V}^T \vec{y}) = \vec{y}^T \mathbf{V} \mathbf{V}^T \vec{y}$

**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 $\vec{y} \in \mathcal{V}$:

$$\|\mathbf{V}^T y\|_2 = \|y\|_2.$$

Proof:

- If $\vec{y} = \sum_i c_i \vec{v}_i$ then $\vec{y} = \mathbf{V}\vec{c}$ where $\vec{c}^T = (c_1, \ldots, c_k)$
- $\|\vec{y}\|_2^2 = \vec{y}^T\vec{y} = (\mathbf{V}\vec{c})^T(\mathbf{V}\vec{c}) = \vec{c}^T\mathbf{V}^T\mathbf{V}\vec{c}$
- $\|\mathbf{V}^T\vec{y}\|_2^2 = (\mathbf{V}^T\vec{y})^T(\mathbf{V}^T\vec{y}) = \vec{y}^T\mathbf{V}\mathbf{V}^T\vec{y} = \vec{c}^T\mathbf{V}^T\mathbf{V}\mathbf{V}^T\mathbf{V}\vec{c}$

# 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 $\vec{y} \in \mathcal{V}$:

$$\|\mathbf{V}^T y\|_2 = \|y\|_2.$$

Proof:

- If $\vec{y} = \sum_i c_i \vec{v}_i$ then $\vec{y} = \mathbf{V}\vec{c}$ where $\vec{c}^T = (c_1, \ldots, c_k)$
- $\|\vec{y}\|_2^2 = \vec{y}^T \vec{y} = (\mathbf{V}\vec{c})^T(\mathbf{V}\vec{c}) = \vec{c}^T\mathbf{V}^T\mathbf{V}\vec{c}$
- $\|\mathbf{V}^T\vec{y}\|_2^2 = (\mathbf{V}^T\vec{y})^T(\mathbf{V}^T\vec{y}) = \vec{y}^T\mathbf{V}\mathbf{V}^T\vec{y} = \vec{c}^T\mathbf{V}^T\mathbf{V}\mathbf{V}^T\mathbf{V}\vec{c}$
- But $\mathbf{V}^T\mathbf{V} = I$

**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 $\vec{y} \in \mathcal{V}$:

$$\|\mathbf{V}^T y\|_2 = \|y\|_2.$$

Proof:

- If $\vec{y} = \sum_i c_i \vec{v}_i$ then $\vec{y} = \mathbf{V}\vec{c}$ where $\vec{c}^T = (c_1, \ldots, c_k)$
- $\|\vec{y}\|_2^2 = \vec{y}^T \vec{y} = (\mathbf{V}\vec{c})^T (\mathbf{V}\vec{c}) = \vec{c}^T \mathbf{V}^T \mathbf{V} \vec{c}$
- $\|\mathbf{V}^T \vec{y}\|_2^2 = (\mathbf{V}^T \vec{y})^T (\mathbf{V}^T \vec{y}) = \vec{y}^T \mathbf{V}\mathbf{V}^T \vec{y} = \vec{c}^T \mathbf{V}^T \mathbf{V}\mathbf{V}^T \mathbf{V} \vec{c}$
- But $\mathbf{V}^T \mathbf{V} = I$ since

$$[\mathbf{V}^T \mathbf{V}]_{i,j} = \vec{v}_i^T \vec{v}_j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

**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 $\vec{y} \in \mathcal{V}$:

$$\|\mathbf{V}^T y\|_2 = \|y\|_2.$$

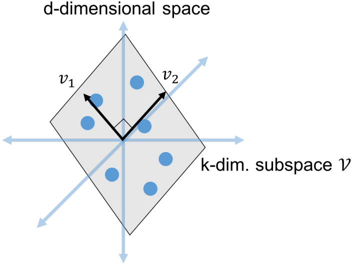
Proof:

- If $\vec{y} = \sum_i c_i \vec{v}_i$ then $\vec{y} = \mathbf{V}\vec{c}$ where $\vec{c}^T = (c_1, \ldots, c_k)$
- $\|\vec{y}\|_2^2 = \vec{y}^T\vec{y} = (\mathbf{V}\vec{c})^T(\mathbf{V}\vec{c}) = \vec{c}^T\mathbf{V}^T\mathbf{V}\vec{c}$
- $\|\mathbf{V}^T\vec{y}\|_2^2 = (\mathbf{V}^T\vec{y})^T(\mathbf{V}^T\vec{y}) = \vec{y}^T\mathbf{V}\mathbf{V}^T\vec{y} = \vec{c}^T\mathbf{V}^T\mathbf{V}\mathbf{V}^T\mathbf{V}\vec{c}$
- But $\mathbf{V}^T\mathbf{V} = I$ since

$$[\mathbf{V}^T\mathbf{V}]_{i,j} = \vec{v}_i^T\vec{v}_j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

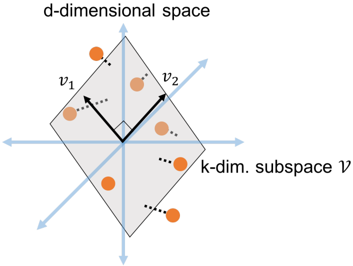
- So $\|\vec{y}\|_2^2 = \vec{c}^T\vec{c} = \|\mathbf{V}^T\vec{y}\|_2^2$.

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



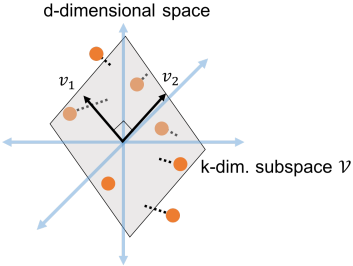
# EMBEDDING WITH ASSUMPTIONS

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



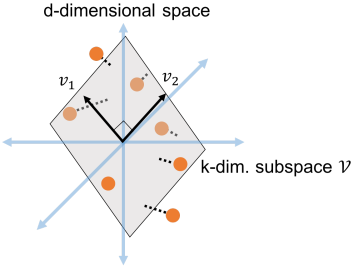
# EMBEDDING WITH ASSUMPTIONS

Now 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 $x_i \in \mathbb{R}^d$.

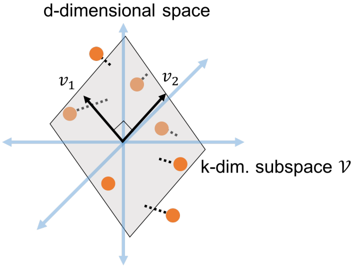
Now assume that data points $\vec{x}_1, \ldots, \vec{x}_n$ lie close to any $k$-dimensional subspace $\mathcal{V}$ of $\mathbb{R}^d$.




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).

# EMBEDDING WITH ASSUMPTIONS

Now 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 $\leq k$.

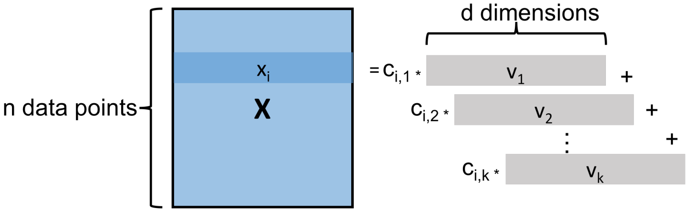
$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for subspace $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

# LOW-RANK 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 any $\vec{x}_i$ as:

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



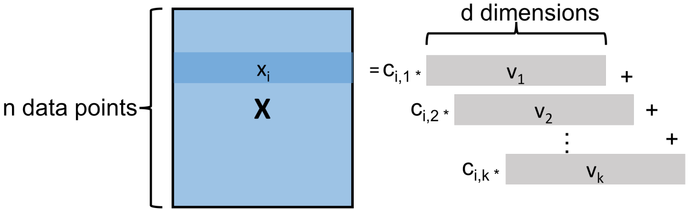

$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for subspace $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

**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 $\leq k$.

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

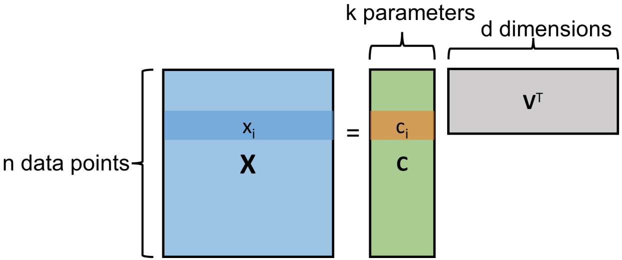
**Claim:** $\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$ lie in a $k$-dimensional subspace $\mathcal{V} \Leftrightarrow$ the data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ has rank $\leq k$.

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

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

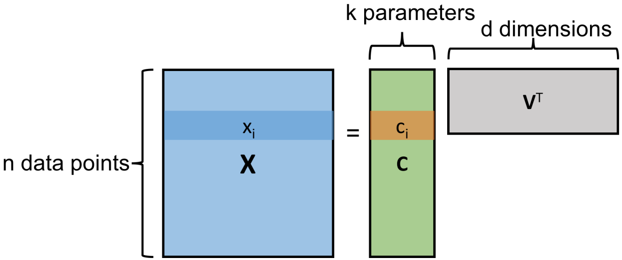
- Every data point $\vec{x}_i$ (row of $\mathbf{X}$) can be written as $\vec{x}_i = \mathbf{V}\vec{c}_i = c_{i,1} \cdot \vec{v}_1 + \ldots + c_{i,k} \cdot \vec{v}_k$.




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

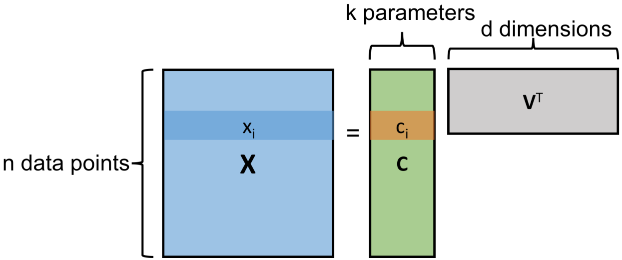



- $\mathbf{X}$ can be represented by $(n+d) \cdot k$ parameters vs. $n \cdot d$.

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

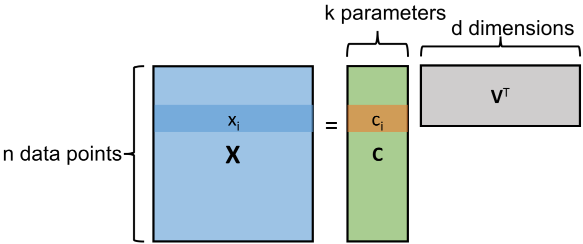
- Every data point $\vec{x}_i$ (row of $\mathbf{X}$) can be written as $\vec{x}_i = \mathbf{V}\vec{c}_i = c_{i,1} \cdot \vec{v}_1 + \ldots + c_{i,k} \cdot \vec{v}_k$.




- $\mathbf{X}$ 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 $\mathbf{C}$.

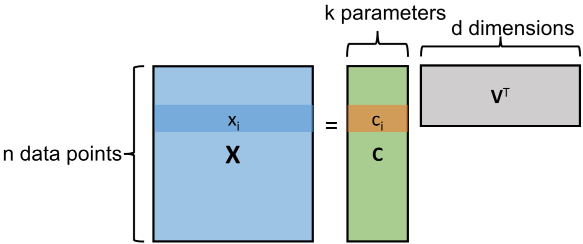
$\vec{x}_1, \ldots, \vec{x}_n$: data points (in $\mathbb{R}^d$), $\mathcal{V}$: $k$-dimensional subspace of $\mathbb{R}^d$, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

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




$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: 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 $\mathbf{X} = \mathbf{C}\mathbf{V}^T$.

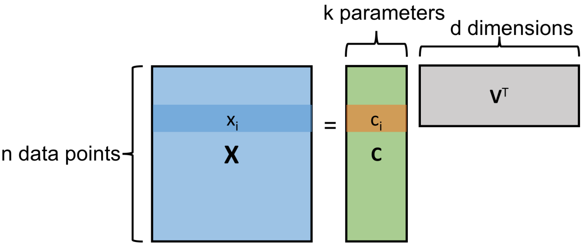



**Exercise:** What is this coefficient matrix **C**? **Hint:** Use that $\mathbf{V}^T\mathbf{V} = \mathbf{I}$.

$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: 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 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 $\mathbf{X} = \mathbf{C}\mathbf{V}^T$.

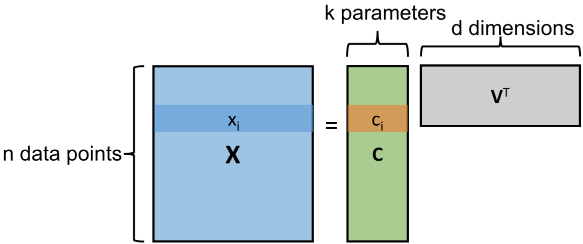



**Exercise:** What is this coefficient matrix $\mathbf{C}$? **Hint:** Use that $\mathbf{V}^T\mathbf{V} = \mathbf{I}$.

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

$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: 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 $\mathbf{X} = \mathbf{C}\mathbf{V}^T$.

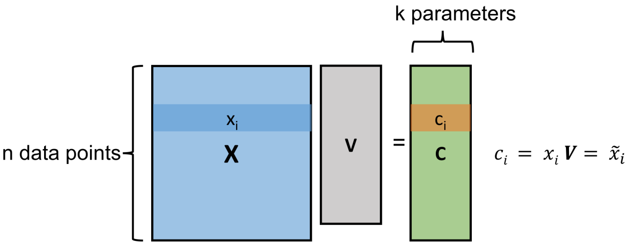



**Exercise:** What is this coefficient matrix $\mathbf{C}$? **Hint:** Use that $\mathbf{V}^T\mathbf{V} = \mathbf{I}$.

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

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

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

$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthonormal 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

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

$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthonormal 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 $\mathcal{V}$.

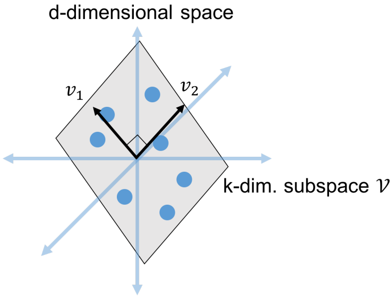
$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthonormal basis for subspace $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

# 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 $\mathcal{V}$.

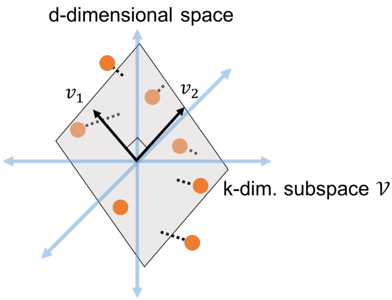



$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthonormal basis for subspace $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

**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

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

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



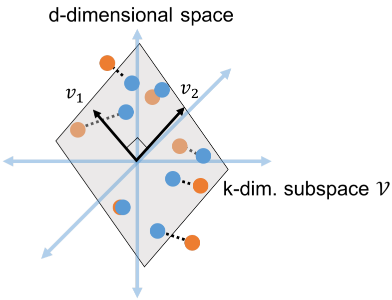

$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthonormal basis for subspace $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

# 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

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

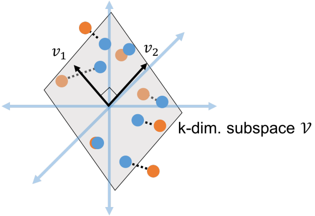
- $\mathbf{V}\mathbf{V}^T$ is a projection matrix, which projects the rows of $\mathbf{X}$ (the data points $\vec{x}_1, \ldots, \vec{x}_n$) onto the subspace $\mathcal{V}$.




$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthonormal 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:

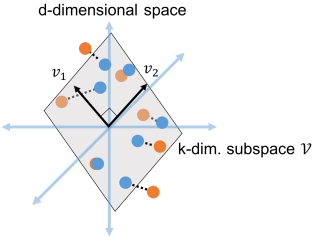
$$\mathbf{X} \approx \mathbf{X}\mathbf{V}\mathbf{V}^T$$




$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthonormal 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:

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

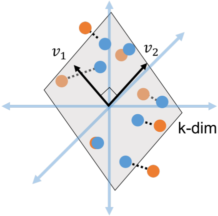



**Note:** $\mathbf{X}\mathbf{V}\mathbf{V}^T$ has rank $k$. It is a low-rank approximation of $\mathbf{X}$.

$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthonormal 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:

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




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

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

$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthonormal 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 $\mathbf{V}$).

$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: 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, $\|(\mathbf{X}\mathbf{V}\mathbf{V}^T)_i - (\mathbf{X}\mathbf{V}\mathbf{V}^T)_j\|_2 = \|[(\mathbf{X}\mathbf{V})_i - (\mathbf{X}\mathbf{V})_j]\mathbf{V}^T\|_2 = \|[(\mathbf{X}\mathbf{V})_i - (\mathbf{X}\mathbf{V})_j]\|_2$.

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

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