CS189: Intro to Machine Learning
Summer 2018

Lecture 7: Dimensionality reduction

Josh Tobin
UC Berkeley EECS

Outline

- Why dimensionality reduction?
- PCA

Outline

- **Why dimensionality reduction?**
- PCA

The curse of dimensionality

Things get weird in higher dimensions

- Numerical instability
- Increased variance (see bias-variance tradeoff)
- Distance functions

Ways to mitigate the CoD

- Regularization (e.g., ridge regression)
- Feature selection (pick only some columns of X)
- Dimensionality reduction - *unsupervised* (usually)

Goals of dimensionality reduction

- Have fewer features
  - Speed things up
  - Reduce variance
- Better numerical stability
- Keep “important” information
- Visualize things (e.g., in 2 or 3 dimensions)
- Anomaly detection (find data that is unusual)

“Important” information?

Many choices

- Euclidean distances
- Inner product distances
- Correlations with target variable
- Etc

Naive dimensionality reduction

Throw out some features

$$\underset{n\times d}{X}\underset{d\times k}{\begin{bmatrix} I \\ 0 \end{bmatrix}} = \underset{n\times k}{X^*}$$

$$k < d$$

Random projection

$$\underset{n\times d}{X}\underset{d\times k}{P} = \underset{n\times k}{X^*}$$

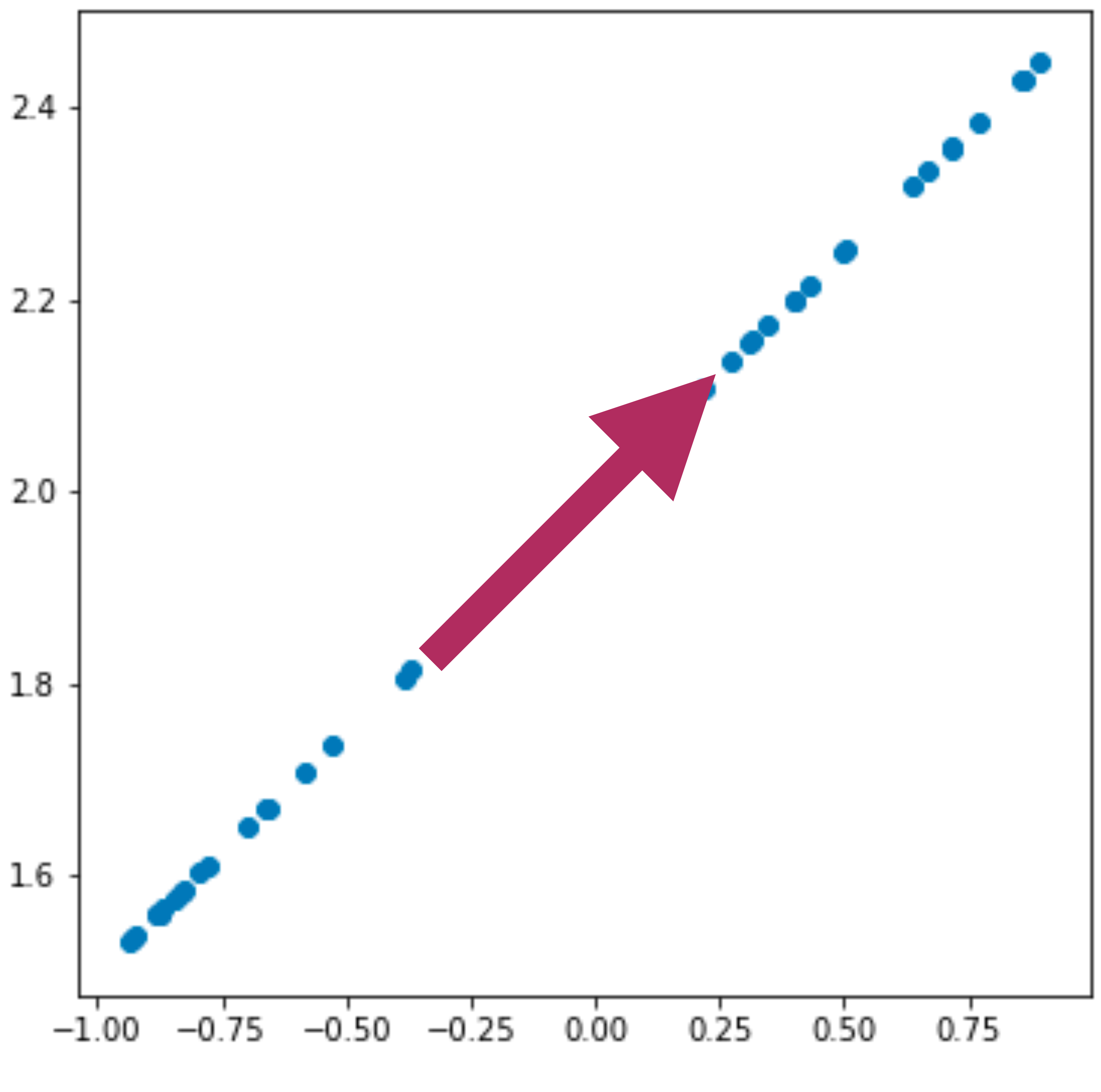
$$k < d$$

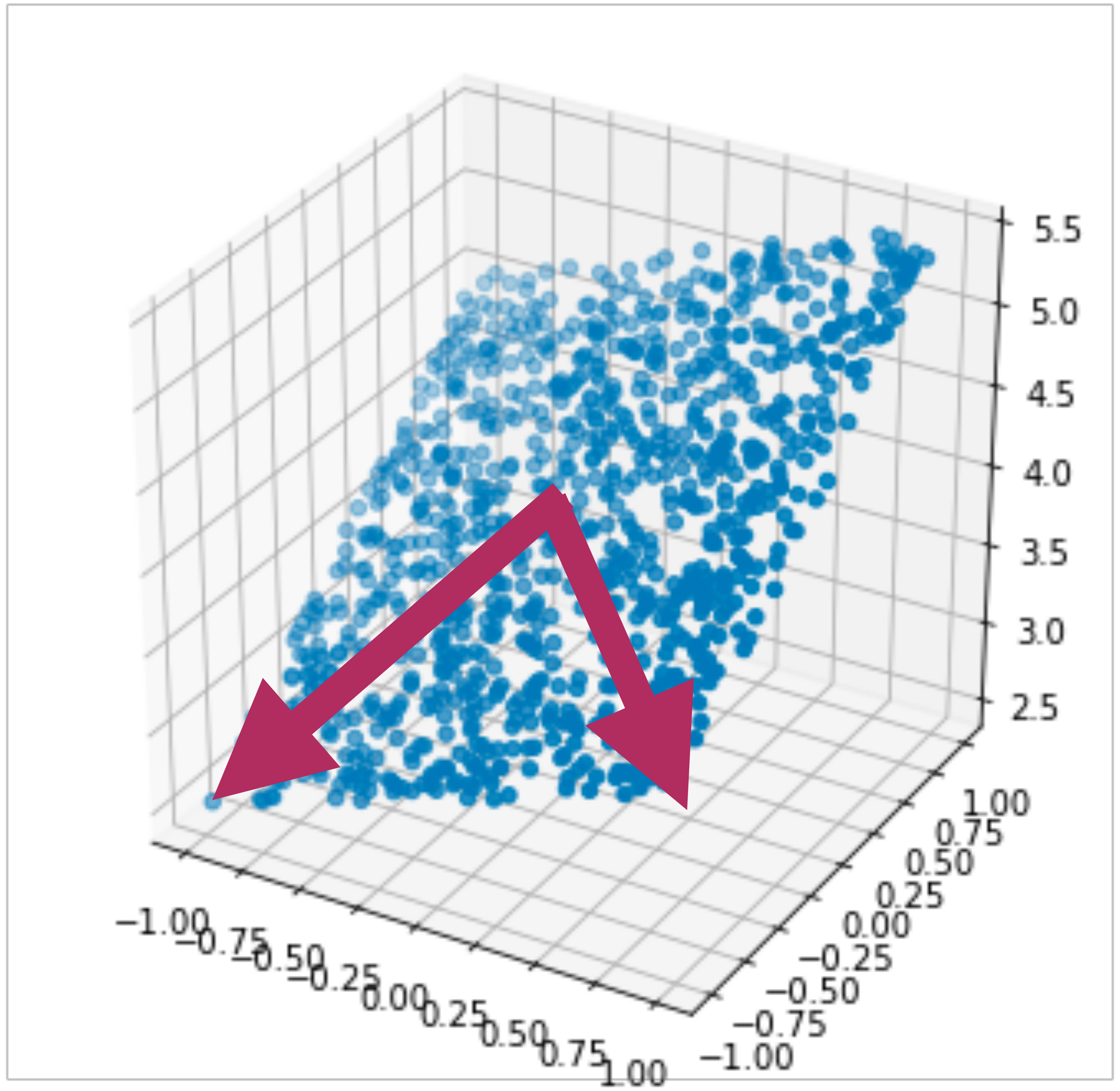
Outline

- Why dimensionality reduction?
- **PCA**

PCA

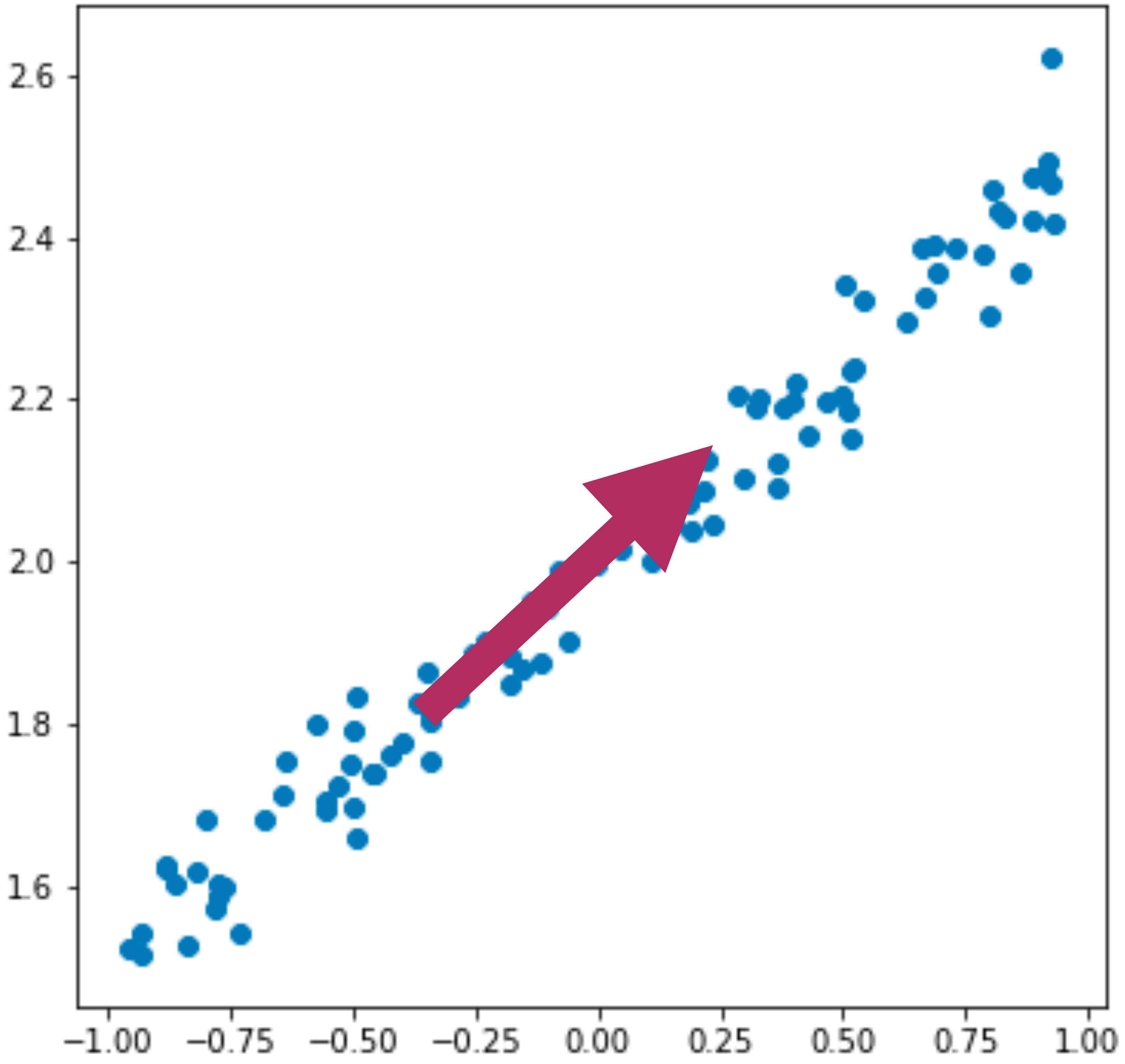
Basic idea: if data lies in a subspace, there are redundant dimensions (i.e., the data looks flat)

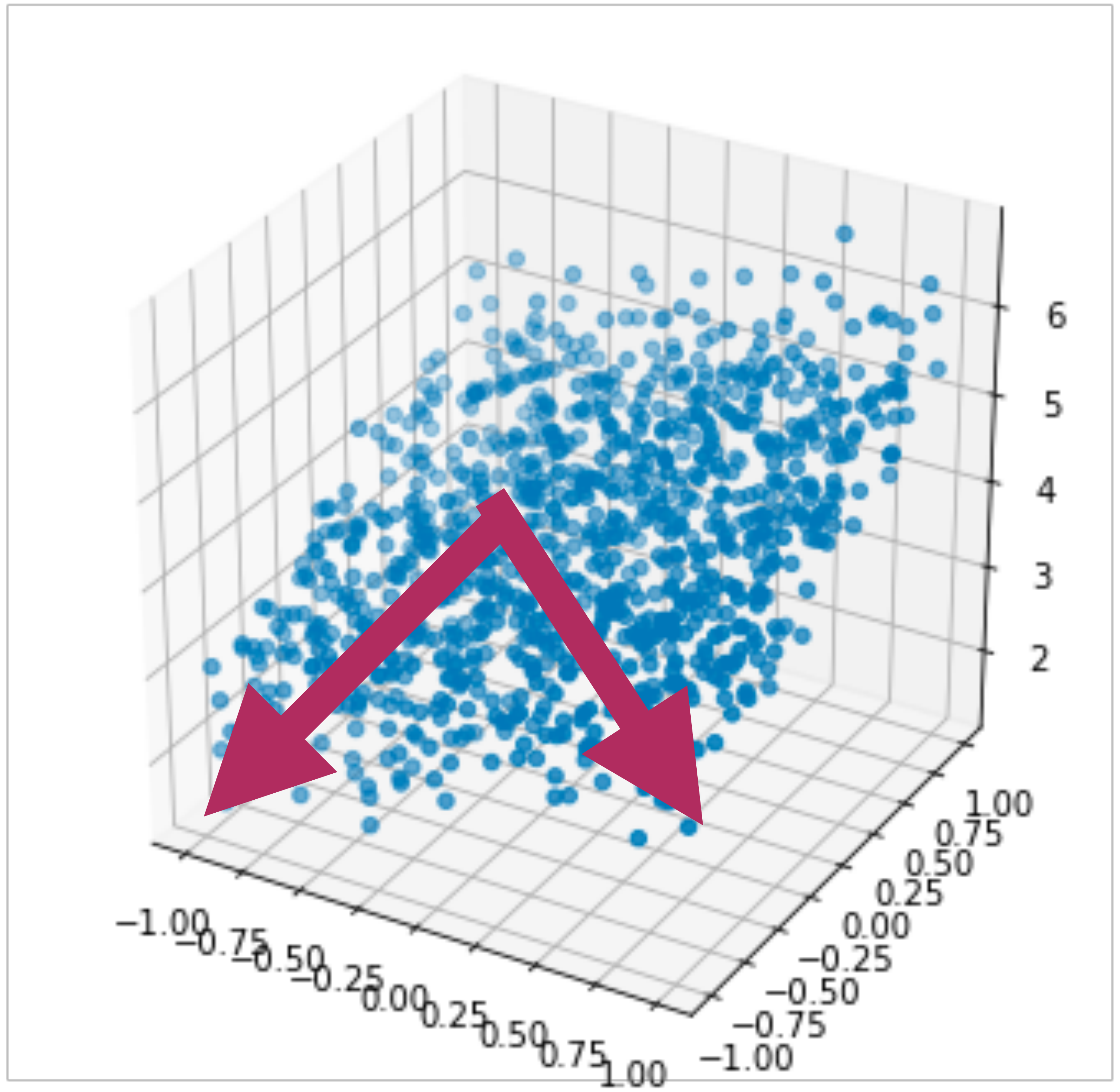




PCA

Basic idea: if data lies in a subspace, there are redundant dimensions (i.e., the data looks flat)





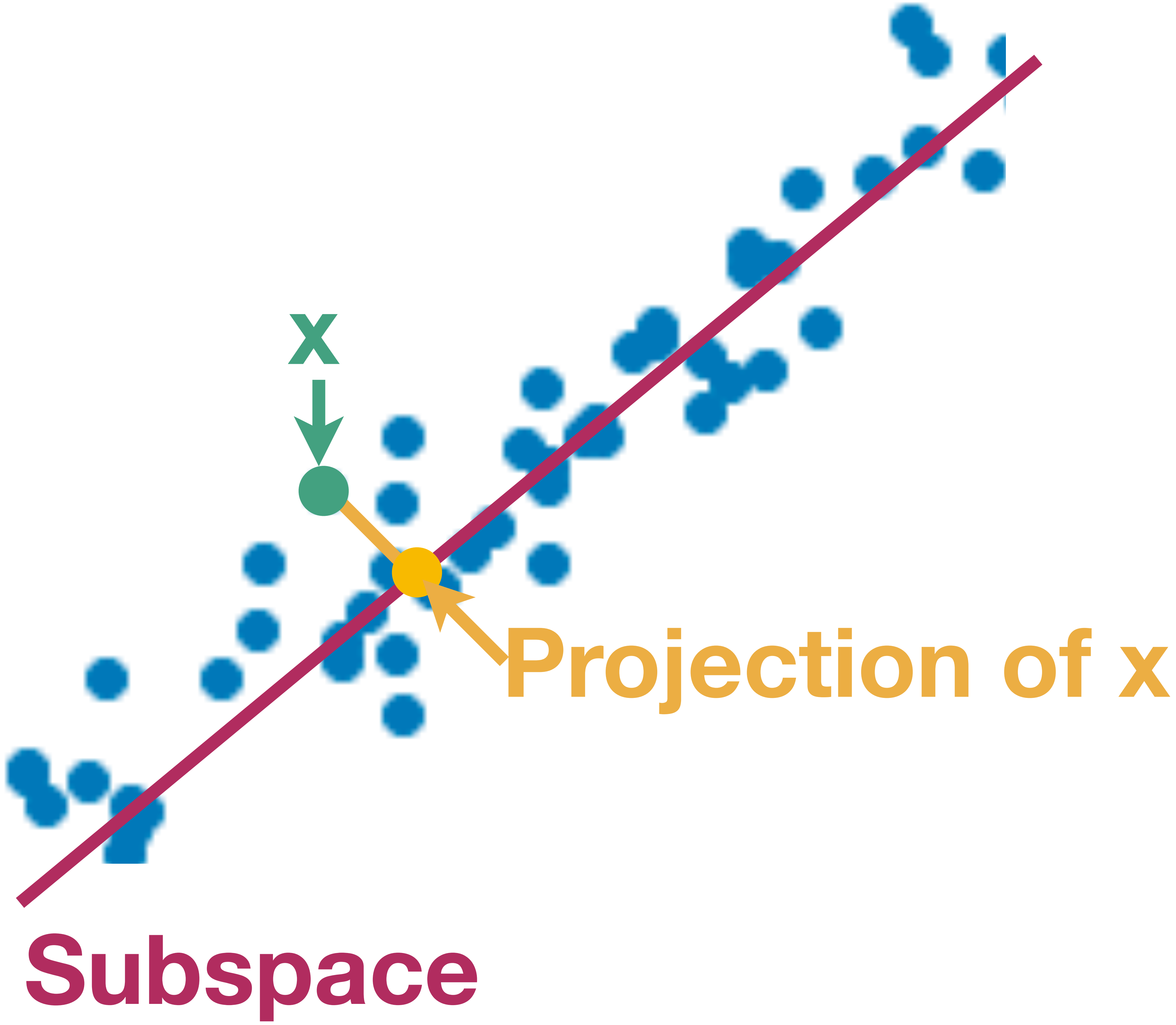
PCA

Basic idea: if data lies in a subspace, there are redundant dimensions (i.e., the data looks flat)

Goal: Find the k-dimensional subspace that minimizes the *projection error*

Projection of x onto unit vector v:

$$P_{\mathbf{v}}\mathbf{x} = (\mathbf{x}^{\top}\mathbf{v})\mathbf{v}$$

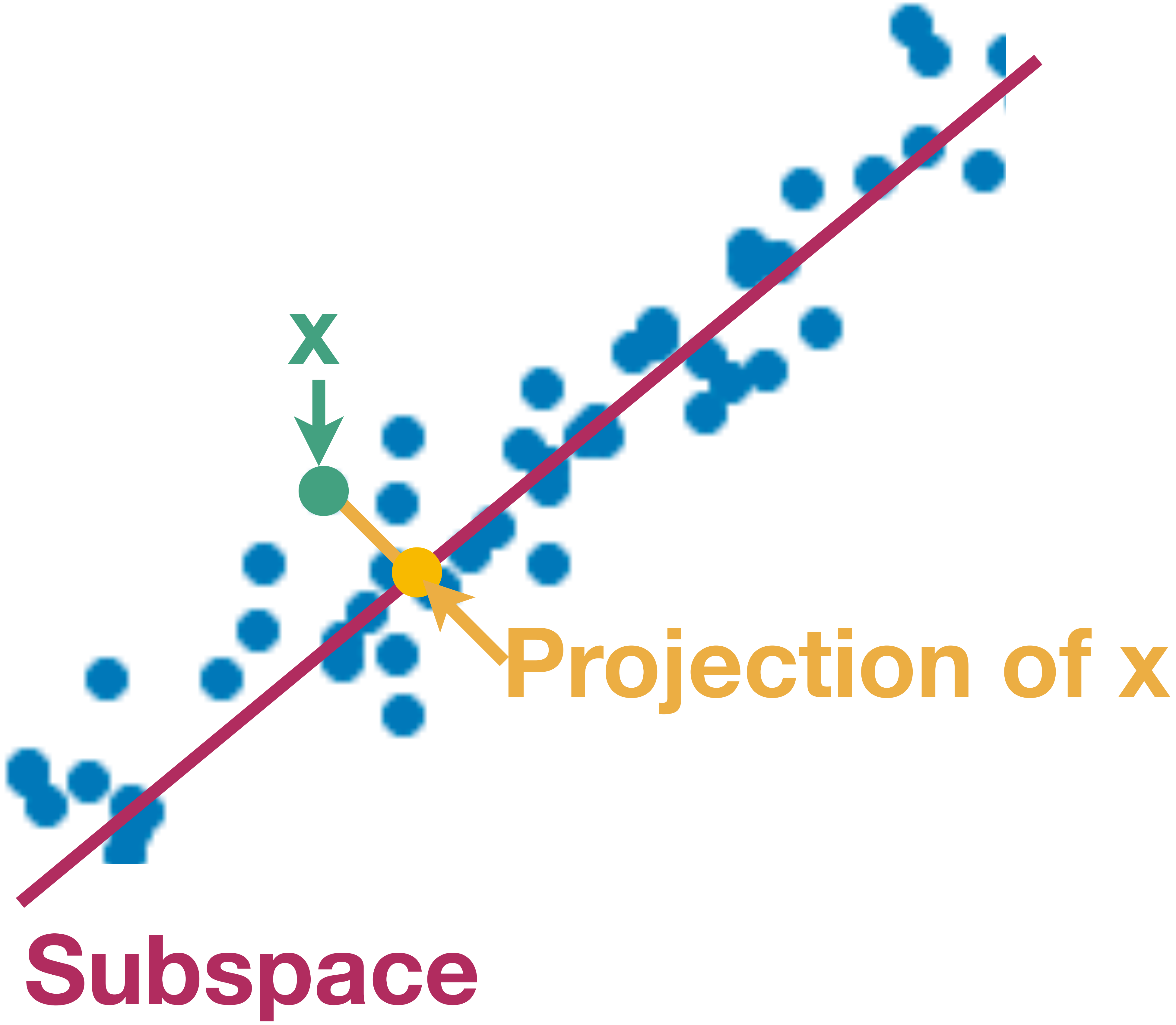


PCA

Goal: Find the k-dimensional subspace that minimizes the *projection error*

An algorithm

1. Start with X (n x d)
2. *Recenter.* Subtract mean from each row: $X_c = X - \text{mean}(X)$
3. *Compute covariance* $C = \frac{1}{n} X_c^T X_c$
4. *k* eigenvectors of C with highest eigenvalues are a basis for the subspace



PCA

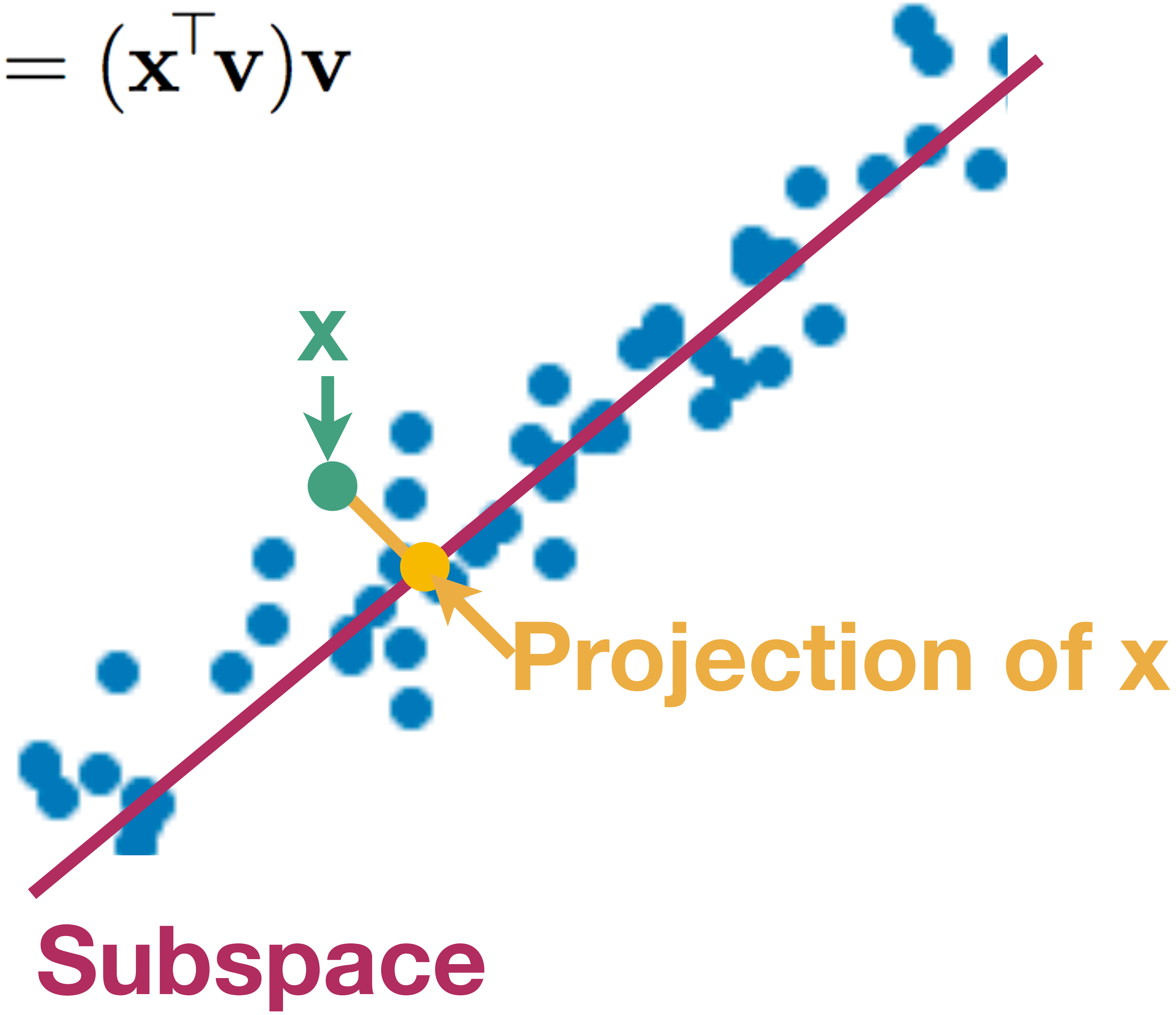
Goal: Find the k-dimensional subspace that minimizes the *projection error*

Projection of x onto unit vector v: $P_{\mathbf{v}}\mathbf{x} = (\mathbf{x}^\top \mathbf{v})\mathbf{v}$

Start with 1-dim subspace

Find: $\arg\min_{\mathbf{v}} \sum_{i=1}^{n} ||\mathbf{x}_i - P_{\mathbf{v}}\mathbf{x}_i||^2$

s.t. $||v|| = 1$



PCA

Find: $\arg\min_{\mathbf{v}} \sum_{i=1}^{n} ||\mathbf{x}_i - P_{\mathbf{v}}\mathbf{x}_i||^2 \quad \text{s.t. } ||v|| = 1$

$\mathbf{x} - P_{\mathbf{v}}\mathbf{x} \perp P_{\mathbf{v}}\mathbf{x}$, so $\|\mathbf{x} - P_{\mathbf{v}}\mathbf{x}\|^2 + \|P_{\mathbf{v}}\mathbf{x}\|^2 = \|\mathbf{x}\|^2$

$$\sum_{i=1}^{n} \|\mathbf{x}_i - P_{\mathbf{v}}\mathbf{x}_i\|^2$$

PCA

Find: $\arg\min_{\mathbf{v}} \sum_{i=1}^{n} ||\mathbf{x}_i - P_{\mathbf{v}}\mathbf{x}_i||^2 \quad \text{s.t.} \ \ ||v|| = 1$

$\mathbf{x} - P_{\mathbf{v}}\mathbf{x} \perp P_{\mathbf{v}}\mathbf{x}$, so $\|\mathbf{x} - P_{\mathbf{v}}\mathbf{x}\|^2 + \|P_{\mathbf{v}}\mathbf{x}\|^2 = \|\mathbf{x}\|^2$

$$\sum_{i=1}^{n} \|\mathbf{x}_i - P_{\mathbf{v}}\mathbf{x}_i\|^2 = \sum_{i=1}^{n} \left(\|\mathbf{x}_i\|^2 - \|P_{\mathbf{v}}\mathbf{x}_i\|^2\right)$$

PCA

Find: $\arg\min_{\mathbf{v}} \sum_{i=1}^{n} ||\mathbf{x}_i - P_{\mathbf{v}}\mathbf{x}_i||^2 \quad \text{s.t. } ||v|| = 1$

$\mathbf{x} - P_{\mathbf{v}}\mathbf{x} \perp P_{\mathbf{v}}\mathbf{x}$, so $\|\mathbf{x} - P_{\mathbf{v}}\mathbf{x}\|^2 + \|P_{\mathbf{v}}\mathbf{x}\|^2 = \|\mathbf{x}\|^2$

$$\begin{aligned}\sum_{i=1}^{n} \|\mathbf{x}_i - P_{\mathbf{v}}\mathbf{x}_i\|^2 &= \sum_{i=1}^{n} \left(\|\mathbf{x}_i\|^2 - \|P_{\mathbf{v}}\mathbf{x}_i\|^2\right) \\ &= \sum_{i=1}^{n} \|\mathbf{x}_i\|^2 - \sum_{i=1}^{n} \|(\mathbf{x}_i{}^\top \mathbf{v})\mathbf{v}\|^2\end{aligned}$$

PCA

Find: $\arg\min_{\mathbf{v}} \sum_{i=1}^{n} ||\mathbf{x}_i - P_{\mathbf{v}}\mathbf{x}_i||^2 \quad \text{s.t. } ||v|| = 1$

$\mathbf{x} - P_{\mathbf{v}}\mathbf{x} \perp P_{\mathbf{v}}\mathbf{x}$, so $\|\mathbf{x} - P_{\mathbf{v}}\mathbf{x}\|^2 + \|P_{\mathbf{v}}\mathbf{x}\|^2 = \|\mathbf{x}\|^2$

$$
\begin{aligned}
\sum_{i=1}^{n} \|\mathbf{x}_i - P_{\mathbf{v}}\mathbf{x}_i\|^2 &= \sum_{i=1}^{n} \left(\|\mathbf{x}_i\|^2 - \|P_{\mathbf{v}}\mathbf{x}_i\|^2\right) \\
&= \sum_{i=1}^{n} \|\mathbf{x}_i\|^2 - \sum_{i=1}^{n} \|(\mathbf{x}_i{}^\top \mathbf{v})\mathbf{v}\|^2 \\
&= \sum_{i=1}^{n} \|\mathbf{x}_i\|^2 - \sum_{i=1}^{n} (\mathbf{x}_i{}^\top \mathbf{v})^2
\end{aligned}
$$

PCA

$$\arg\min_{\mathbf{v}} \sum_{i=1}^{n} ||\mathbf{x}_i - P_{\mathbf{v}}\mathbf{x}_i||^2 = \arg\min_{\mathbf{v}} \sum_{i=1}^{n} ||\mathbf{x}_i||^2 - \sum_{i=1}^{n} (\mathbf{x}_i^T \mathbf{v})^2$$

$$= \arg\max_{\mathbf{v}} \sum_{i=1}^{n} (\mathbf{x}_i^T \mathbf{v})^2$$

$$= \arg\max_{\mathbf{v}} (X\mathbf{v})^T X\mathbf{v}$$

$$= \arg\max_{\mathbf{v}} \mathbf{v}^T X^T X\mathbf{v}$$

$$\text{s.t. } ||v|| = 1$$

PCA

$$\arg\min_{\mathbf{v}} \sum_{i=1}^{n} ||\mathbf{x}_i - P_{\mathbf{v}}\mathbf{x}_i||^2 \longrightarrow \arg\max_{\mathbf{v}} \mathbf{v}^T X^T X \mathbf{v} \text{ s.t. } ||\mathbf{v}|| = 1$$

$$\text{s.t. } ||v|| = 1$$

Lagrange multipliers

$$\mathcal{L}(\mathbf{v}, \lambda) = \mathbf{v}^T X^T X \mathbf{v} - \lambda(\mathbf{v}^T \mathbf{v} - 1)$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = 1 - \mathbf{v}^T \mathbf{v} = 0$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{v}} = 2X^T X \mathbf{v} - 2\lambda \mathbf{v} = 0$$

$$\implies \mathbf{v}^T \mathbf{v} = 1,\ X^T X \mathbf{v} = \lambda \mathbf{v}$$

So pick the eigenvector with biggest eigenvalue

PCA

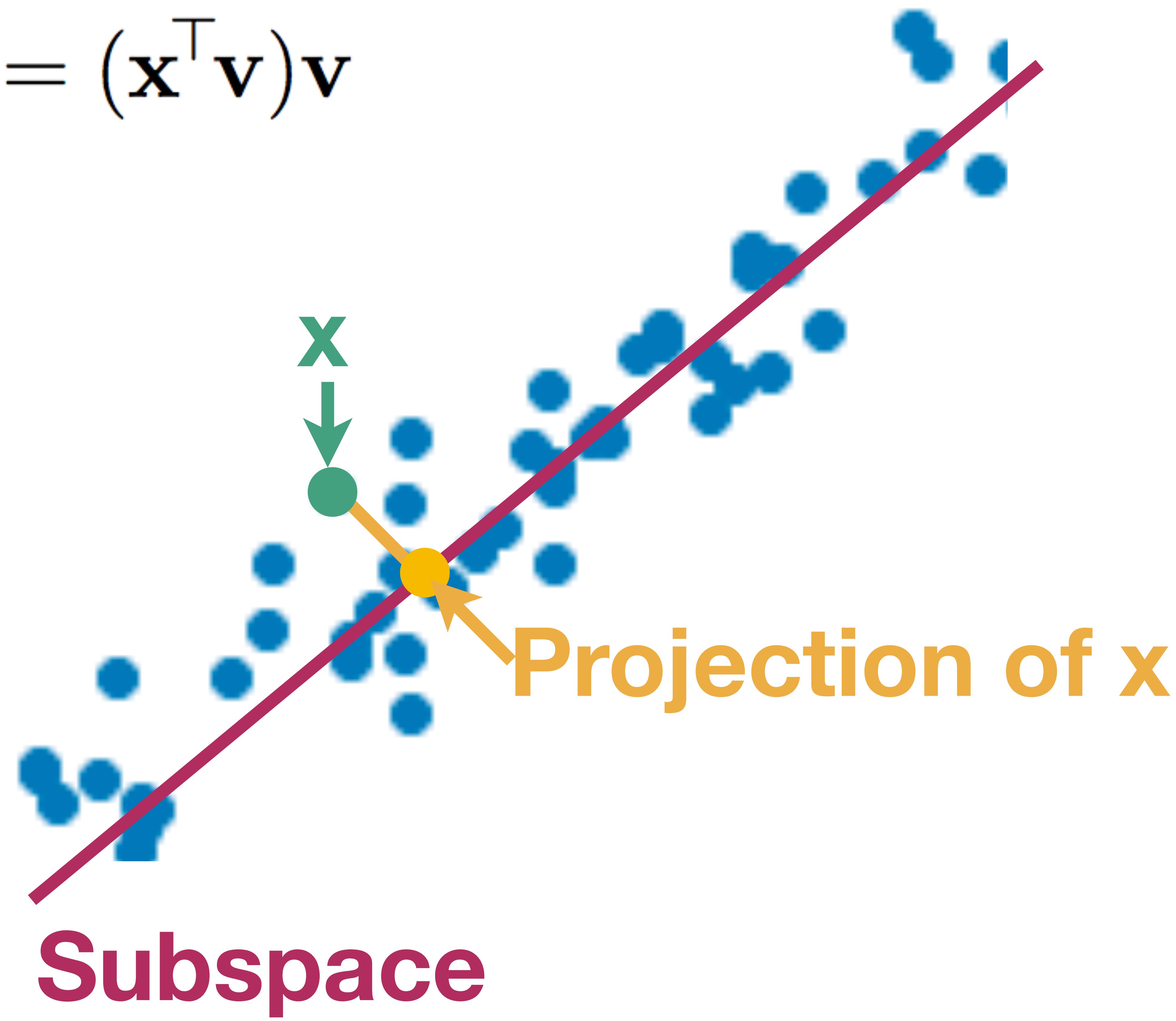
Goal: Find the k-dimensional subspace that minimizes the *projection error*

Projection of x onto unit vector v: $P_{\mathbf{v}}\mathbf{x} = (\mathbf{x}^\top \mathbf{v})\mathbf{v}$

Start with 1-dim subspace

Find: $\arg\min_{\mathbf{v}} \sum_{i=1}^{n} ||\mathbf{x}_i - P_{\mathbf{v}}\mathbf{x}_i||^2$

Eigenvector with largest eigenvalue



PCA

We argued that the ‘best’ 1-dimensional projection is onto the eigenvector with the largest eigenvalue

What about k-dimensional?

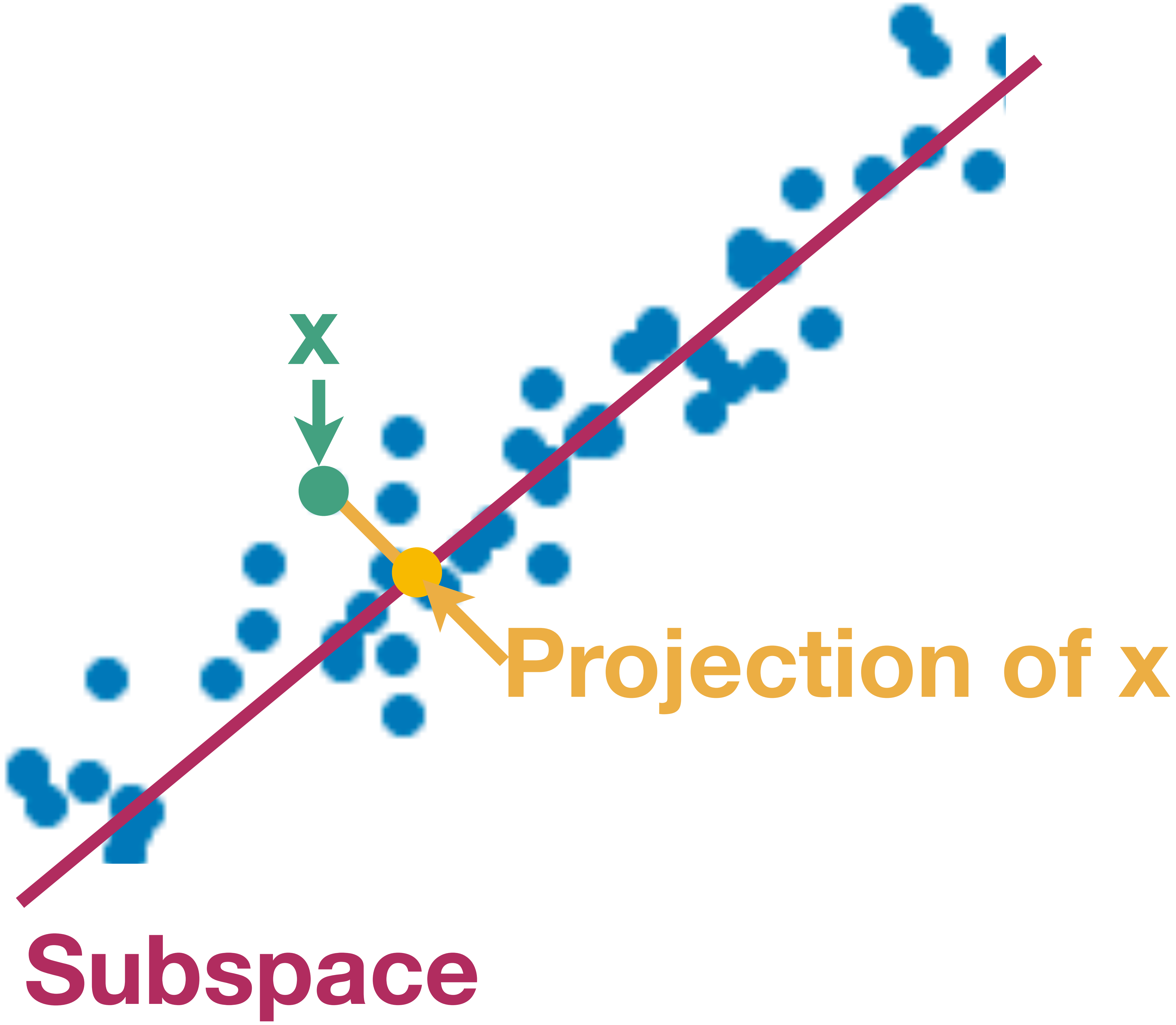
X^T X is SPD, so it has an orthonormal basis of eigenvectors. Hence the best projection orthogonal to the first is the eigenvector with the second largest eigenvalue, etc

PCA

Goal: Find the k-dimensional subspace that minimizes the *projection error*

An algorithm

1. Start with X (n x d)
2. *Recenter.* Subtract mean from each row: $X_c = X - \text{mean}(X)$
3. *Compute covariance* $C = \frac{1}{n} X_c^T X_c$
4. *k* eigenvectors of C with highest eigenvalues are basis for the subspace



PCA intuition

Goal: Find the k-dimensional subspace that minimizes the *projection error*

1. Minimizing projection error is the same as finding the directions with largest variance

$$\arg\min_{\mathbf{v}} \sum_{i=1}^{n} ||\mathbf{x}_i - P_{\mathbf{v}}\mathbf{x}_i||^2 \ = \arg\max_{\mathbf{v}} \mathbf{v}^T X^T X \mathbf{v}$$

variance of data = captured variance + reconstruction error

2. Reprojecting by the transpose of the projection matrix gives the best rank-k approximate to X

Another way of computing PCA

Method 1: eigendecomposition

PCs are eigenvectors of covariance matrix $C = (1/n)\ X^T X$

Computational complexity: $O(n\ d^2)$

Method 2: Singular value decomposition

$X = U\ S\ V^T$

V are principal components

Computational complexity: $O(n\ d\ k)$ [k is dim you’re reducing to]