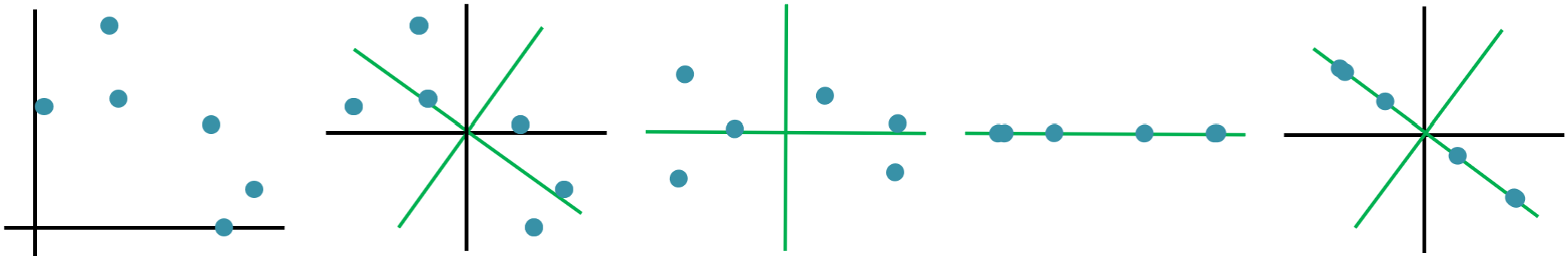


07-380 AI & ML II

Low-rank Optimization & PCA



Instructors: Pat Virtue & Mohammad Salameh

Slide credits: CMU AI with drawings from <http://ai.berkeley.edu>

Some Definitions to Start

Dimensionality Reduction

Transforming data from a **high-dimensional** representation to a **lower-dimensional** representation, while preserving important information (and removing distracting unimportant information)

Low-rank Optimization

Finding a low-rank matrix A' that best approximates a given matrix A :

$$\begin{aligned} \min_{A'} \quad & \|A - A'\|_2^2 \\ \text{s.t.} \quad & \text{rank}(A') \leq r \end{aligned}$$

Some Definitions to Start

*(pardon the simplification of
linear algebra terms)*

Basis

An M -D coordinate system defined by a set of M vectors

Change of basis

Converting data from one coordinate system to another

Orthogonal basis

Basis defined by a set of orthogonal vectors

Principal components

Given a dataset, the principal components are an ordered sequence of orthogonal basis vectors, where for all $K \in \{1, \dots, M\}$, the first K vectors span the “best” subspace of the data.

PCA Example

MRI Growth Plate

Growth Plate Imaging

Growth Plate Disruption and Limb Length Discrepancy



8 year-old boy with previous fracture
and 4cm leg length discrepancy

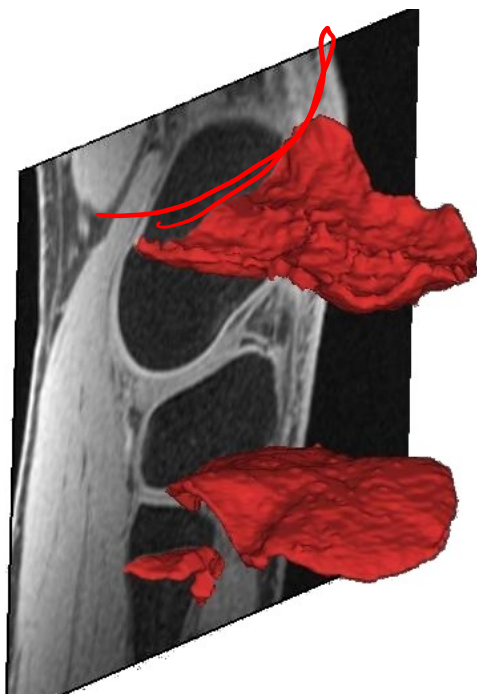


Images Courtesy
H. Potter, H.S.S.

Growth Plate Imaging

Growth Plate Disruption and Limb Length Discrepancy

8 year-old boy with previous fracture
and 4cm leg length discrepancy

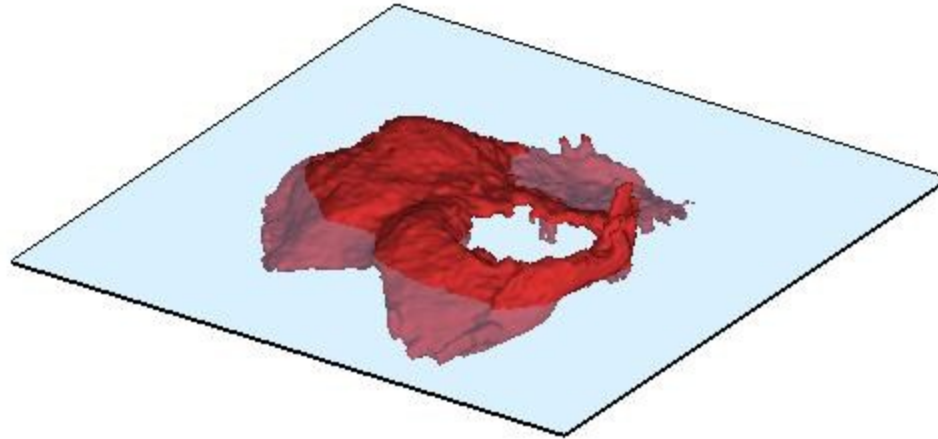


Images Courtesy
H. Potter, H.S.S.

Growth Plate Imaging

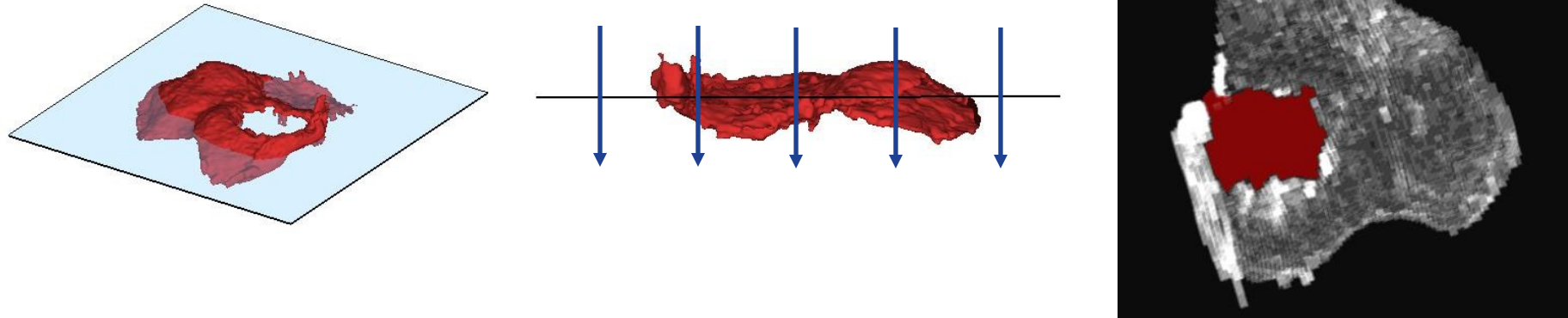
Area Measurement

Spoiler:
The PCA axes are eigenvectors!



Growth Plate Imaging

Area Measurement



Flatten Growth Plate to Enable 2D Area Measurement

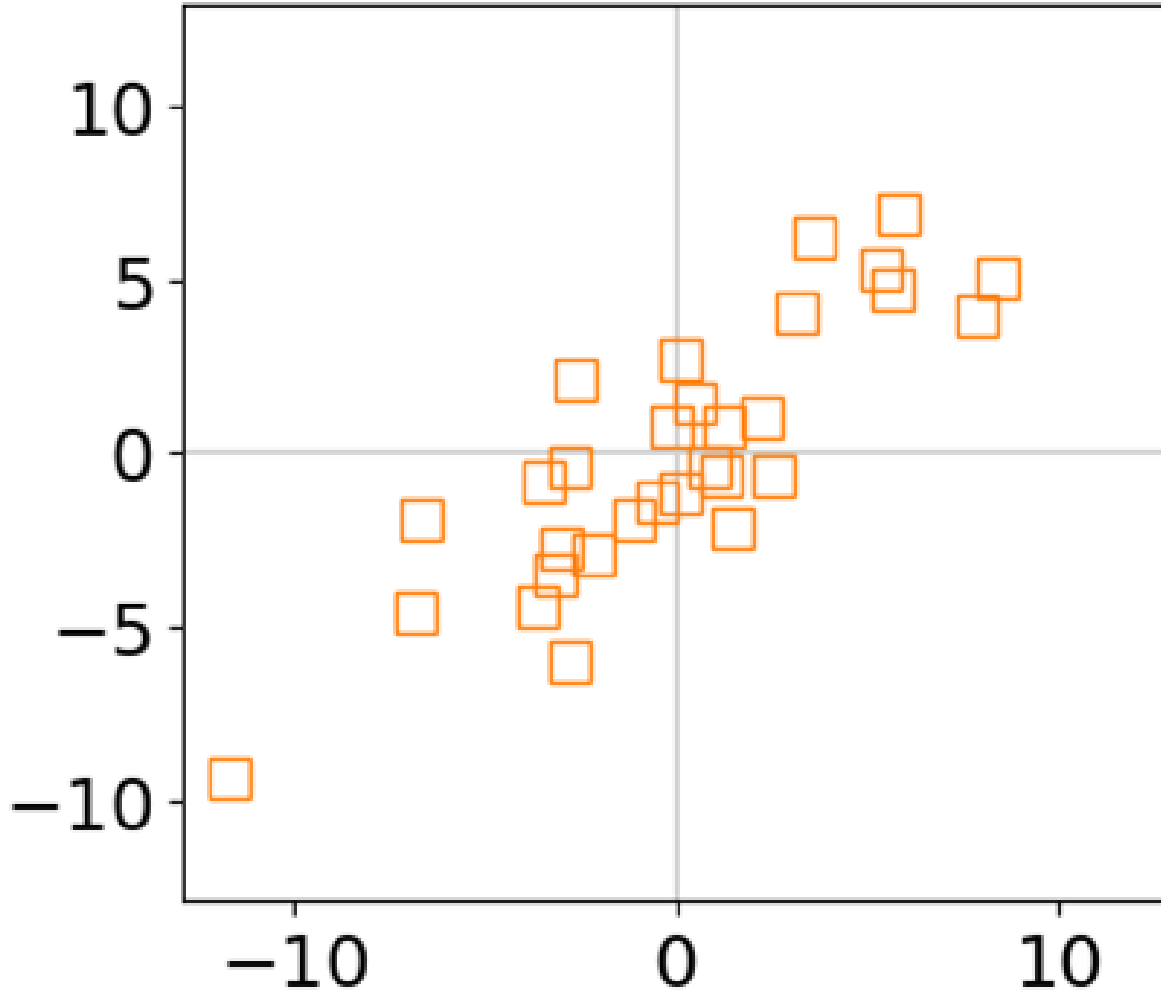
PCA Geometry Examples

2D: 1st and 2nd principal components

3D: 1st, 2nd, 3rd principal components

Principal Component Axes

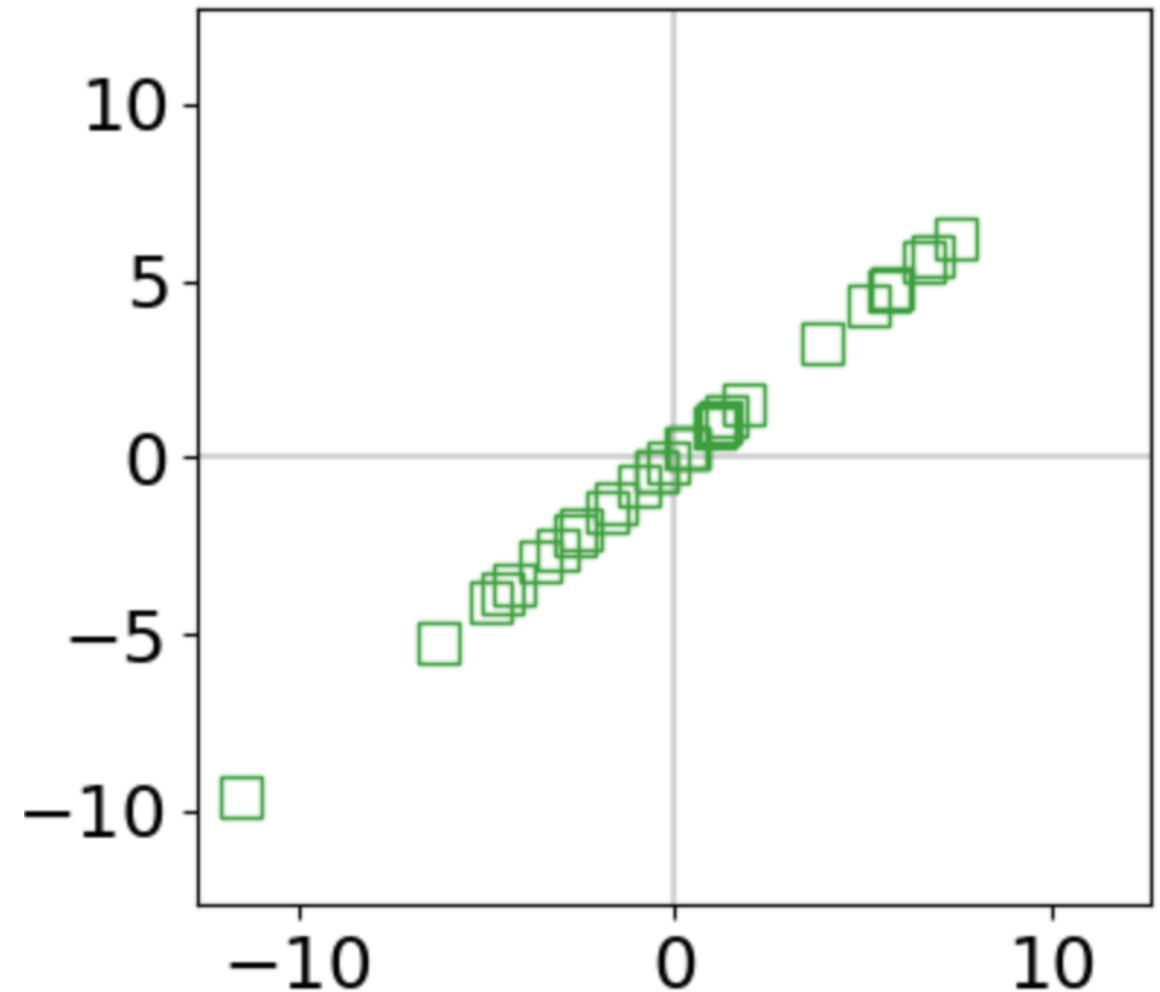
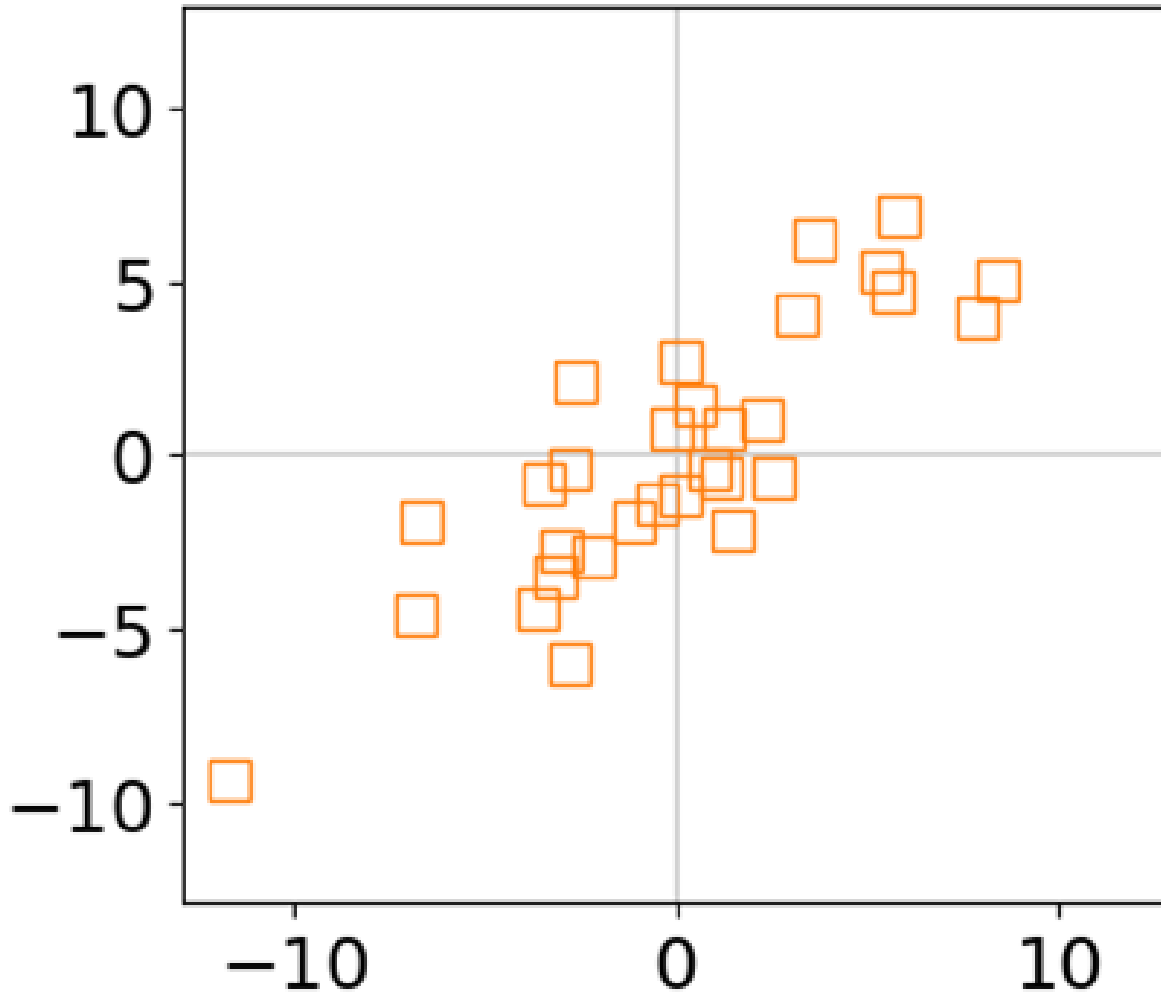
2-D Gaussian Data: 1st and 2nd principal component axes



Spoiler: The PCA axes can be found using eigenvectors!

PCA Dimensionality Reduction

2-D Gaussian Data: Reduced along 1st principal component



PCA Axes

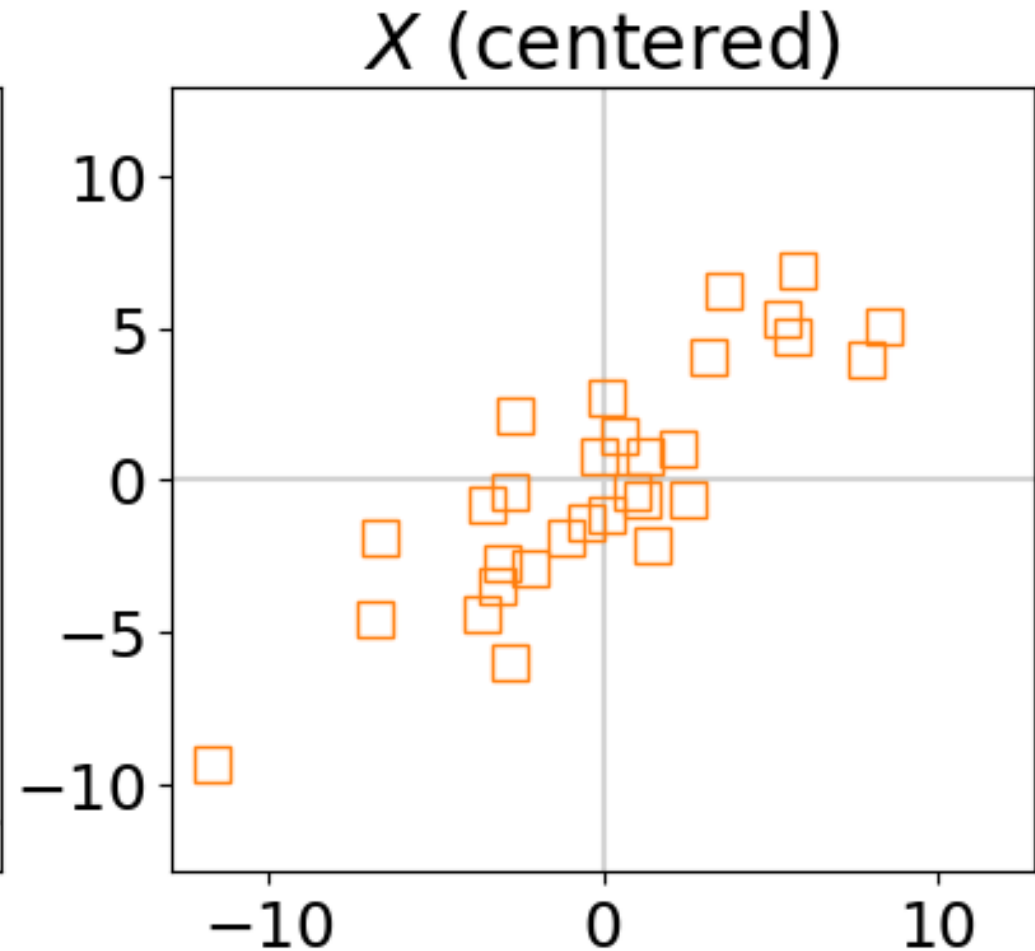
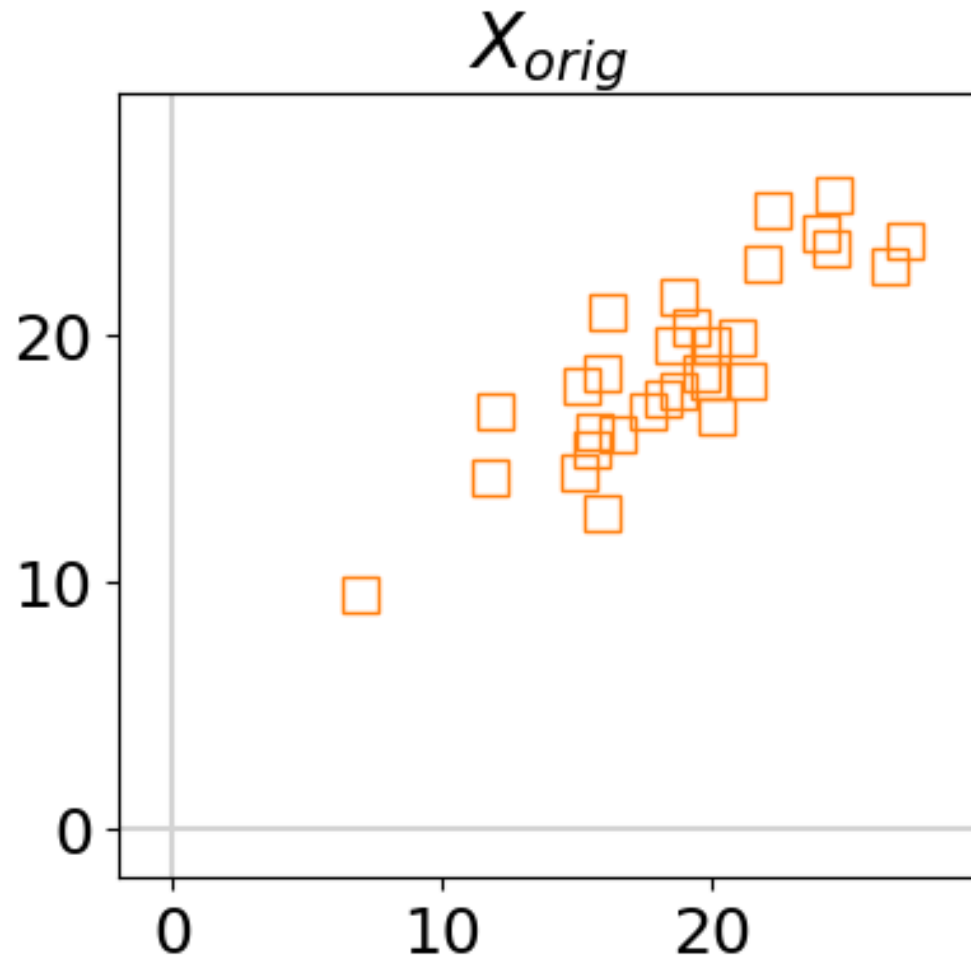
3-D Data

PCA Prepwork

Centering the data

PCA: Pre-processing

What if the data isn't centered



PCA: Centering Data

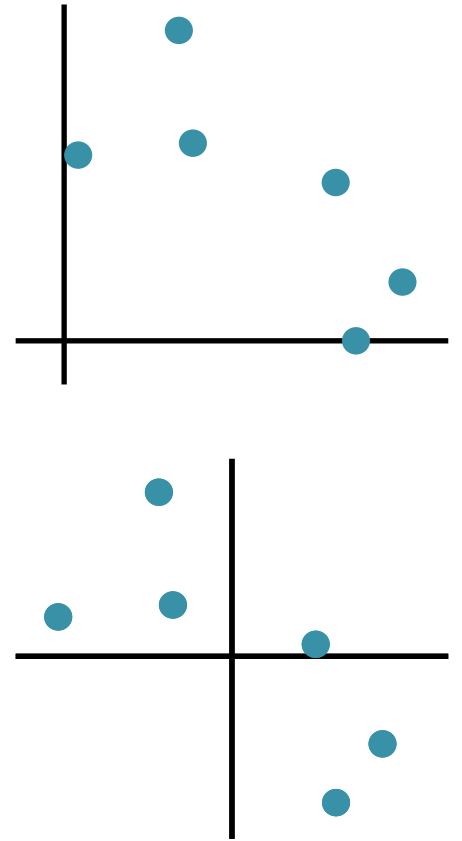
$$\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^N \quad \mathbf{X} = \begin{bmatrix} (\mathbf{x}^{(1)})^T \\ (\mathbf{x}^{(2)})^T \\ \vdots \\ (\mathbf{x}^{(N)})^T \end{bmatrix}$$

We assume the data is **centered**

$$\mu = \frac{1}{N} \sum_{i=1}^N \mathbf{x}^{(i)} = \mathbf{0}$$

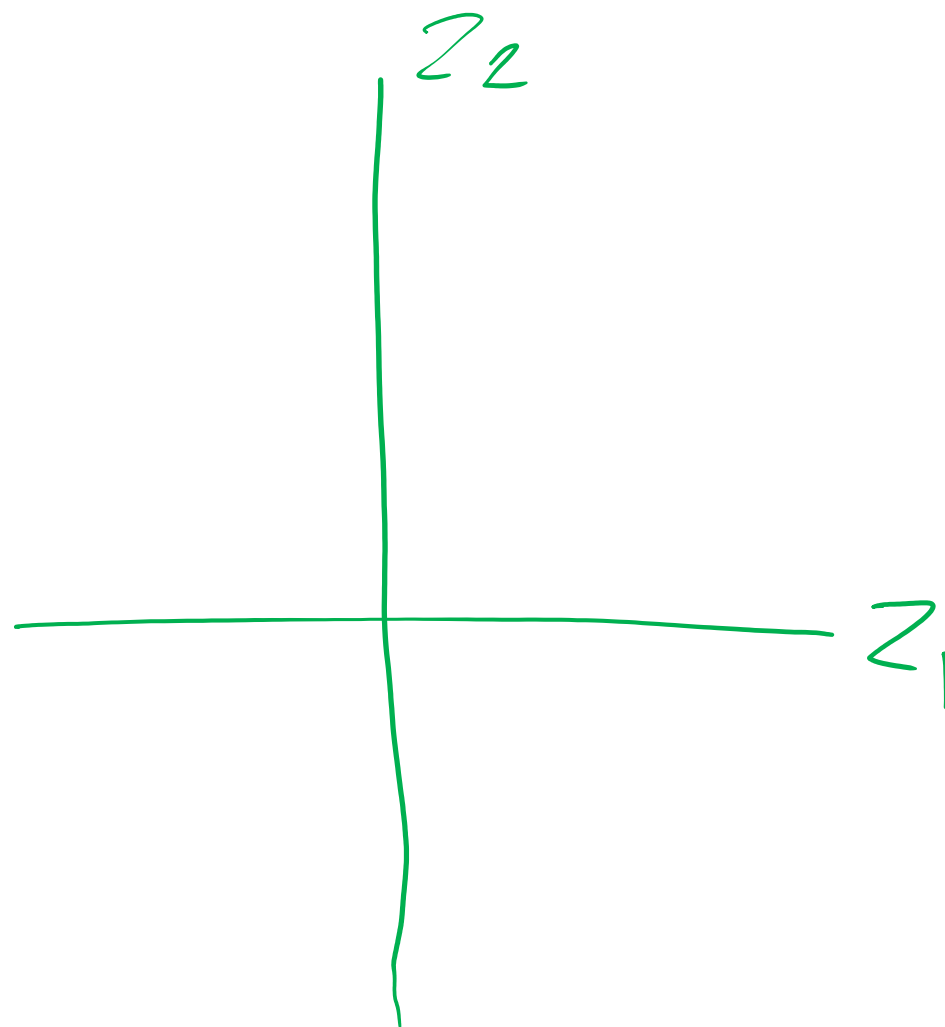
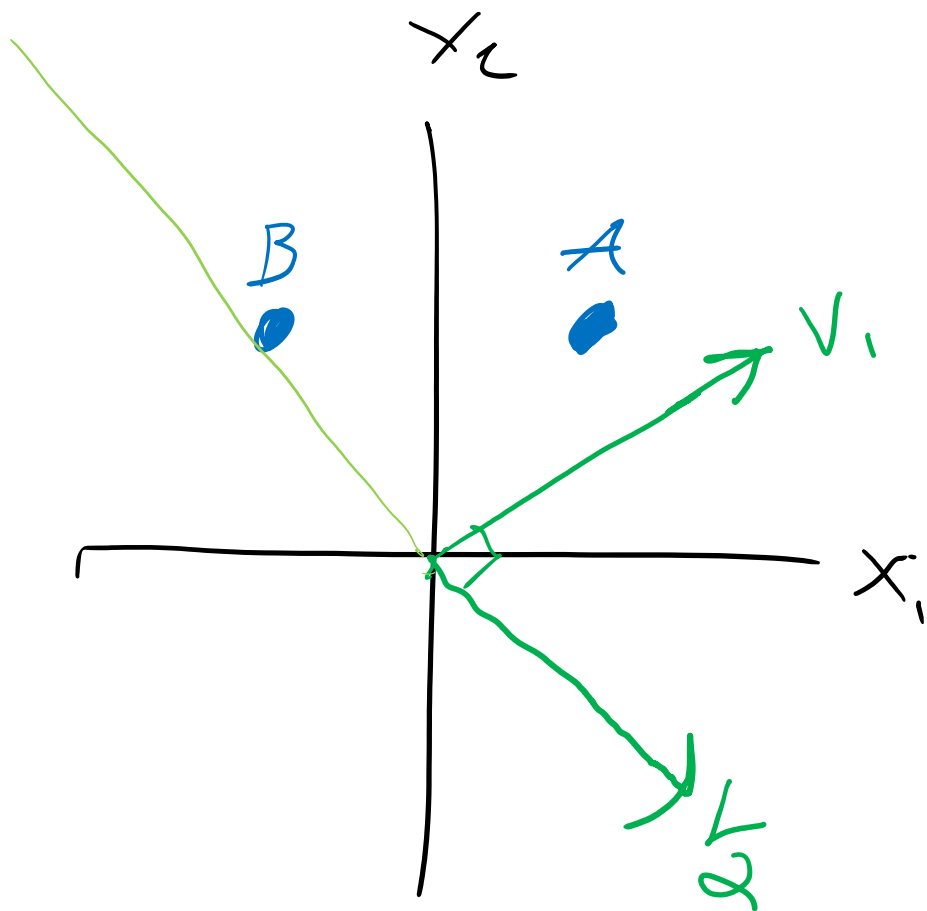
Q: What if
your data is
not centered?

A: Subtract
off the
sample mean



PCA Coordinate Transforms

Sketching by hand



PCA Coordinate Transforms

Using Linear Algebra

Poll 1

What is the projection of point $\mathbf{x}$ onto vector $\mathbf{v}$, assuming that $\|\mathbf{v}\|_2 = 1$?

A. $\mathbf{vx}$

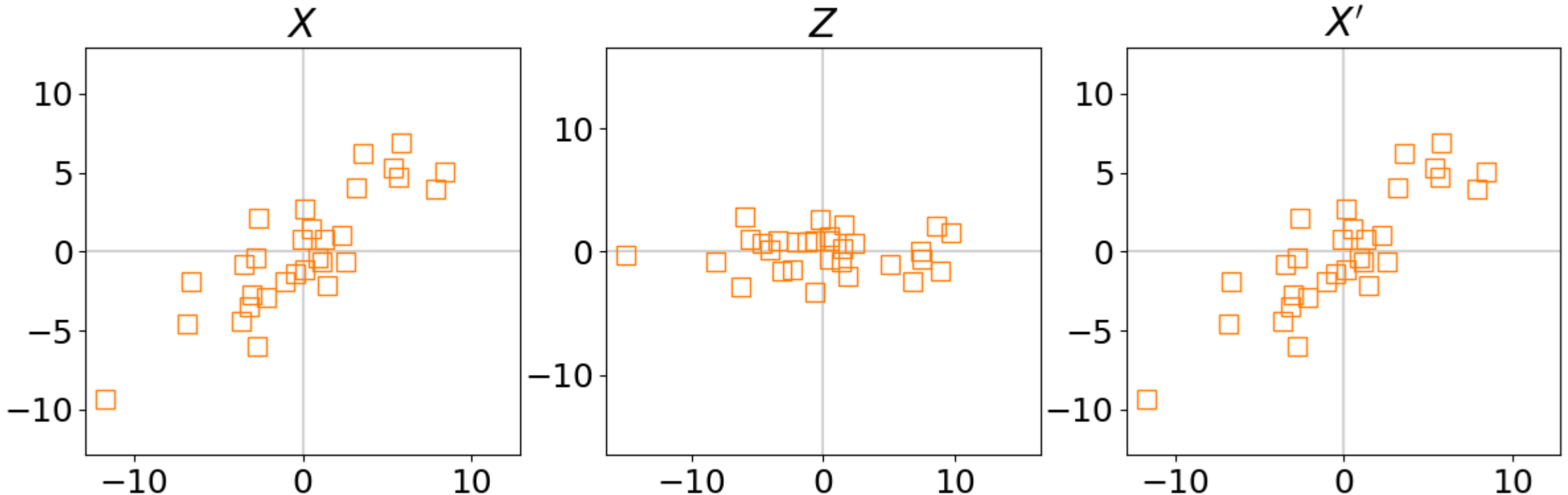
B. $\mathbf{v}^T \mathbf{x}$

C. $(\mathbf{v}^T \mathbf{x}) \mathbf{v}$

D. $\mathbf{v}^T \mathbf{x} \mathbf{x}^T \mathbf{v}$

Rotation of Data (and back)

1. For any orthogonal matrix $V \in \mathbb{R}^{M \times M}$
2. Rotate to new space: $\mathbf{z}^{(i)} = V^T \mathbf{x}^{(i)} \quad \forall i$
3. (Un)rotate back: $\mathbf{x}'^{(i)} = V \mathbf{z}^{(i)}$



PCA Algorithm

PCA Algorithm

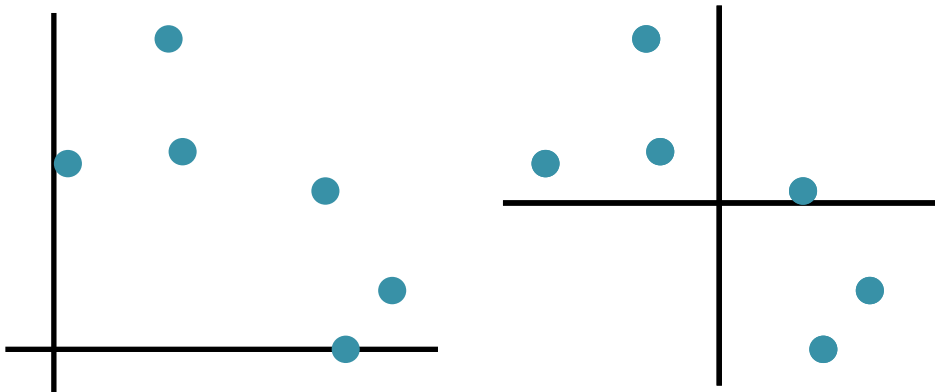
Input: X, X_{test}, K

1. Center data (and scale each axis) based on training data $\rightarrow X, X_{test}$
2. $V = \text{eigenvectors}(X^T X)$
3. Keep only the top K eigenvectors: V_K

PCA Algorithm

Input: X, X_{test}, K

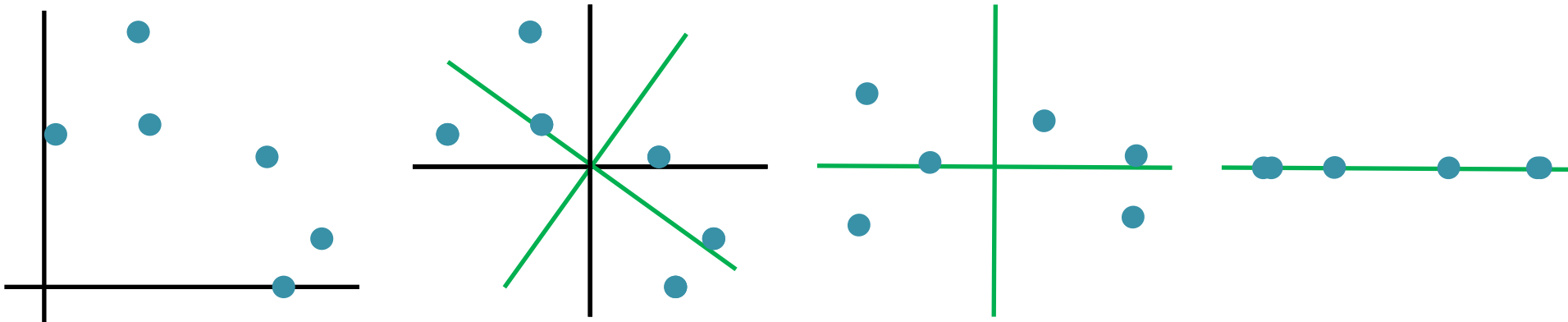
1. Center data (and scale each axis) based on training data $\rightarrow X, X_{test}$
2. $V = \text{eigenvectors}(X^T X)$
3. Keep only the top K eigenvectors: V_K



PCA Algorithm

Input: X, X_{test}, K

1. Center data (and scale each axis) based on training data $\rightarrow X, X_{test}$
2. $V = \text{eigenvectors}(X^T X)$
3. Keep only the top K eigenvectors: V_K
4. $Z_{test} = X_{test} V_K$

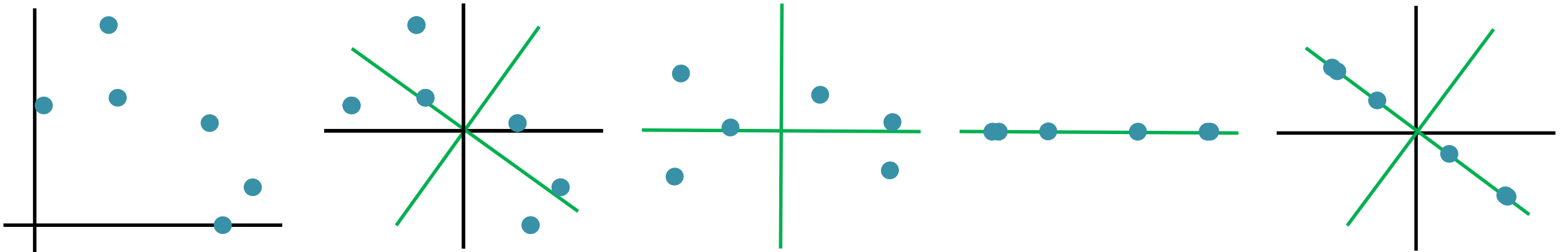


PCA Algorithm

Input: X, X_{test}, K

1. Center data (and scale each axis) based on training data $\rightarrow X, X_{test}$
2. $V = \text{eigenvectors}(X^T X)$
3. Keep only the top K eigenvectors: V_K
4. $Z_{test} = X_{test} V_K$

Optionally, use V_K^T to rotate Z_{test} back to original subspace X'_{test} and uncenter

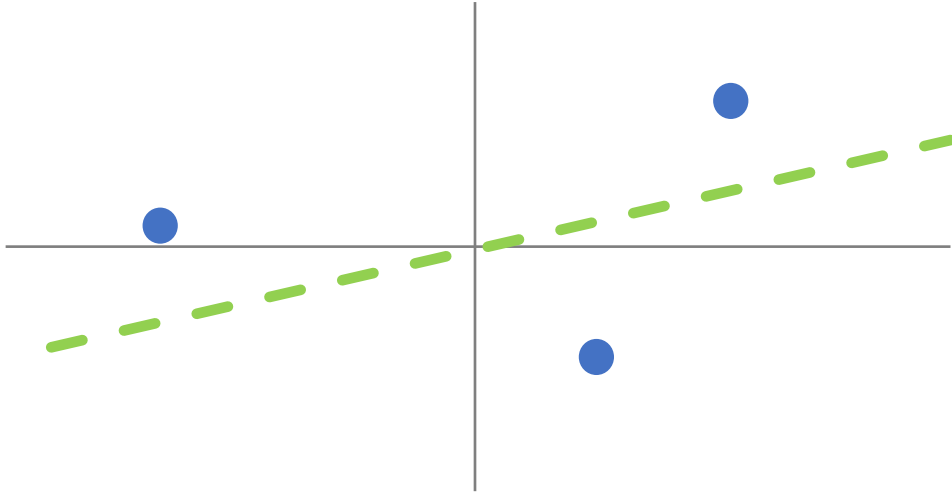


PCA Objective for Finding the 1st Principal Component

- Minimizing Reconstruction Error
- Maximizing Variance of (scalar) Projection

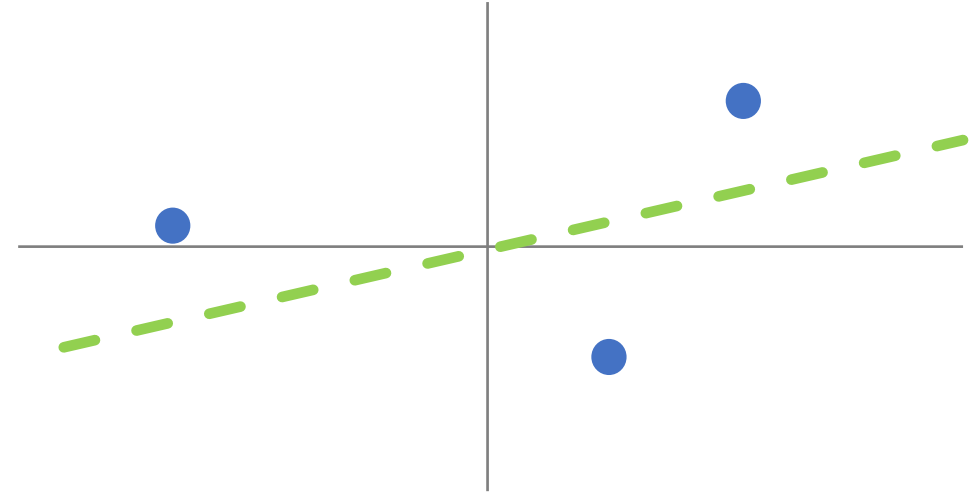
Select “Best” Vector

Reconstruction Error vs Variance of Projection



Reconstruction Error

$$\mathbf{v}^* = \underset{\substack{\mathbf{v} \\ s.t. \|\mathbf{v}\|_2=1}}{\operatorname{argmin}} \frac{1}{N} \sum_1^N \left\| \mathbf{x}^{(i)} - (\mathbf{x}^{\top(i)} \mathbf{v}) \mathbf{v} \right\|_2^2$$



(Sample) Variance of (scalar) Projection

$$\operatorname{Var}[\mathbf{d}] = \frac{1}{N} \sum_{i=1}^N (d^{(i)} - \bar{d})^2$$
$$\mathbf{v}^* = \underset{\substack{\mathbf{v} \\ s.t. \|\mathbf{v}\|_2=1}}{\operatorname{argmax}} \frac{1}{N} \sum_1^N (\mathbf{x}^{\top(i)} \mathbf{v})^2$$

Poll 2

Demo!

Play with the projection vector v in the Desmos Demo

What is the minimum Reconstruction MSE that you can find?

What is the maximum Variance of Scalar Projection that you can find?

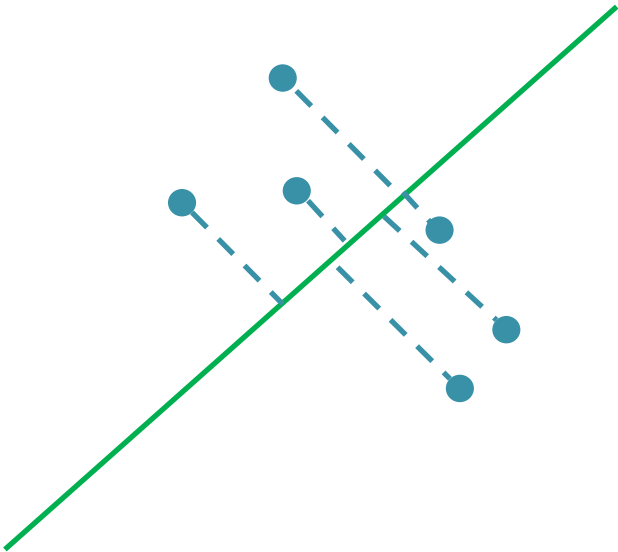
Poll 3

Consider the two projections below

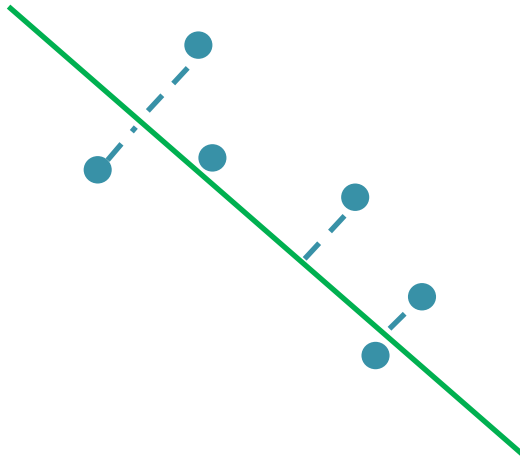
Q1) Which minimizes the reconstruction error?

Q2) Which maximizes the variance?

Option A



Option B



PCA Objective Equivalence

Equivalence of Maximizing Variance and Minimizing Reconstruction Error

Claim: Minimizing the reconstruction error is equivalent to maximizing the variance.

Proof: First, note that:

$$\|\mathbf{x}^{(i)} - (\mathbf{v}^T \mathbf{x}^{(i)})\mathbf{v}\|^2 = \|\mathbf{x}^{(i)}\|^2 - (\mathbf{v}^T \mathbf{x}^{(i)})^2 \quad (1)$$

since $\mathbf{v}^T \mathbf{v} = \|\mathbf{v}\|^2 = 1$.

Substituting into the minimization problem, and removing the extraneous terms, we obtain the maximization problem.

$$\mathbf{v}^* = \operatorname{argmin}_{\mathbf{v}: \|\mathbf{v}\|^2=1} \frac{1}{N} \sum_{i=1}^N \|\mathbf{x}^{(i)} - (\mathbf{v}^T \mathbf{x}^{(i)})\mathbf{v}\|^2 \quad (2)$$

$$= \operatorname{argmin}_{\mathbf{v}: \|\mathbf{v}\|^2=1} \frac{1}{N} \sum_{i=1}^N \|\mathbf{x}^{(i)}\|^2 - (\mathbf{v}^T \mathbf{x}^{(i)})^2 \quad (3)$$

$$= \operatorname{argmax}_{\mathbf{v}: \|\mathbf{v}\|^2=1} \frac{1}{N} \sum_{i=1}^N (\mathbf{v}^T \mathbf{x}^{(i)})^2 \quad (4)$$

Sketch of PCA

1. Select "best" $V \in \mathbb{R}^{M \times K}$
2. Project down: $\mathbf{z}^{(i)} = V^\top \mathbf{x}^{(i)} \quad \forall i$
3. Reconstruct up: $\mathbf{x}'^{(i)} = V \mathbf{z}^{(i)} \quad \forall i$

Definition of PCA

1. Select $\mathbf{v}_1$ that best explains data
2. Select next $\mathbf{v}_j$ that
 - i. Is orthogonal to $\mathbf{v}_1, \dots, \mathbf{v}_{j-1}$
 - ii. Best explains remaining data
3. Repeat 2 until desired amount of data is explained

Eigenvectors of Covariance Matrix
are the Principal Components!

PCA Eigenvalues and Eigenvectors

PCA: The First Principal Component

Use method of Lagrange multipliers to show that the first principal component is an eigenvalue of the covariance matrix

PCA: the First Principal Component

To find the first principal component, we wish to solve the following constrained optimization problem (variance maximization).

$$\mathbf{v}_1 = \operatorname{argmax}_{\mathbf{v}: \|\mathbf{v}\|^2=1} \mathbf{v}^T \boldsymbol{\Sigma} \mathbf{v} \quad (1)$$

So we turn to the method of Lagrange multipliers. The Lagrangian is:

$$\mathcal{L}(\mathbf{v}, \lambda) = \mathbf{v}^T \boldsymbol{\Sigma} \mathbf{v} - \lambda(\mathbf{v}^T \mathbf{v} - 1) \quad (2)$$

Taking the derivative of the Lagrangian and setting to zero gives:

$$\frac{d}{d\mathbf{v}} (\mathbf{v}^T \boldsymbol{\Sigma} \mathbf{v} - \lambda(\mathbf{v}^T \mathbf{v} - 1)) = 0 \quad (3)$$

$$\boldsymbol{\Sigma} \mathbf{v} - \lambda \mathbf{v} = 0 \quad (4)$$

$$\boldsymbol{\Sigma} \mathbf{v} = \lambda \mathbf{v} \quad (5)$$

Recall: For a square matrix $\mathbf{A}$, the vector $\mathbf{v}$ is an **eigenvector** iff there exists **eigenvalue** λ such that:

$$\mathbf{A} \mathbf{v} = \lambda \mathbf{v} \quad (6)$$

PCA: Choosing K

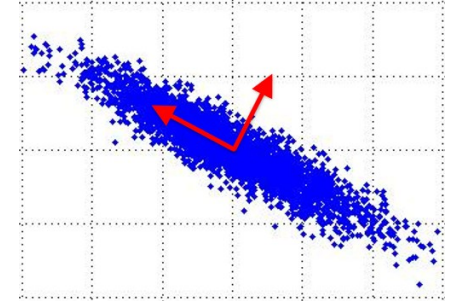
Look at the eigen values!

Principal Component Analysis (PCA)

$(X^T X) \mathbf{v} = \lambda \mathbf{v}$, so $\mathbf{v}$ (the first PC) is the eigenvector of sample covariance matrix $X^T X$

Sample variance of projection $\mathbf{v}^T X^T X \mathbf{v} = \lambda \mathbf{v}^T \mathbf{v} = \lambda$

Thus, the eigenvalue λ denotes the amount of variability captured along that dimension (aka amount of energy along that dimension).

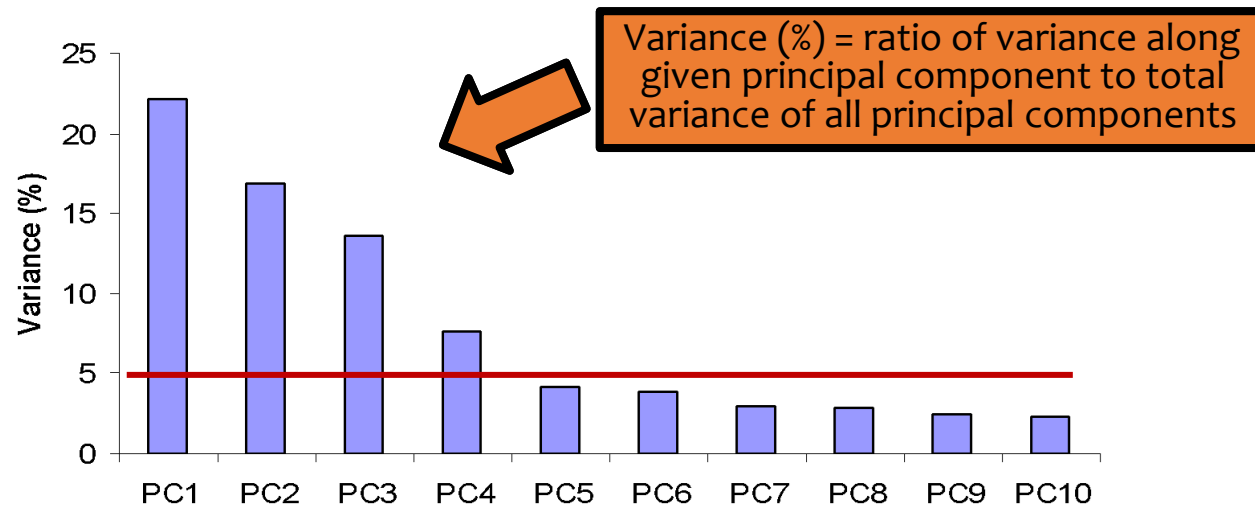


Eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \dots$

- The 1st PC $\mathbf{v}_1$ is the eigenvector of the sample covariance matrix $X^T X$ associated with the largest eigenvalue
- The 2nd PC $\mathbf{v}_2$ is the eigenvector of the sample covariance matrix $X^T X$ associated with the second largest eigenvalue
- And so on ...

How Many PCs?

- ⑩ For M original dimensions, sample covariance matrix is $M \times M$, and has up to M eigenvectors. So M PCs.
- ⑩ Where does dimensionality reduction come from?
Can *ignore* the components of lesser significance.



- You do *lose some information*, but if the eigenvalues are small, you don't lose much
 - M dimensions in original data
 - calculate M eigenvectors and eigenvalues
 - choose only the first D eigenvectors, based on their eigenvalues
 - final data set has only D dimensions

SVD for PCA

SVD matrix factorization

$$X = USV^T, A \in \mathbb{R}^{N \times M}$$

U : $N \times N$ orthogonal matrix

- Columns of U are *left* singular vectors of X
- Columns of U are eigenvectors of XX^T

V : $M \times M$ orthogonal matrix

- Columns of V are *right* singular vectors of X
- Columns of V are eigenvectors of $X^T X$

S : $N \times M$ diagonal matrix

- Diagonal entries are singular values of X , σ_k
- Each σ_k^2 are the eigenvalues of both XX^T and $X^T X$!!

PCA versus other Linear Transforms

Where else have we seen linear transforms?