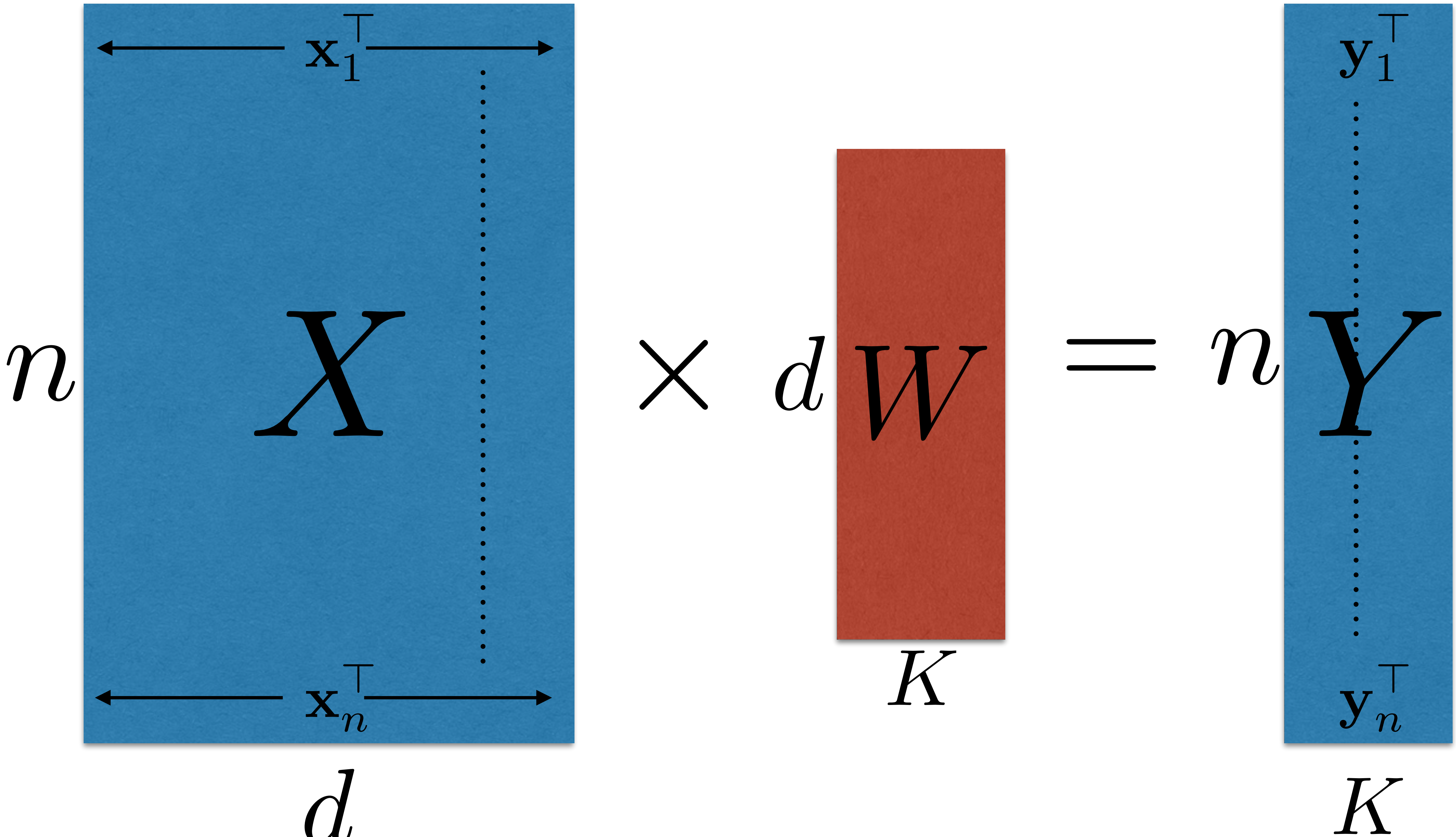
Machine Learning for Data Science (CS4786)
Lecture 3

Principal Component Analysis

Course Webpage :
http://www.cs.cornell.edu/Courses/cs4786/2016fa/

Dim Reduction: Linear Transformation

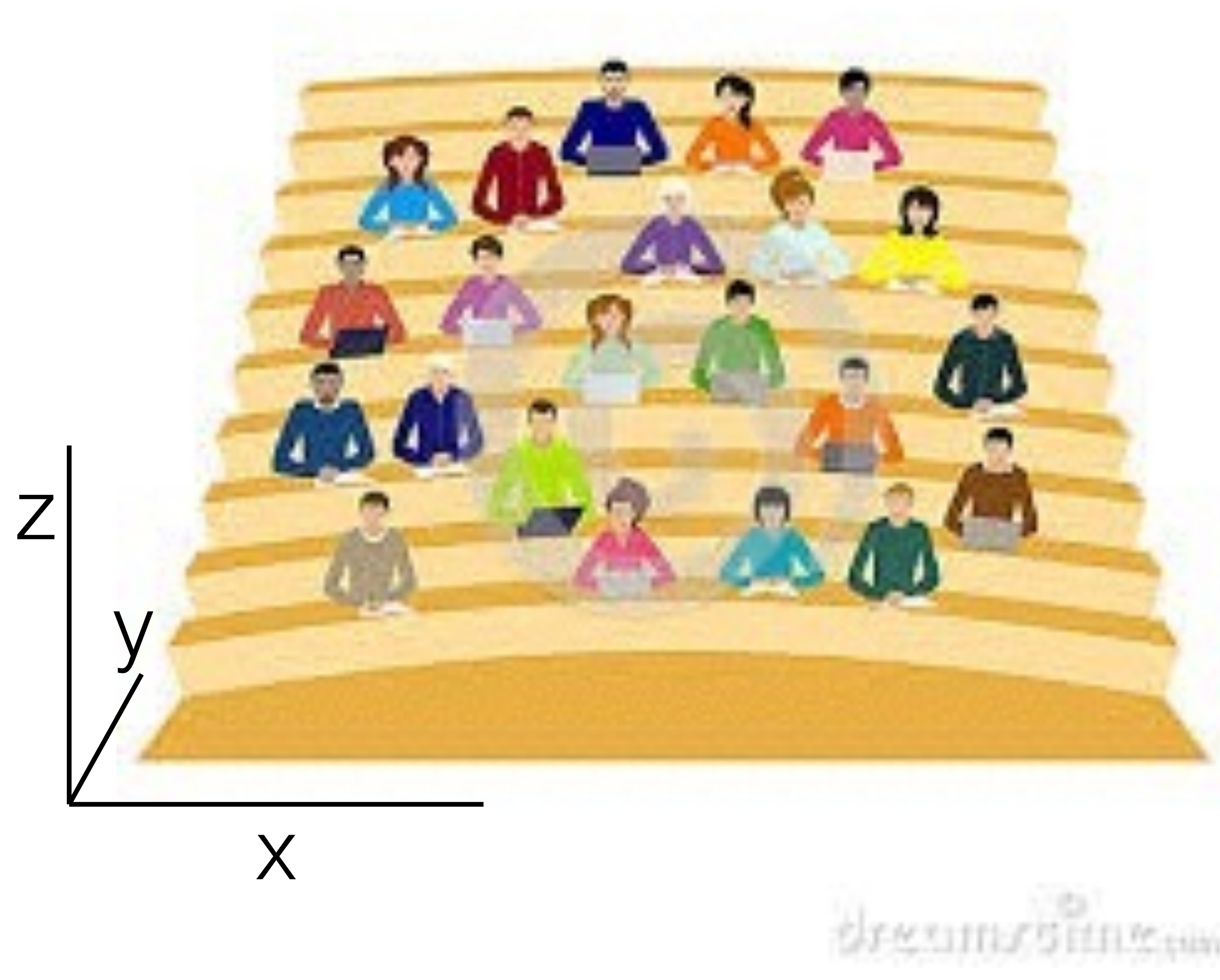


$$\mathbf{y}_i^\top = \mathbf{x}_i^\top W$$

Questions

1. Σ is (an $n \times n$) covariance matrix, how many eigenvectors does it have?
   (a) 1 (b) n (c) As many as underlying dimensionality of data
2. X is an $n \times d$ data matrix such that each of the $j \in [d]$ variance on that coordinate is 1. Which of the following are true
   (a) Covariance matrix is the identity matrix
   (b) Covariance matrix can be any arbitrary symmetric matrix
   (c) All diagonal elements of the covariance matrix are 1
   (d) Off-diagonal elements can have magnitude at most 1
3. We have data matrix X and another matrix X' obtained by rotating X. Consider using PCA (say with $K = 1$).
   (a) $\mathbf{w}_1$ the first component for X and $\mathbf{w}_1'$ for X' are the same
   (b) Y obtained from $PCA(X)$ and Y' obtained from $PCA(X')$ are same
   (c) Both (a) and (b) are true
   (d) Both (a) and (b) are false

Example:
Students in classroom



Review

- Review covariance
- Review Eigen vectors

PCA: Variance Maximization

Covariance matrix:

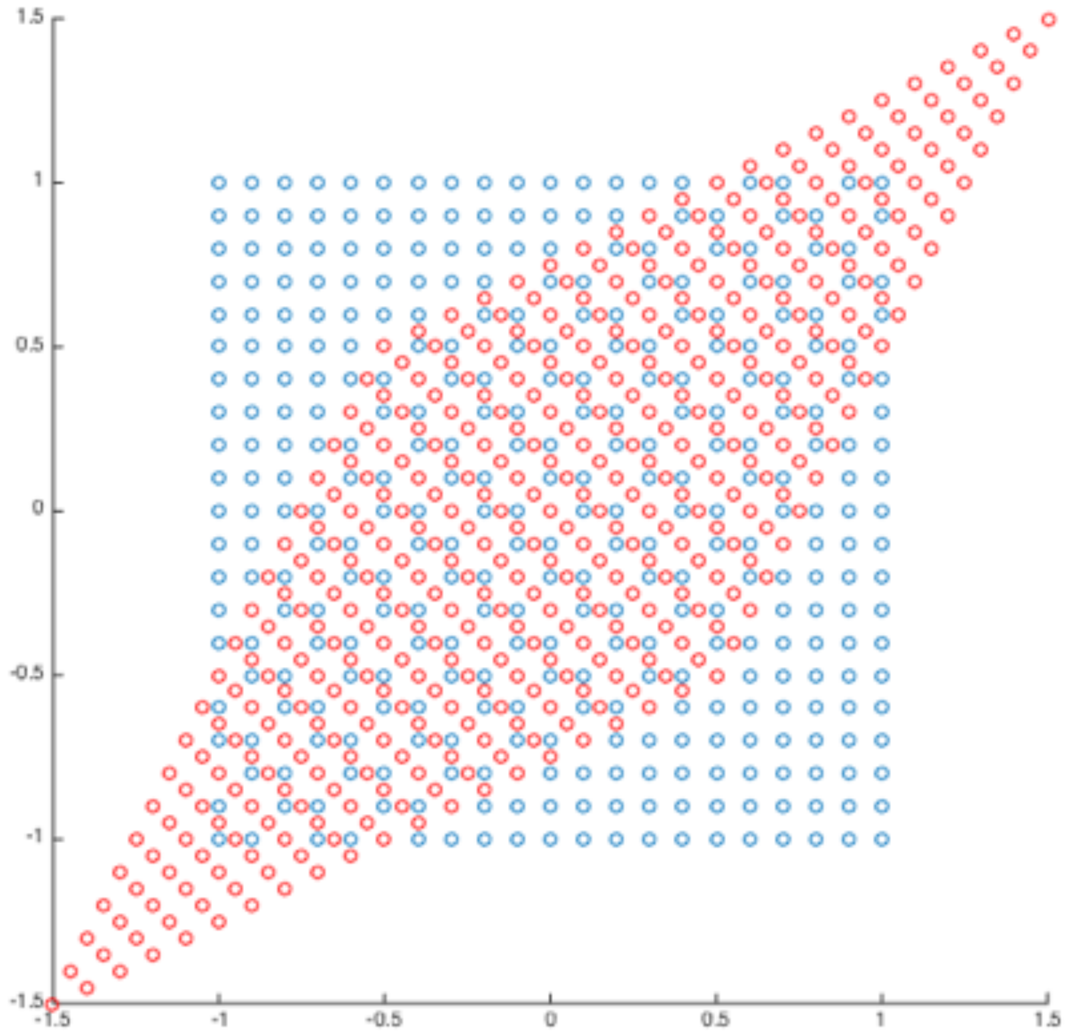
$$\Sigma = \frac{1}{n}\sum_{t=1}^{n}(\mathbf{x}_t - \mu)(\mathbf{x}_t - \mu)^\top$$

- Its a $d \times d$ matrix, $\Sigma[i, j]$ measures "covariance" of features i and j

$$\Sigma[i, j] = \frac{1}{n}\sum_{t=1}^{n}(\mathbf{x}_t[i] - \mu[i])(\mathbf{x}_t[j] - \mu[j])$$

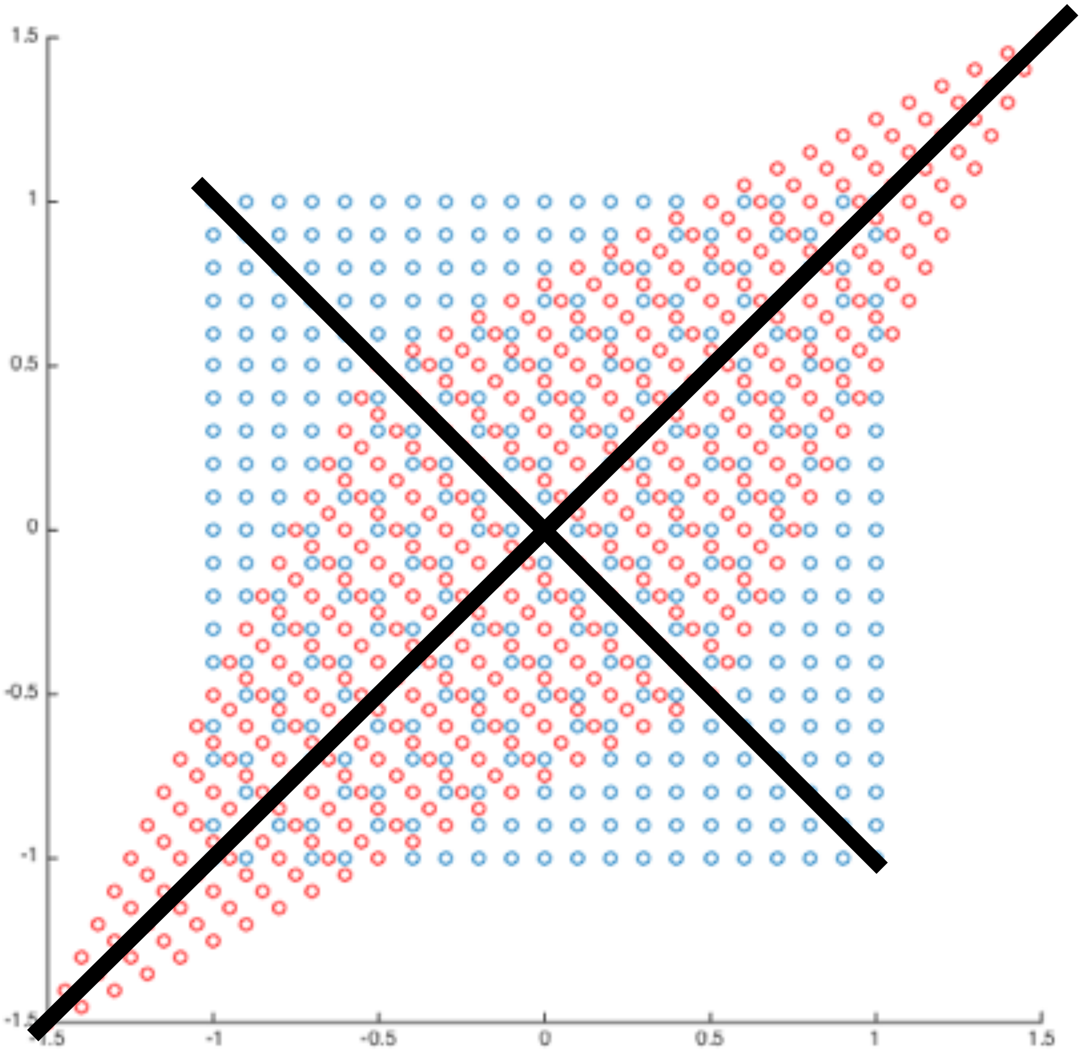
What are Eigen Vectors?

$$x \mapsto Ax$$



What are Eigen Vectors?

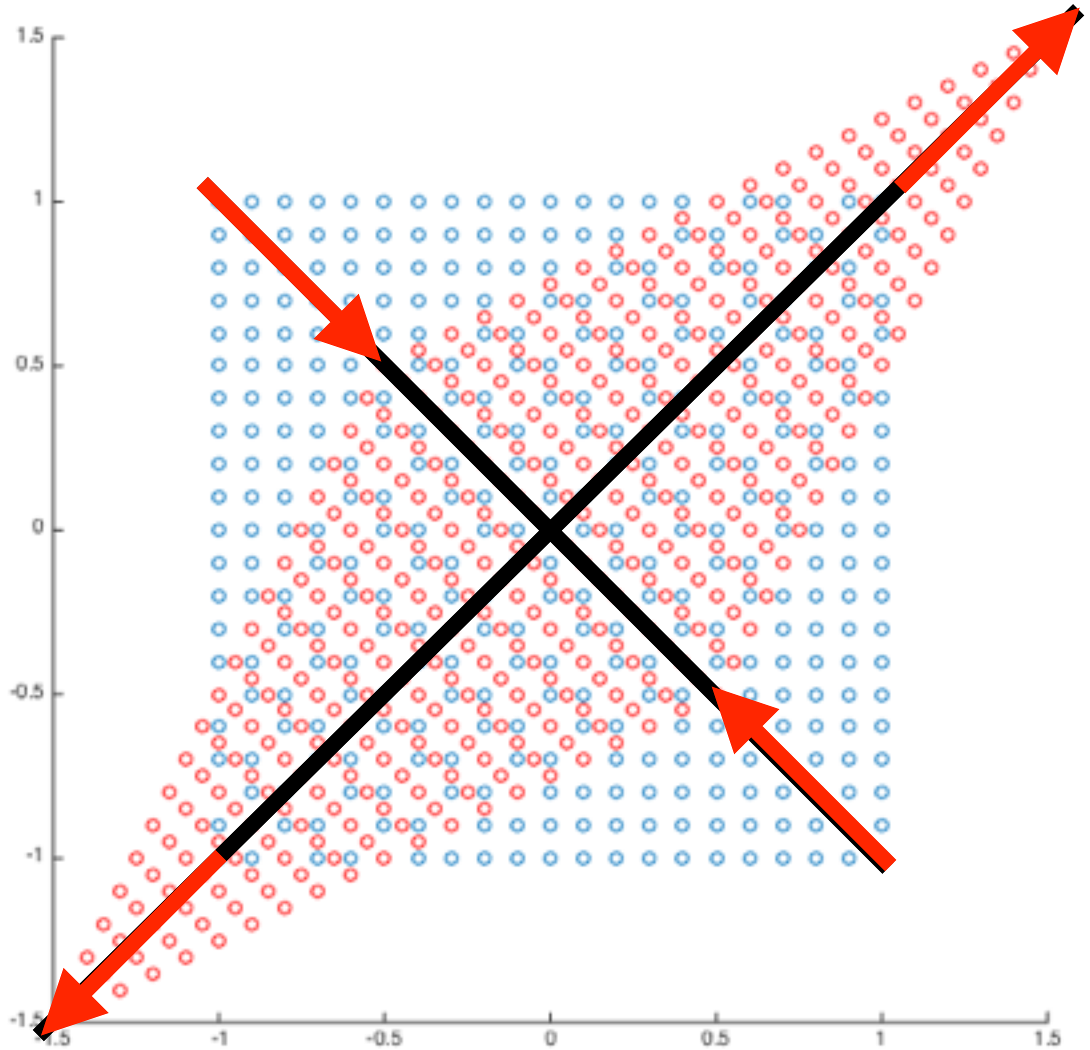
$$x \mapsto Ax$$



$$A\mathbf{x} = \lambda\mathbf{x}$$

What are Eigen Vectors?

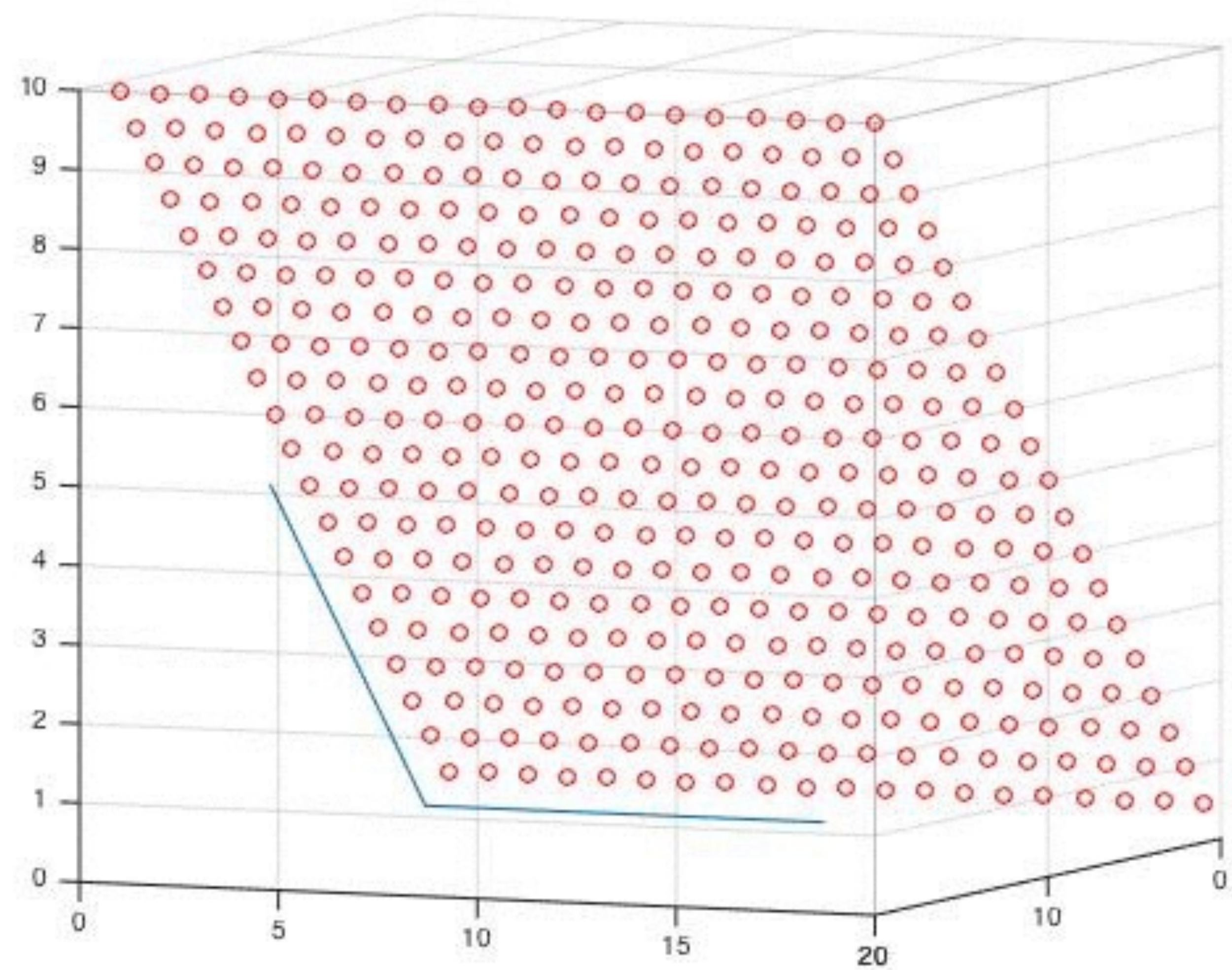
$$x \mapsto Ax$$



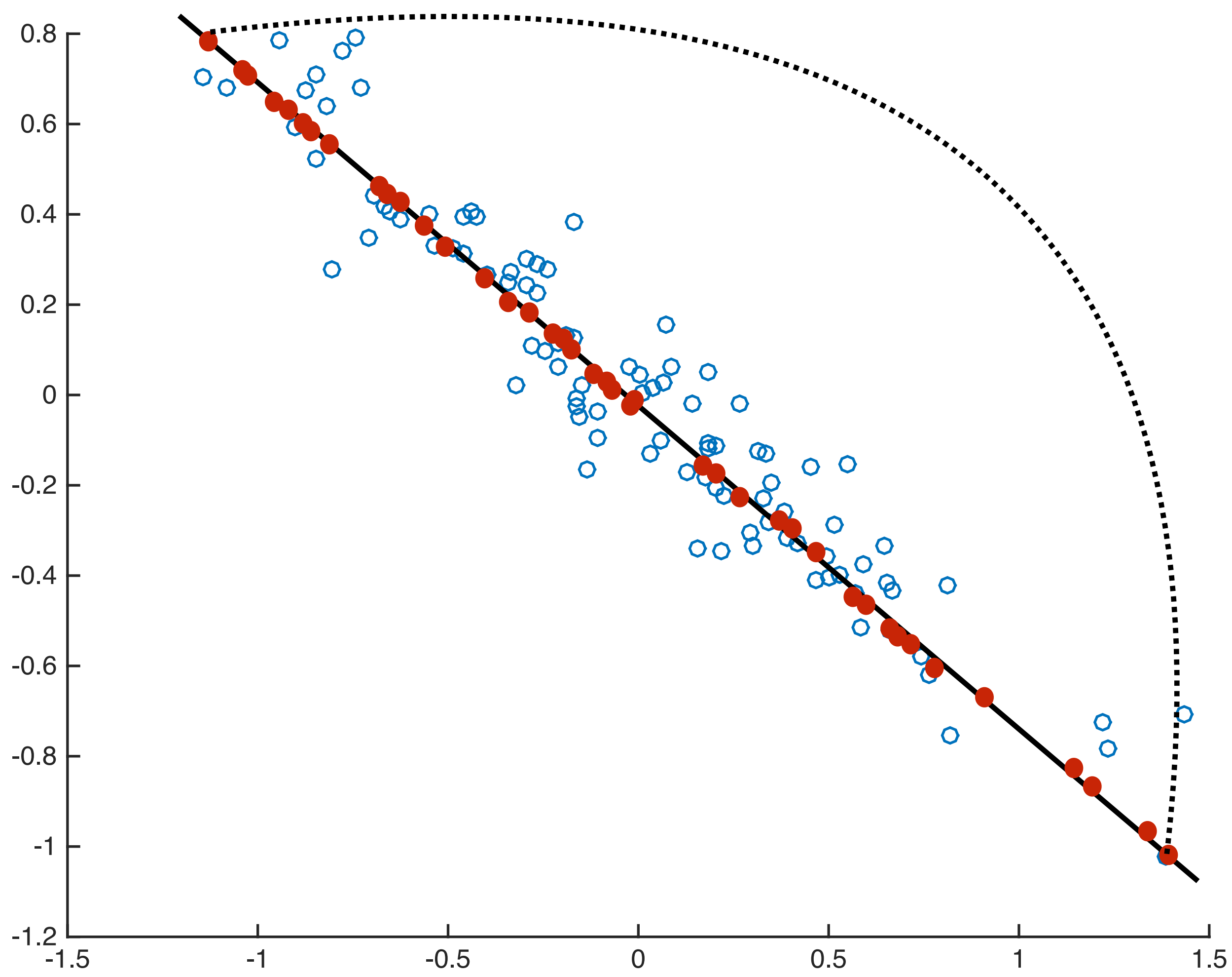
$$A\mathbf{x} = \lambda\mathbf{x}$$

Example:
Students in classroom

PCA on the Example



PCA: Variance Maximization



Work out variance on board

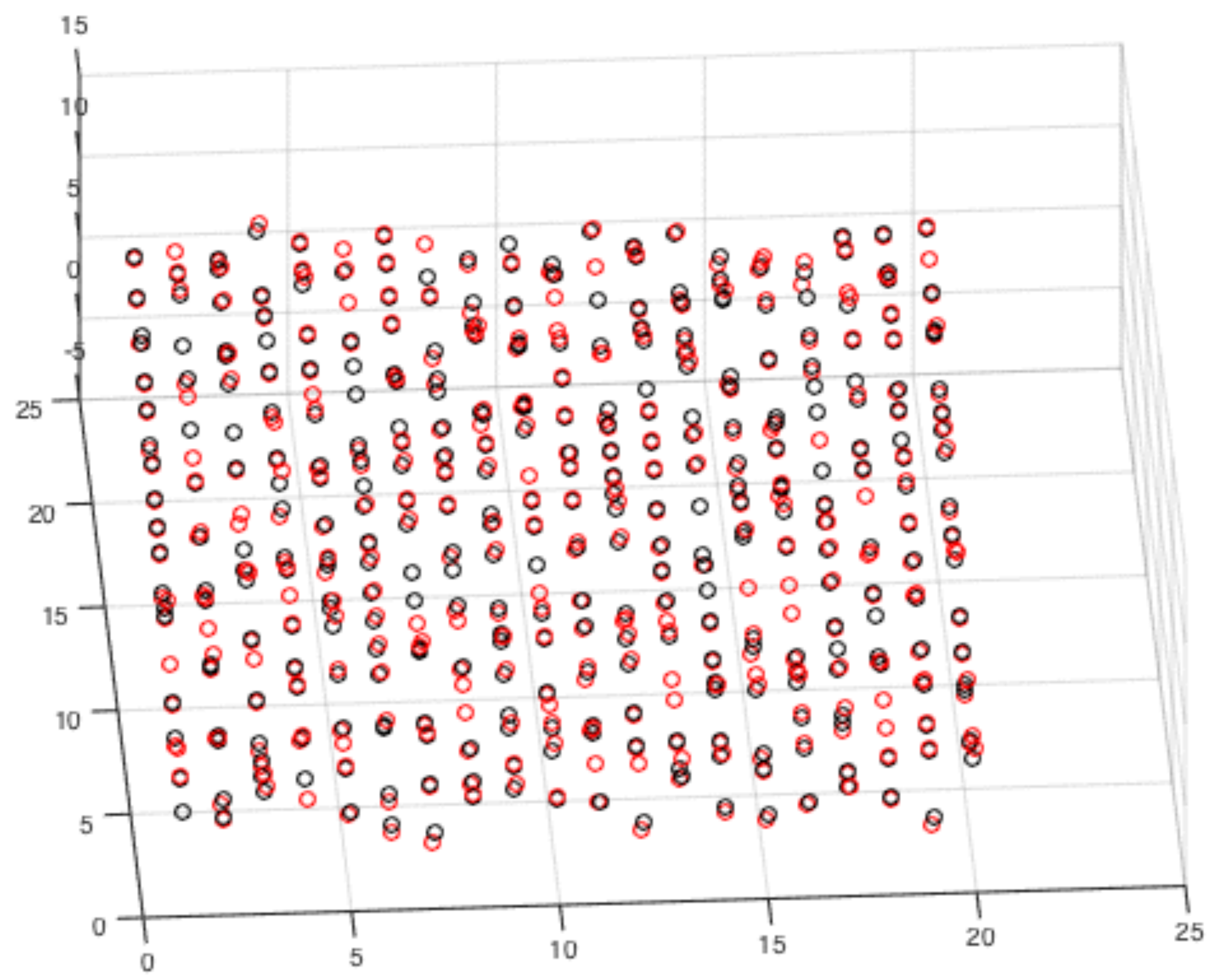
- What if we want more than one number for each data point?

- That is we want to reduce to $K > 1$ dimensions?

PCA: Variance Maximization

- How do we find the K components?

  Ans: Maximize sum of spread in the K directions



PCA: Variance Maximization

- How do we find the K components?
- We are looking for orthogonal directions that maximize total spread in each direction
- Find orthonormal W that maximizes $\sum_{k=1}^{d} \mathbf{w}_i[k]\mathbf{w}_j[k] = 0 \ \ \& \ \ \sum_{k=1}^{d} \mathbf{w}_i[k] = 1$

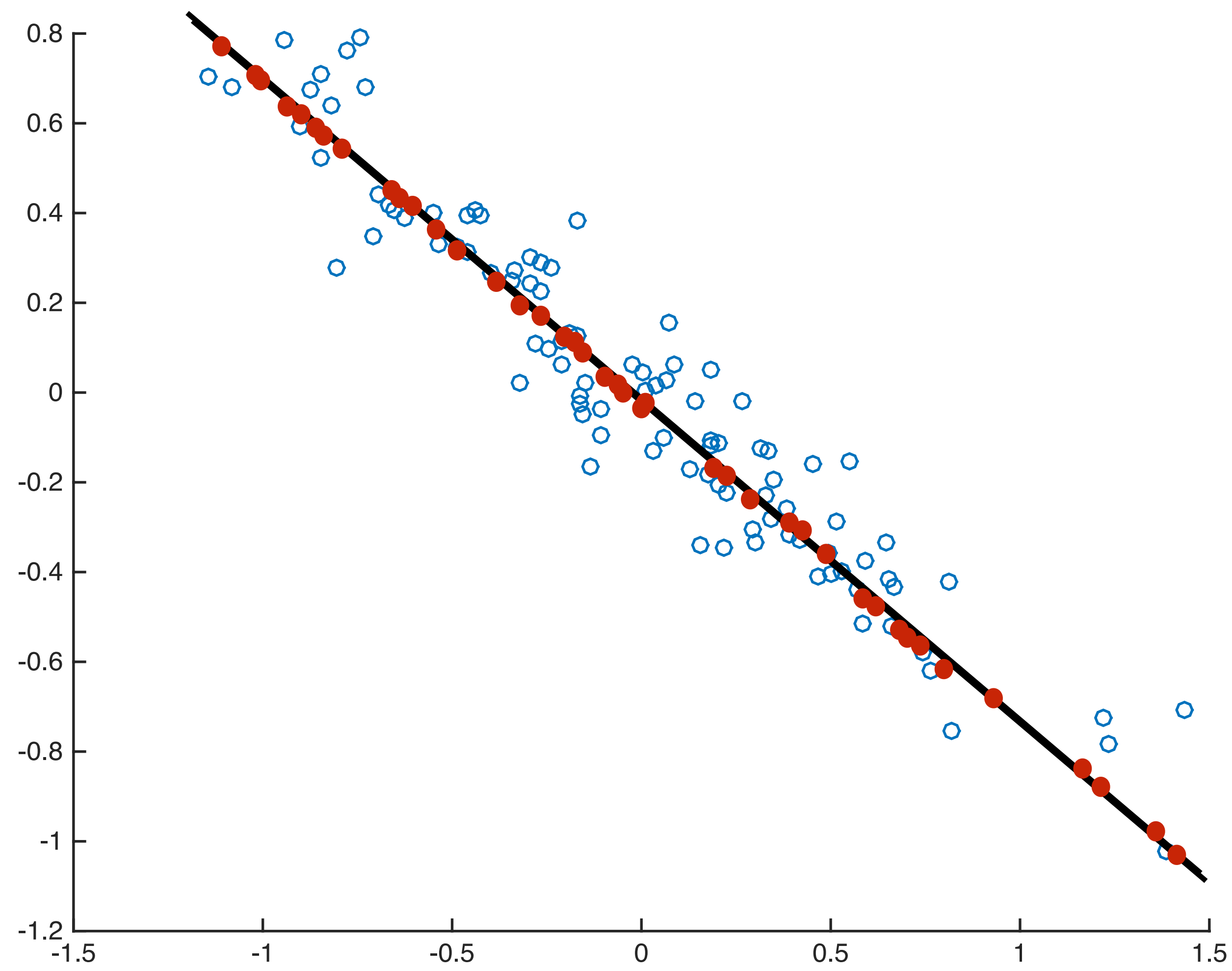
$$\sum_{j=1}^{K} \frac{1}{n} \sum_{t=1}^{n} \left(\mathbf{y}_t[j] - \frac{1}{n} \sum_{t=1}^{n} \mathbf{y}_t[j] \right)^2 = \sum_{j=1}^{K} \frac{1}{n} \sum_{t=1}^{n} \left(\mathbf{w}_j^\top \left(\mathbf{x}_t - \frac{1}{n} \sum_{t=1}^{n} \mathbf{x}_t \right) \right)^2$$

$$= \sum_{j=1}^{K} \mathbf{w}_j^\top \Sigma \mathbf{w}_j$$

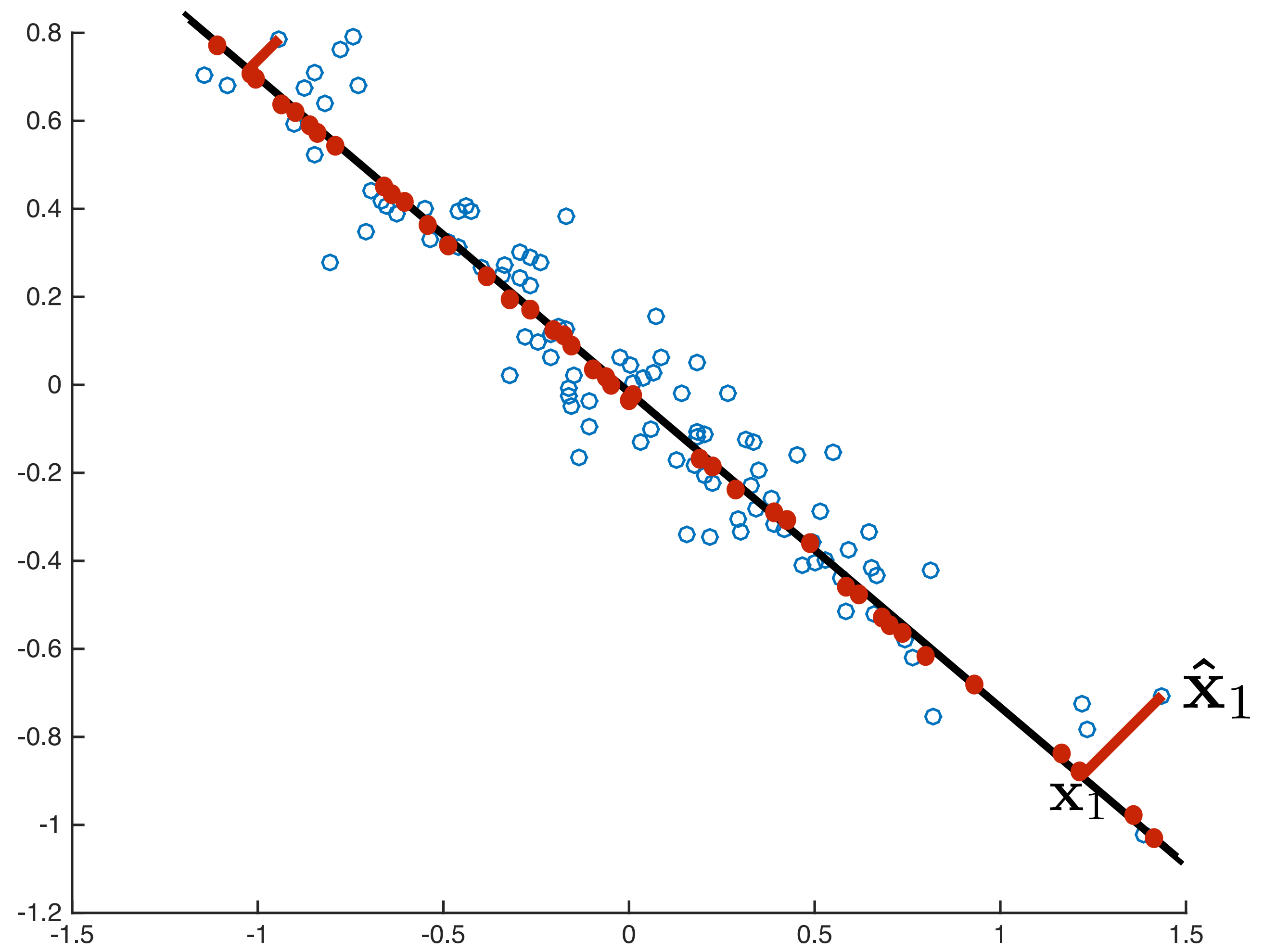
- This solutions is given by W = Top K eigenvectors of Σ

An Alternative View

PCA: Minimizing Reconstruction Error

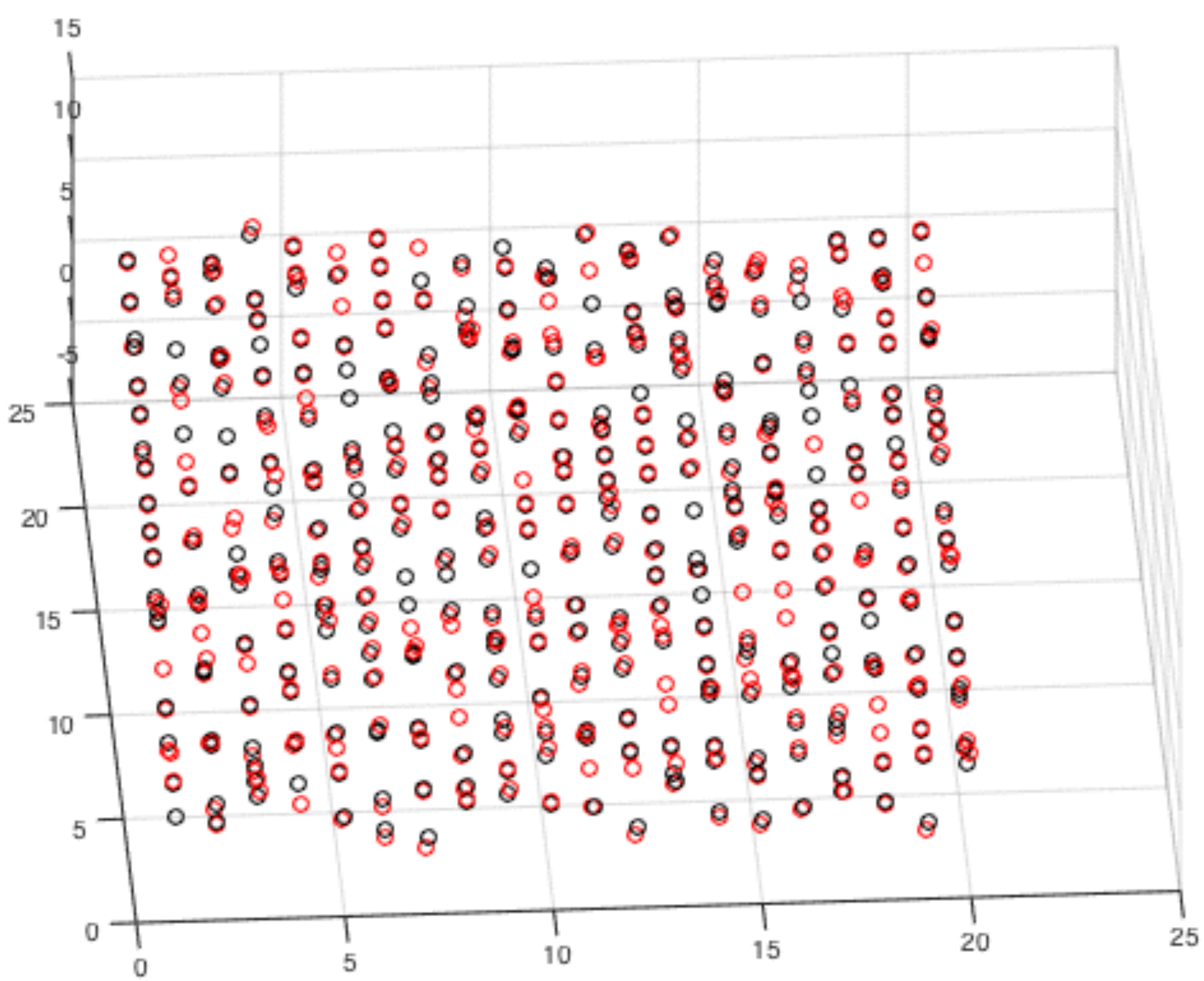


PCA: Minimizing Reconstruction Error

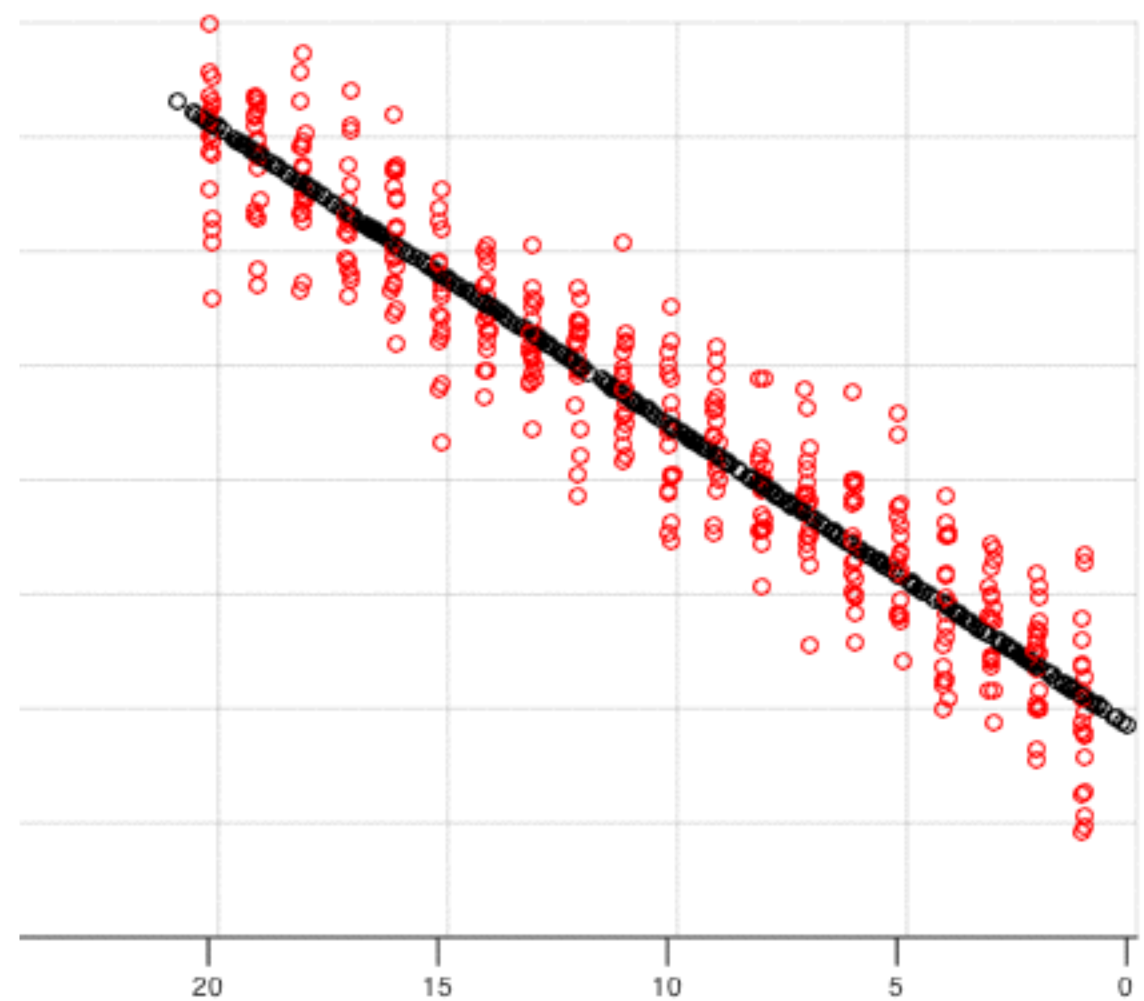


$$\sum_{t=1}^{n} \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|^2$$

Maximize Spread



Minimize Reconstruction Error



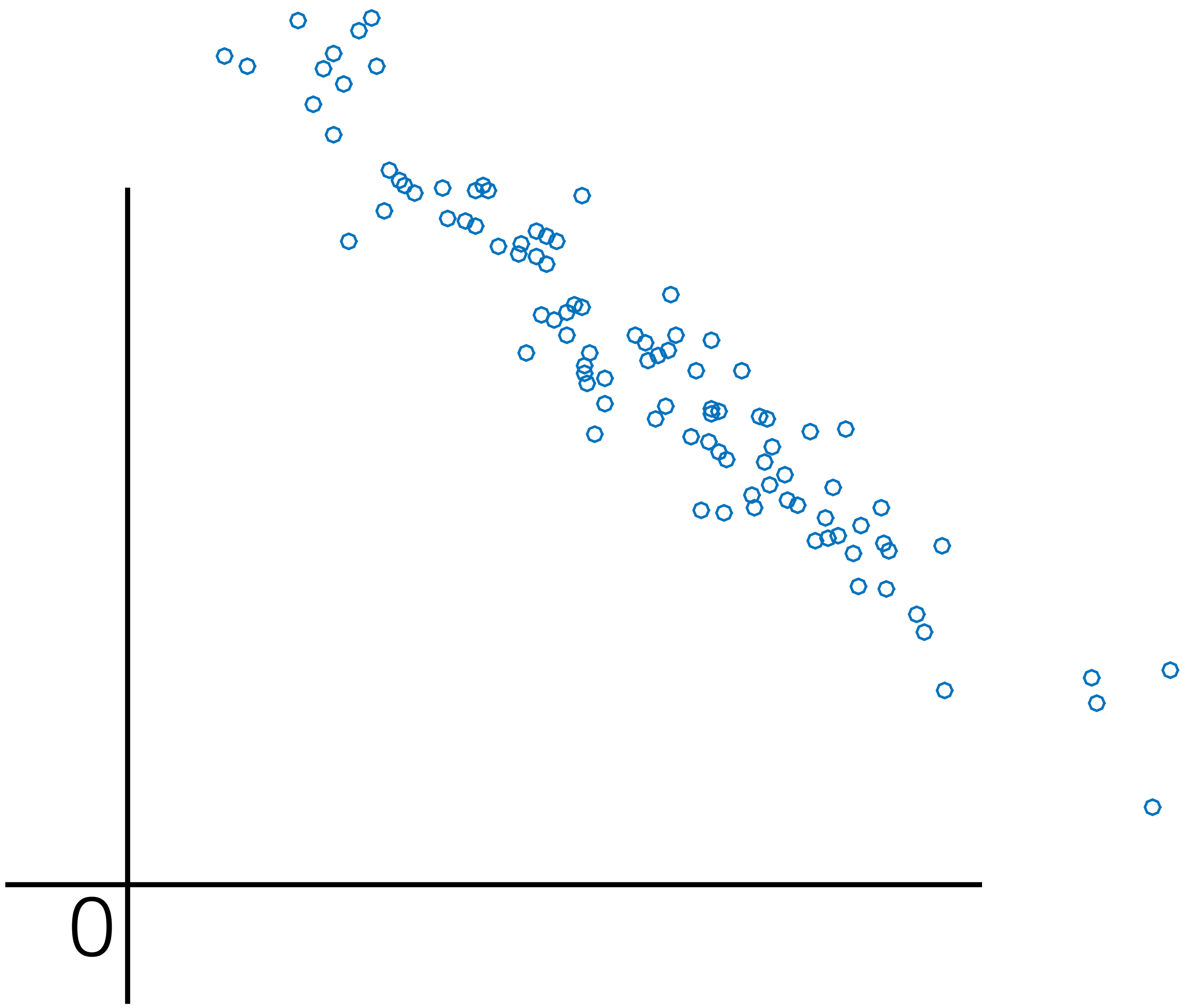
Orthonormal Projections

- Think of $\mathbf{w}_1, \ldots, \mathbf{w}_K$ as coordinate system for PCA (in a K dimensional subspace)
- $\mathbf{y}$ values provide coefficients in this system
- Without loss of generality, $\mathbf{w}_1, \ldots, \mathbf{w}_K$ can be orthonormal, i.e. $\mathbf{w}_i \perp \mathbf{w}_j$ & $\|\mathbf{w}_i\| = 1$.

$$\|\mathbf{w}_i\|_2^2 = \sum_{k=1}^{d} \mathbf{w}_i[k]^2$$

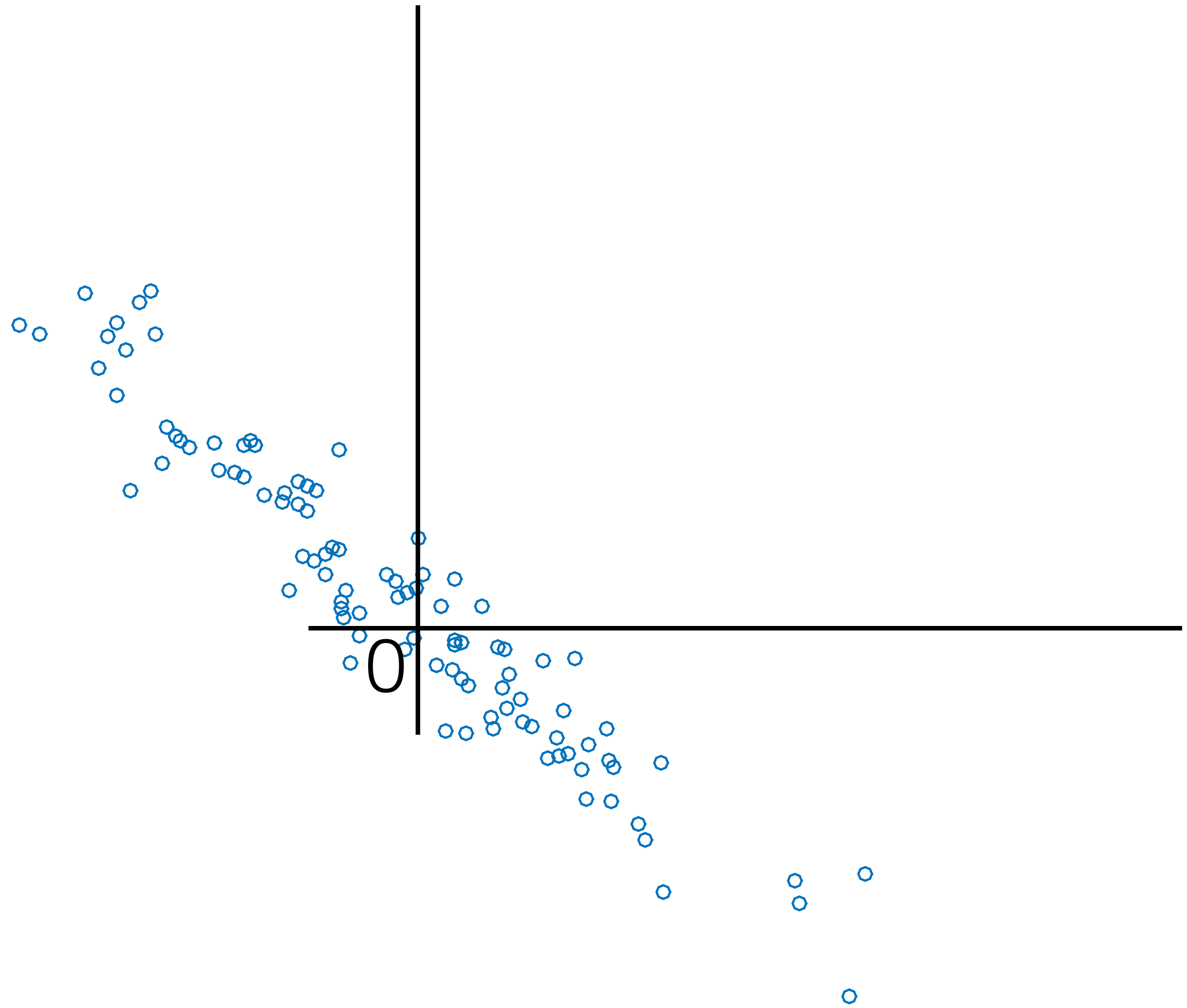
$$\mathbf{w}_i \perp \mathbf{w}_j \Rightarrow \sum_{k=1}^{d} \mathbf{w}_i[k]\mathbf{w}_j[k] = 0$$

Centering Data



Compressing these data points…

Centering Data



… is same as compressing these.

Orthonormal Projections

- (Centered) Data-points as linear combination of some orthonormal basis, i.e.

$$\mathbf{x}_t = \mu + \sum_{j=1}^{d} \mathbf{y}_t[j]\mathbf{w}_j$$

where $\mathbf{w}_1, \ldots, \mathbf{w}_d \in \mathbb{R}^d$ are the orthonormal basis and $\mu = \frac{1}{n}\sum_{t=1}^{n} x_t$.

- Represent data as linear combination of just K orthonormal basis,

$$\hat{\mathbf{x}}_t = \mu + \sum_{j=1}^{K} \mathbf{y}_t[j]\mathbf{w}_j$$

PCA: Minimizing Reconstruction Error

- Goal: find the basis that minimizes reconstruction error,

$$
\begin{aligned}
\sum_{t=1}^{n} \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|_2^2 &= \sum_{t=1}^{n} \left\| \sum_{j=1}^{K} \mathbf{y}_t[j]\mathbf{w}_j + \mu - \mathbf{x}_t \right\|_2^2 \\
&= \sum_{t=1}^{n} \left\| \sum_{j=1}^{K} \mathbf{y}_t[j]\mathbf{w}_j + \mu - \sum_{j=1}^{d} \mathbf{y}_t[j]\mathbf{w}_j - \mu \right\|_2^2 \\
&= \sum_{t=1}^{n} \left\| \sum_{j=K+1}^{d} \mathbf{y}_t[j]\mathbf{w}_j \right\|_2^2 \quad (\text{but } \|a+b\|_2^2 = \|a\|_2^2 + \|b\|_2^2 + 2a^\top b) \\
&= \sum_{t=1}^{n} \left(\sum_{j=K+1}^{d} \mathbf{y}_t[j]^2 \|\mathbf{w}_j\|_2^2 + 2 \sum_{j=K+1}^{d} \sum_{i=j+1}^{d} \mathbf{y}_t[j]\mathbf{y}_t[i]\mathbf{w}_j^\top \mathbf{w}_i \right) \\
&= \sum_{t=1}^{n} \sum_{j=k+1}^{d} \mathbf{y}_t[j]^2 \|\mathbf{w}_j\|_2^2 \quad (\text{last step because } \mathbf{w}_j \perp \mathbf{w}_i)
\end{aligned}
$$

PCA: Minimizing Reconstruction Error

$$
\begin{aligned}
\frac{1}{n}\sum_{t=1}^{n}\|\hat{\mathbf{x}}_t - \mathbf{x}_t\|_2^2 &= \frac{1}{n}\sum_{t=1}^{n}\sum_{j=k+1}^{d}\mathbf{y}_t[j]^2\|\mathbf{w}_j\|_2^2 \qquad (\text{but } \|\mathbf{w}_j\| = 1) \\
&= \frac{1}{n}\sum_{t=1}^{n}\sum_{j=k+1}^{d}\mathbf{y}_t[j]^2 \qquad (\text{now } \mathbf{y}_j = \mathbf{w}_j^\top(\mathbf{x}_t - \mu)) \\
&= \frac{1}{n}\sum_{t=1}^{n}\sum_{j=k+1}^{d}\left(\mathbf{w}_j^\top(\mathbf{x}_t - \mu)\right)^2 \\
&= \frac{1}{n}\sum_{t=1}^{n}\sum_{j=k+1}^{d}\mathbf{w}_j^\top(\mathbf{x}_t - \mu)(\mathbf{x}_t - \mu)^\top\mathbf{w}_j \\
&= \sum_{j=k+1}^{d}\mathbf{w}_j^\top\Sigma\mathbf{w}_j
\end{aligned}
$$

PCA: Minimizing Reconstruction Error

Minimize w.r.t. $\mathbf{w}_1, \ldots, \mathbf{w}_K$'s that are orthonormal,

$$\underset{\forall j,\ \|\mathbf{w}_j\|_2=1}{\operatorname{argmin}} \sum_{j=k+1}^{d} \mathbf{w}_j^\top \Sigma \mathbf{w}_j$$

- Solution, (discard) $\mathbf{w}_{K+1}, \ldots, \mathbf{w}_d$ are bottom $d - K$ eigenvectors
- Hence $\mathbf{w}_1, \ldots, \mathbf{w}_K$ are the top K eigenvectors

Principal Component Analysis

1. $$\Sigma = \mathrm{cov}\left(X \right)$$

2. $$W = \mathrm{eigs}\left(\Sigma, K \right)$$

3. $$Y = X - \mu \times W$$

Reconstruction

4. $$\hat{X} = Y \times W^{\top} + \mu$$

When $d >> n$

- If $d >> n$ then Σ is large
- But we only need top K eigen vectors.
- Idea: use SVD

$$X - \mu = UDV^\top$$

Then note that, $\Sigma = (X - \mu)^\top (X - \mu) = VD^2V$

- Hence, matrix V is the same as matrix W got from eigen decomposition of Σ, eigenvalues are diagonal elements of D^2
- Alternative algorithm:

$$[U, V] = \mathrm{SVD}(X - \mu, K) \quad W = V$$

Principal Component Analysis: Demo