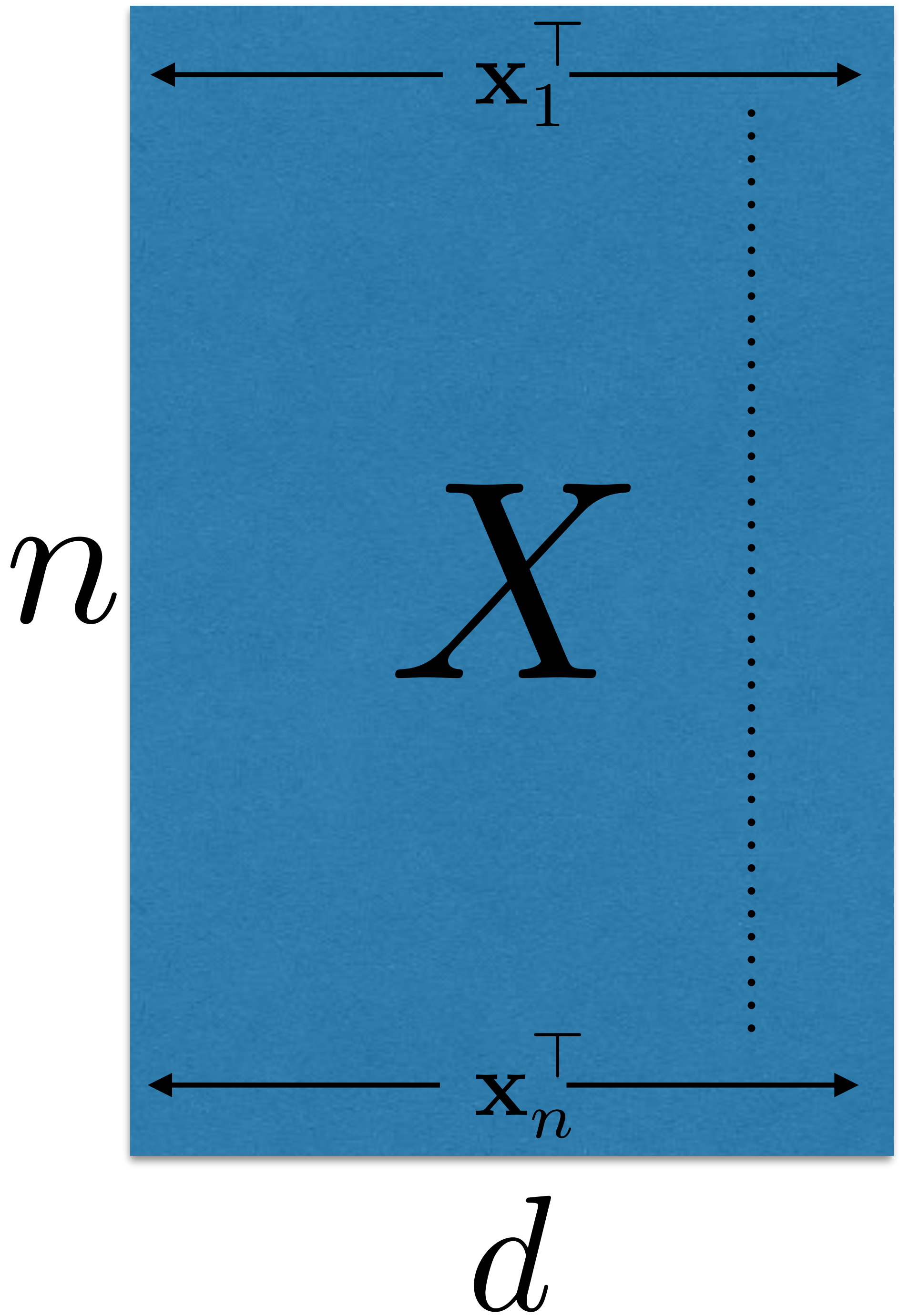
Machine Learning for Data Science (CS4786)
Lecture 3

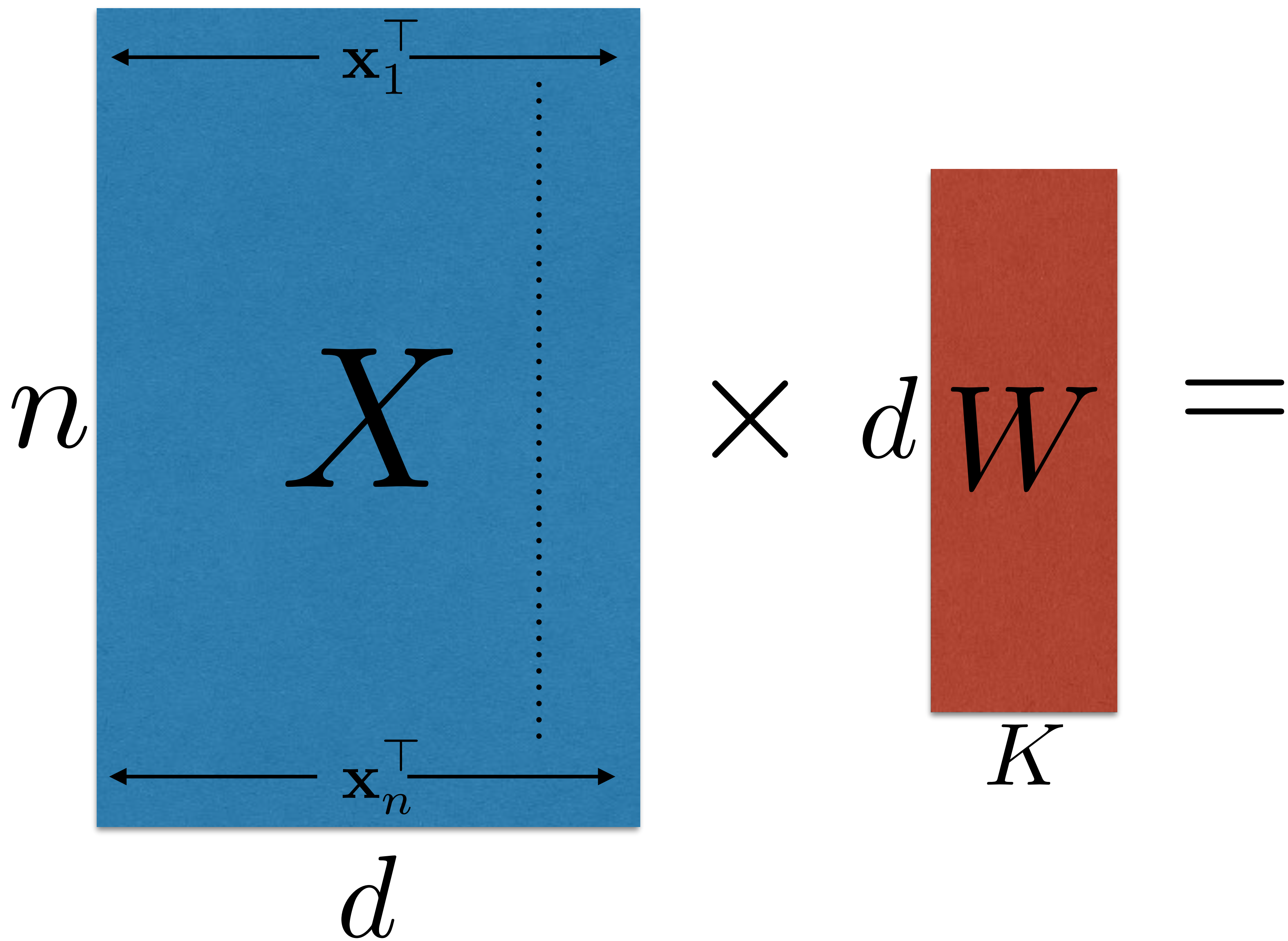
Principal Component Analysis

Dim Reduction: Linear Transformation

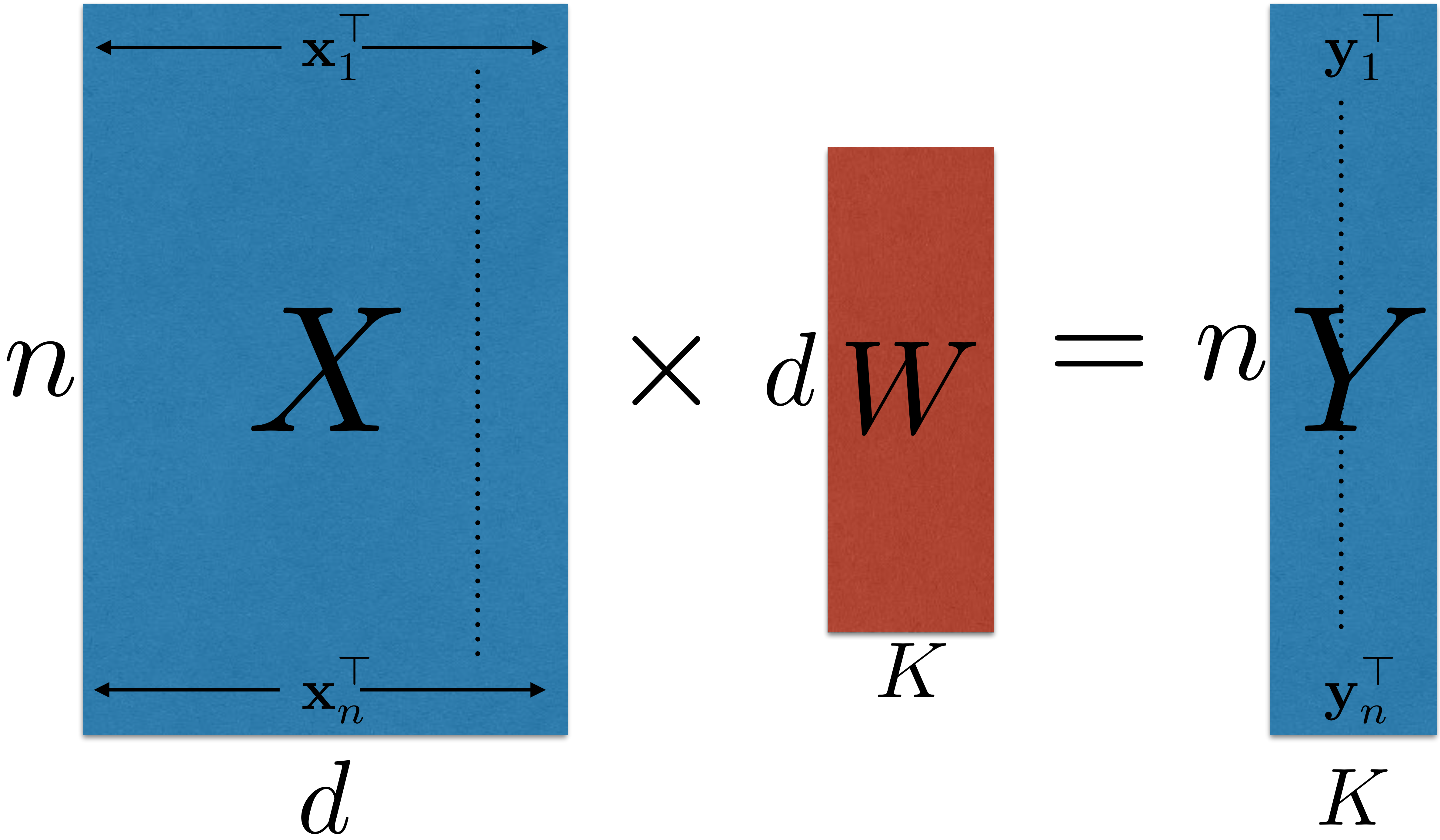
Dim Reduction: Linear Transformation



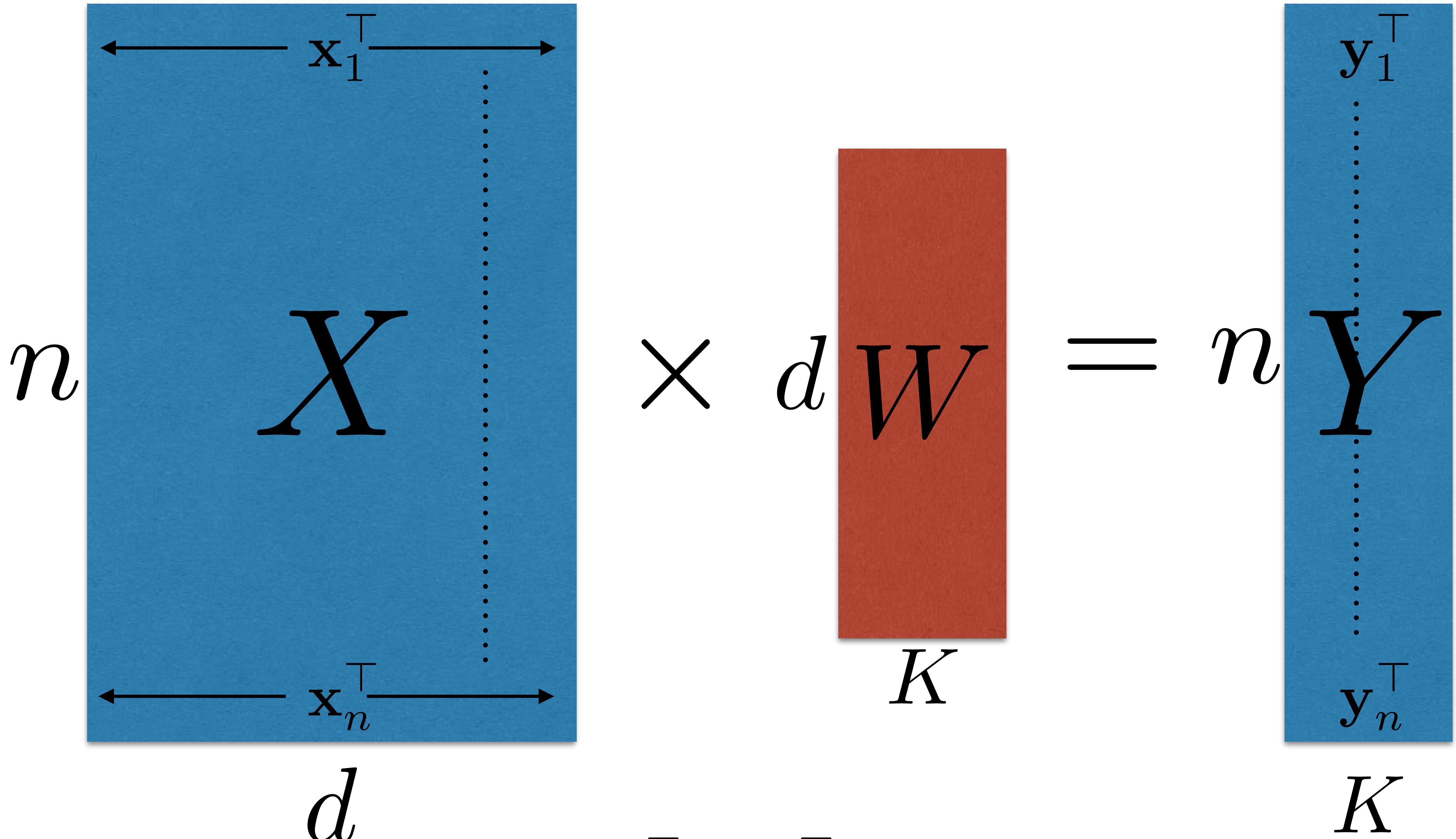
DIM REDUCTION: LINEAR TRANSFORMATION



Dim Reduction: Linear Transformation



Dim Reduction: Linear Transformation

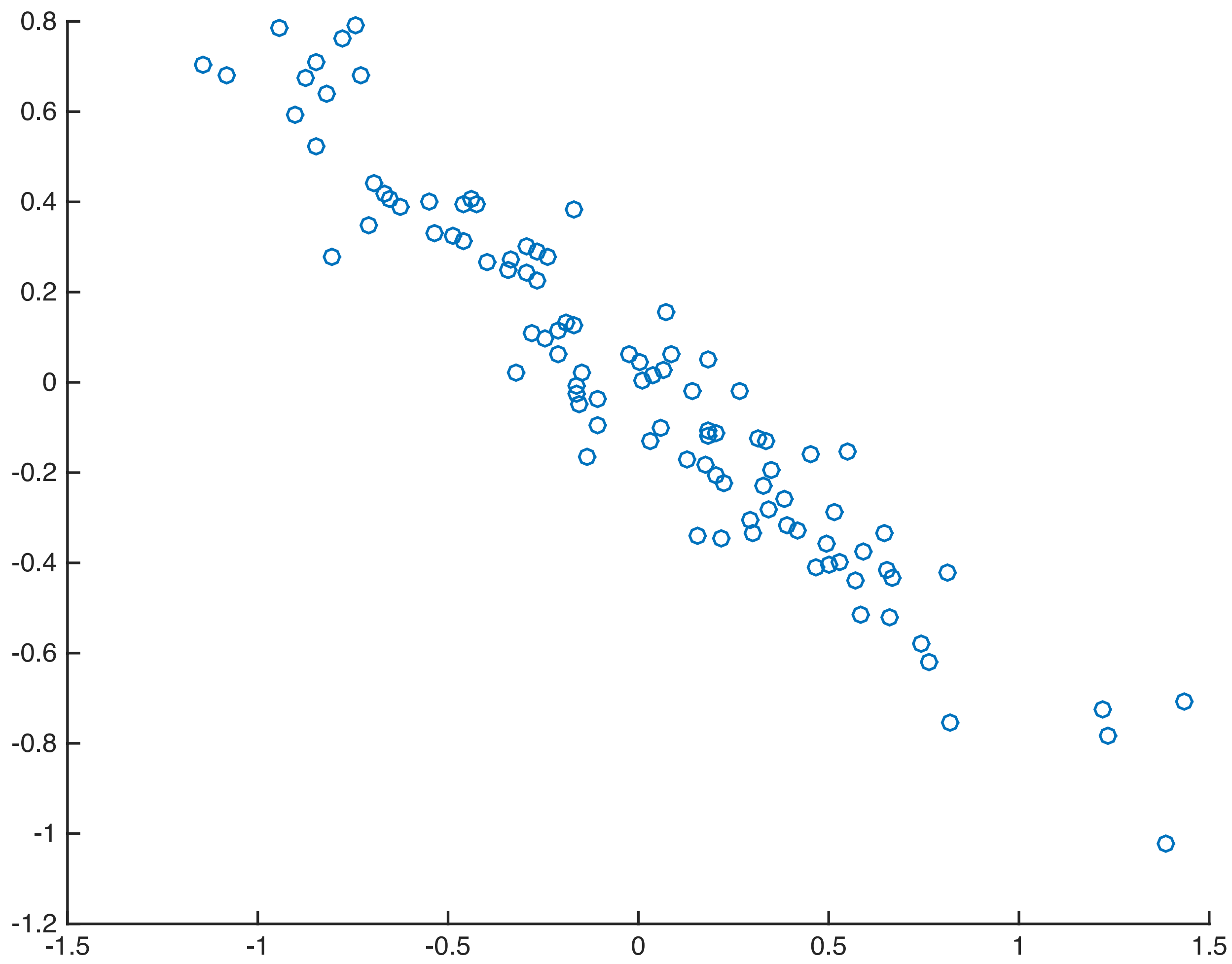


$$\mathbf{y}_i^\top = \mathbf{x}_i^\top W$$

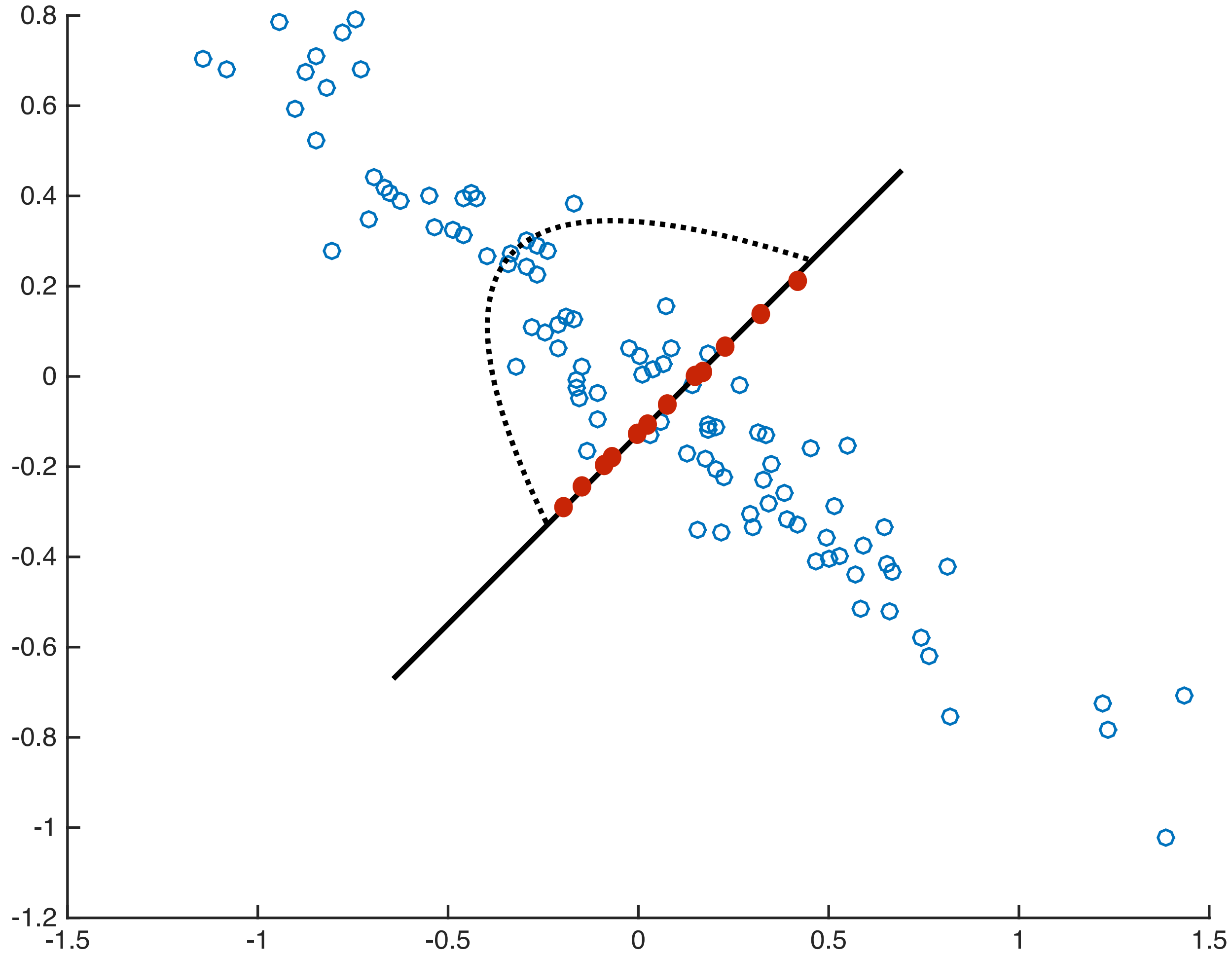
Prelude: Reducing to 1 Dim

- W is a d x 1 matrix (d dimensional vector)
- Each data point is compressed to a single number
- How do we pick this W?

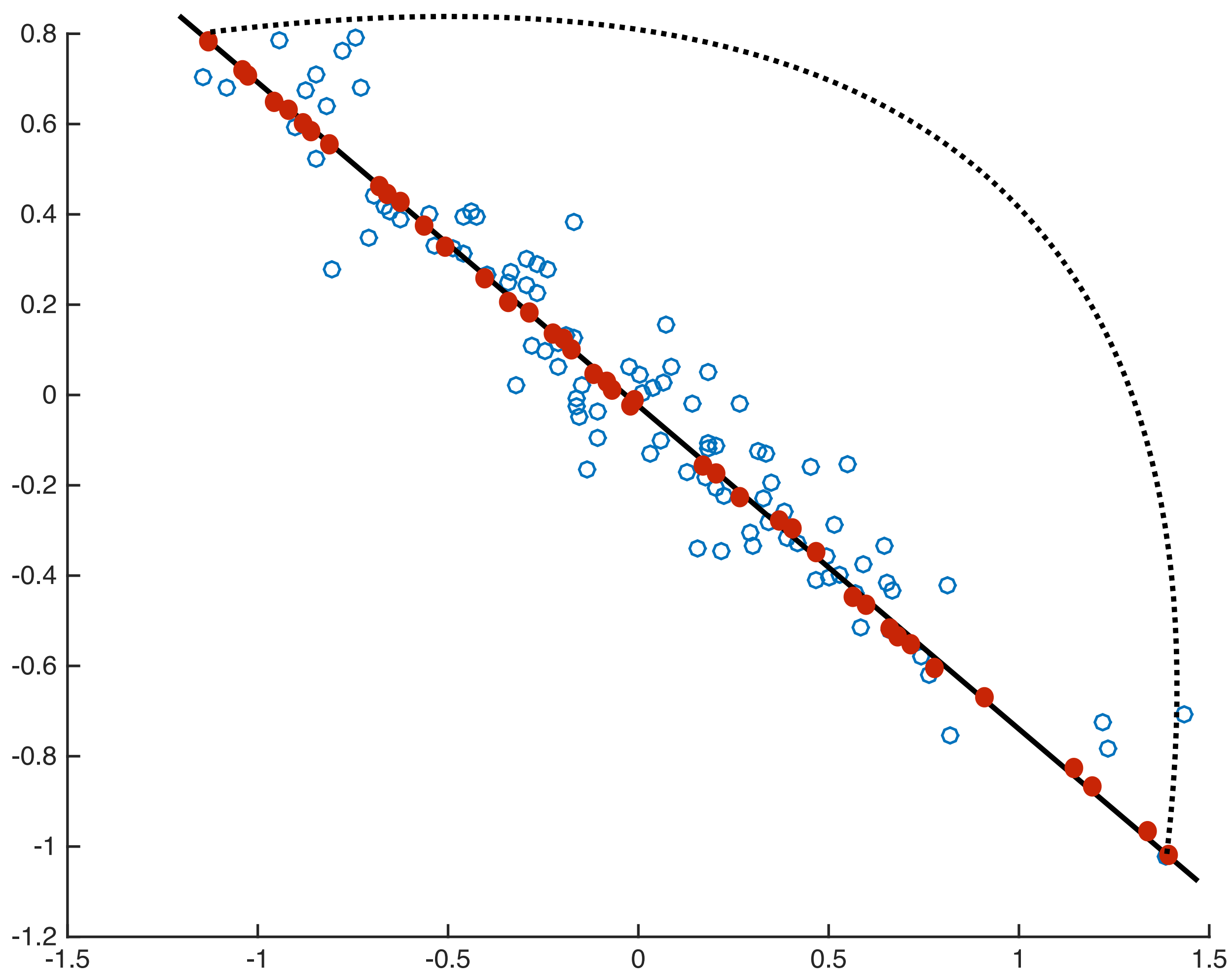
PCA: Variance Maximization



PCA: Variance Maximization



PCA: Variance Maximization



PCA: Variance Maximization

- Pick directions along which data varies the most

PCA: Variance Maximization

- Pick directions along which data varies the most

$$\text{Variance} = \frac{1}{n}\sum_{t=1}^{n}\left(y_t - \frac{1}{n}\sum_{s=1}^{n} y_s\right)^2$$

PCA: Variance Maximization

- Pick directions along which data varies the most

$$\text{Variance} = \frac{1}{n}\sum_{t=1}^{n}\left(y_t - \frac{1}{n}\sum_{s=1}^{n} y_s\right)^2$$

$$= \frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{w}^\top \mathbf{x}_t - \frac{1}{n}\sum_{s=1}^{n} \mathbf{w}^\top \mathbf{x}_s\right)^2$$

PCA: Variance Maximization

- Pick directions along which data varies the most

$$
\begin{aligned}
\text{Variance} &= \frac{1}{n}\sum_{t=1}^{n}\left(y_t - \frac{1}{n}\sum_{s=1}^{n} y_s\right)^2 \\
&= \frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{w}^\top \mathbf{x}_t - \frac{1}{n}\sum_{s=1}^{n} \mathbf{w}^\top \mathbf{x}_s\right)^2 \\
&= \frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{w}^\top \mathbf{x}_t - \mathbf{w}^\top\left(\frac{1}{n}\sum_{s=1}^{n} \mathbf{x}_s\right)\right)^2
\end{aligned}
$$

PCA: Variance Maximization

- Pick directions along which data varies the most

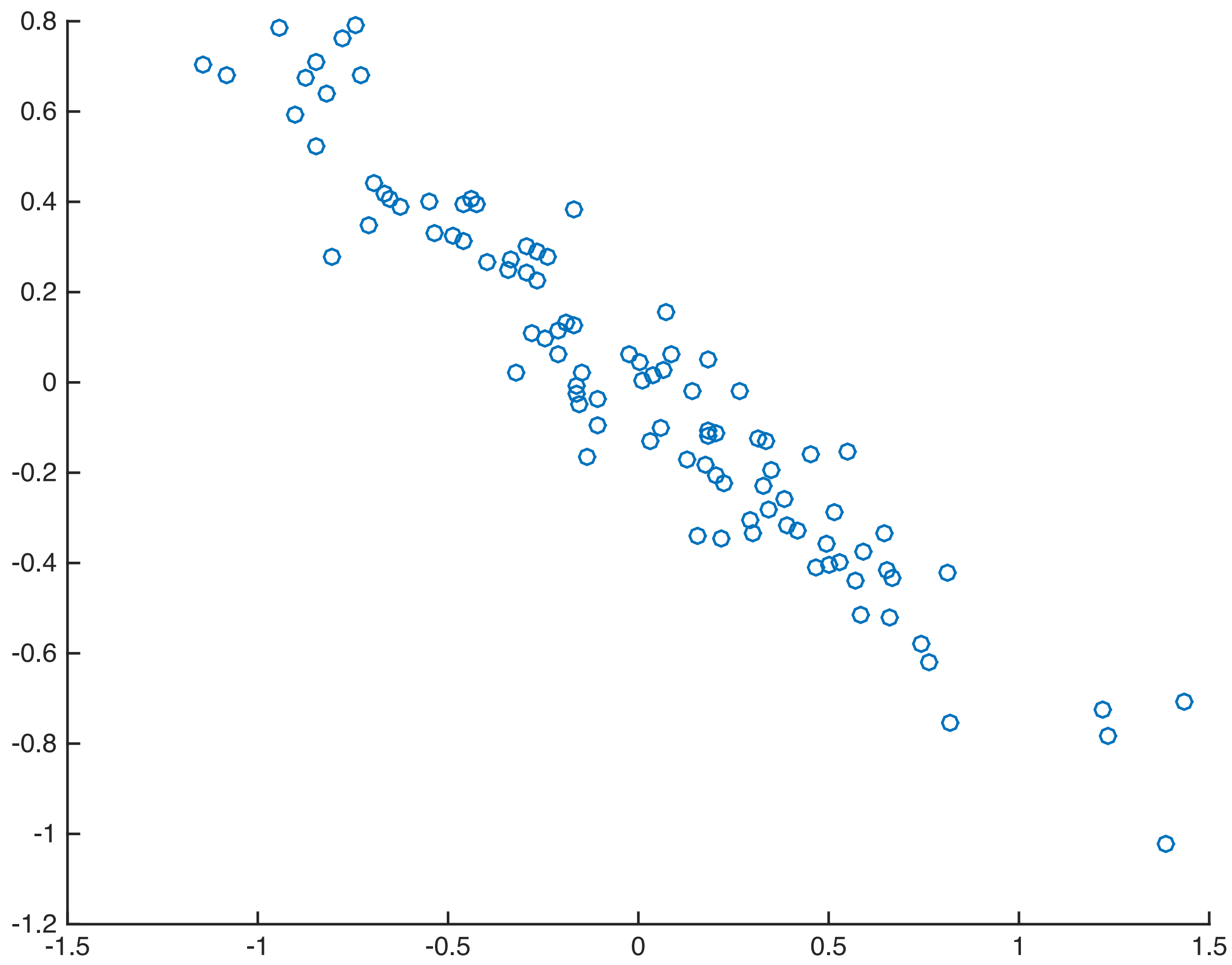
$$
\begin{aligned}
\text{Variance} &= \frac{1}{n}\sum_{t=1}^{n}\left(y_t - \frac{1}{n}\sum_{s=1}^{n} y_s\right)^2 \\
&= \frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{w}^\top \mathbf{x}_t - \frac{1}{n}\sum_{s=1}^{n} \mathbf{w}^\top \mathbf{x}_s\right)^2 \\
&= \frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{w}^\top \mathbf{x}_t - \mathbf{w}^\top \left(\frac{1}{n}\sum_{s=1}^{n} \mathbf{x}_s\right)\right)^2 \\
&= \frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{w}^\top (\mathbf{x}_t - \mu)\right)^2
\end{aligned}
$$

PCA: Variance Maximization

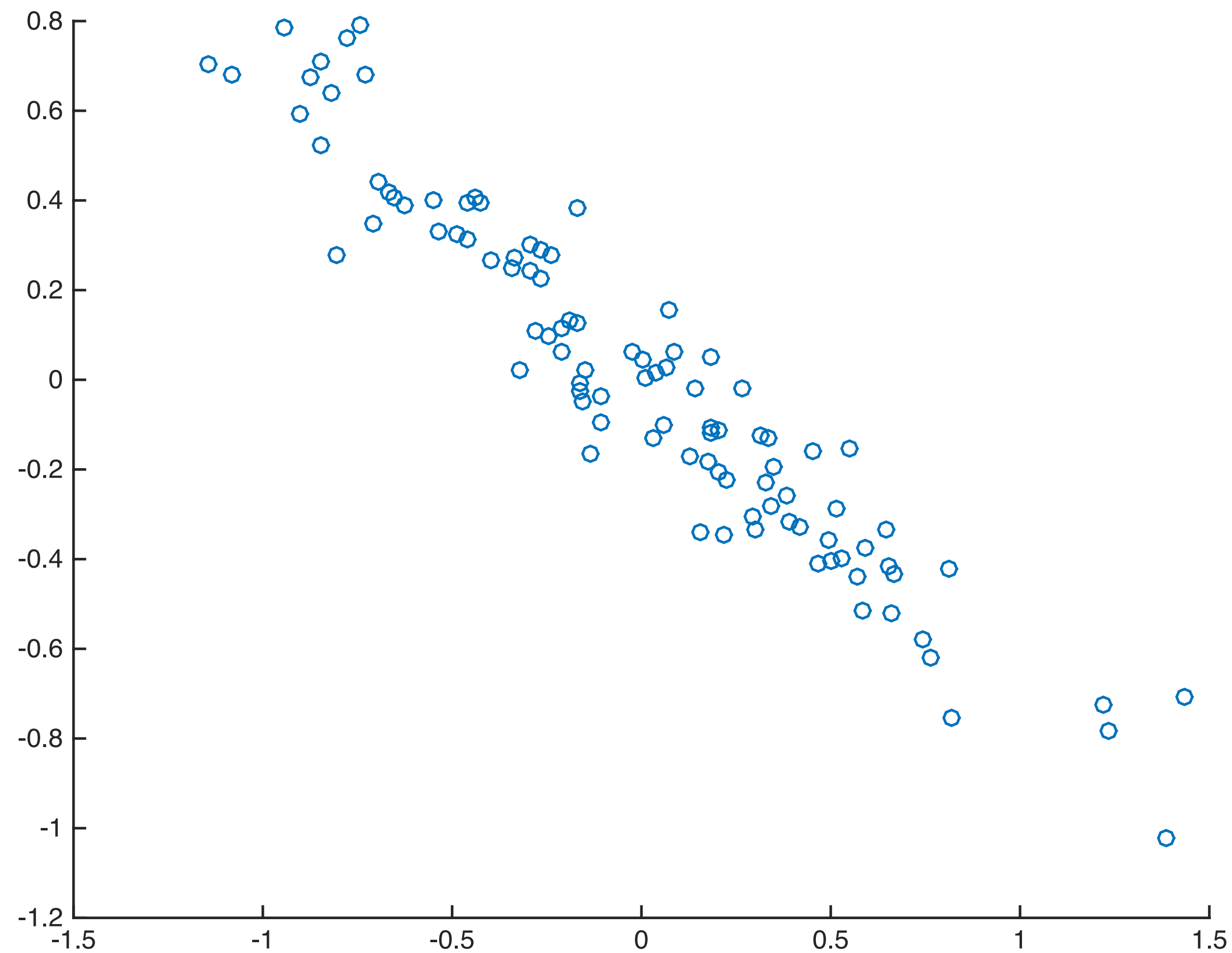
- Pick directions along which data varies the most

$$
\begin{aligned}
\text{Variance} &= \frac{1}{n}\sum_{t=1}^{n}\left(y_t - \frac{1}{n}\sum_{s=1}^{n} y_s\right)^2 \\
&= \frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{w}^\top \mathbf{x}_t - \frac{1}{n}\sum_{s=1}^{n} \mathbf{w}^\top \mathbf{x}_s\right)^2 \\
&= \frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{w}^\top \mathbf{x}_t - \mathbf{w}^\top \left(\frac{1}{n}\sum_{s=1}^{n} \mathbf{x}_s\right)\right)^2 \\
&= \frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{w}^\top (\mathbf{x}_t - \mu)\right)^2 \\
&= \text{ average squared inner product}
\end{aligned}
$$

Which Direction?

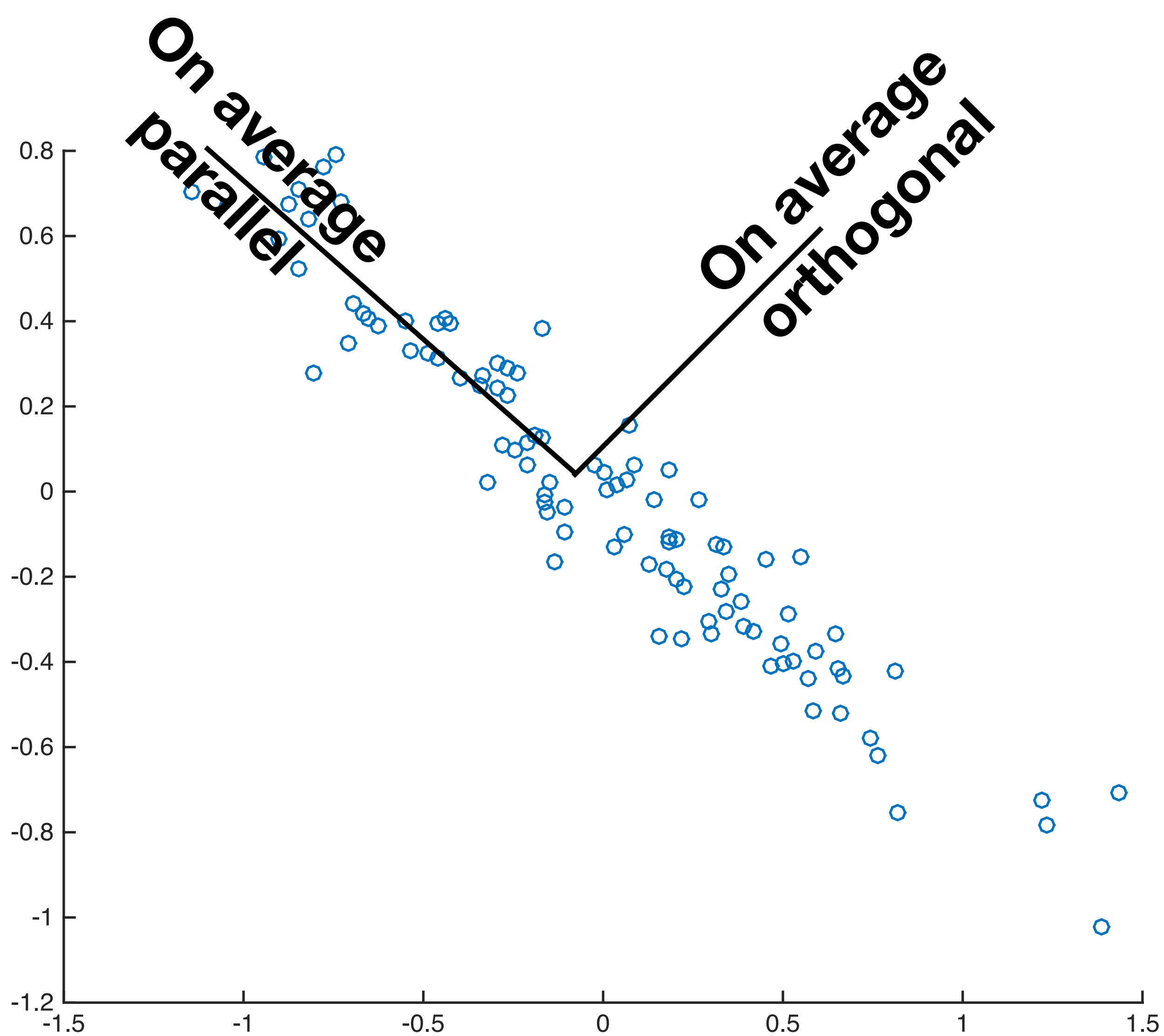


Which Direction?



$$\frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{w}^\top(\mathbf{x}_t-\mu)\right)^2 = \frac{1}{n}\sum_{t=1}^{n}\|\mathbf{x}_t-\mu\|^2 \text{cosine}(w, x_t-\mu)$$

Which Direction?



$$\frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{w}^\top(\mathbf{x}_t-\mu)\right)^2 = \frac{1}{n}\sum_{t=1}^{n}\|\mathbf{x}_t-\mu\|^2\text{cosine}(w, x_t-\mu)$$

PCA: Variance Maximization

- Pick directions along which data varies the most
- First principal component:

$$\mathbf{w}_1 = \arg\max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{w}^\top \mathbf{x}_t - \frac{1}{n}\sum_{t=1}^{n}\mathbf{w}^\top \mathbf{x}_t\right)^2$$

$$= \arg\max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{w}^\top(\mathbf{x}_t - \mu)\right)^2$$

PCA: Variance Maximization

- Pick directions along which data varies the most
- First principal component:

$$
\begin{aligned}
\mathbf{w}_1 &= \arg\max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{w}^\top \mathbf{x}_t - \frac{1}{n}\sum_{t=1}^{n}\mathbf{w}^\top \mathbf{x}_t\right)^2 \\
&= \arg\max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{w}^\top(\mathbf{x}_t - \mu)\right)^2 \\
&= \arg\max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \frac{1}{n}\sum_{t=1}^{n}\mathbf{w}^\top(\mathbf{x}_t - \mu)(\mathbf{x}_t - \mu)^\top \mathbf{w}
\end{aligned}
$$

PCA: Variance Maximization

- Pick directions along which data varies the most
- First principal component:

$$\begin{aligned}\mathbf{w}_1 &= \arg\max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{w}^\top \mathbf{x}_t - \frac{1}{n}\sum_{t=1}^{n}\mathbf{w}^\top \mathbf{x}_t\right)^2 \\ &= \arg\max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{w}^\top(\mathbf{x}_t-\mu)\right)^2 \\ &= \arg\max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \frac{1}{n}\sum_{t=1}^{n}\mathbf{w}^\top(\mathbf{x}_t-\mu)(\mathbf{x}_t-\mu)^\top\mathbf{w} \\ &= \arg\max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \mathbf{w}^\top\Sigma\mathbf{w}\end{aligned}$$

Σ is the covariance matrix

PCA: Variance Maximization

- Pick directions along which data varies the most
- First principal component:

$$\mathbf{w}_1 = \arg\max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \mathbf{w}^\top \Sigma \mathbf{w}$$

Σ is the covariance matrix

PCA: Variance Maximization

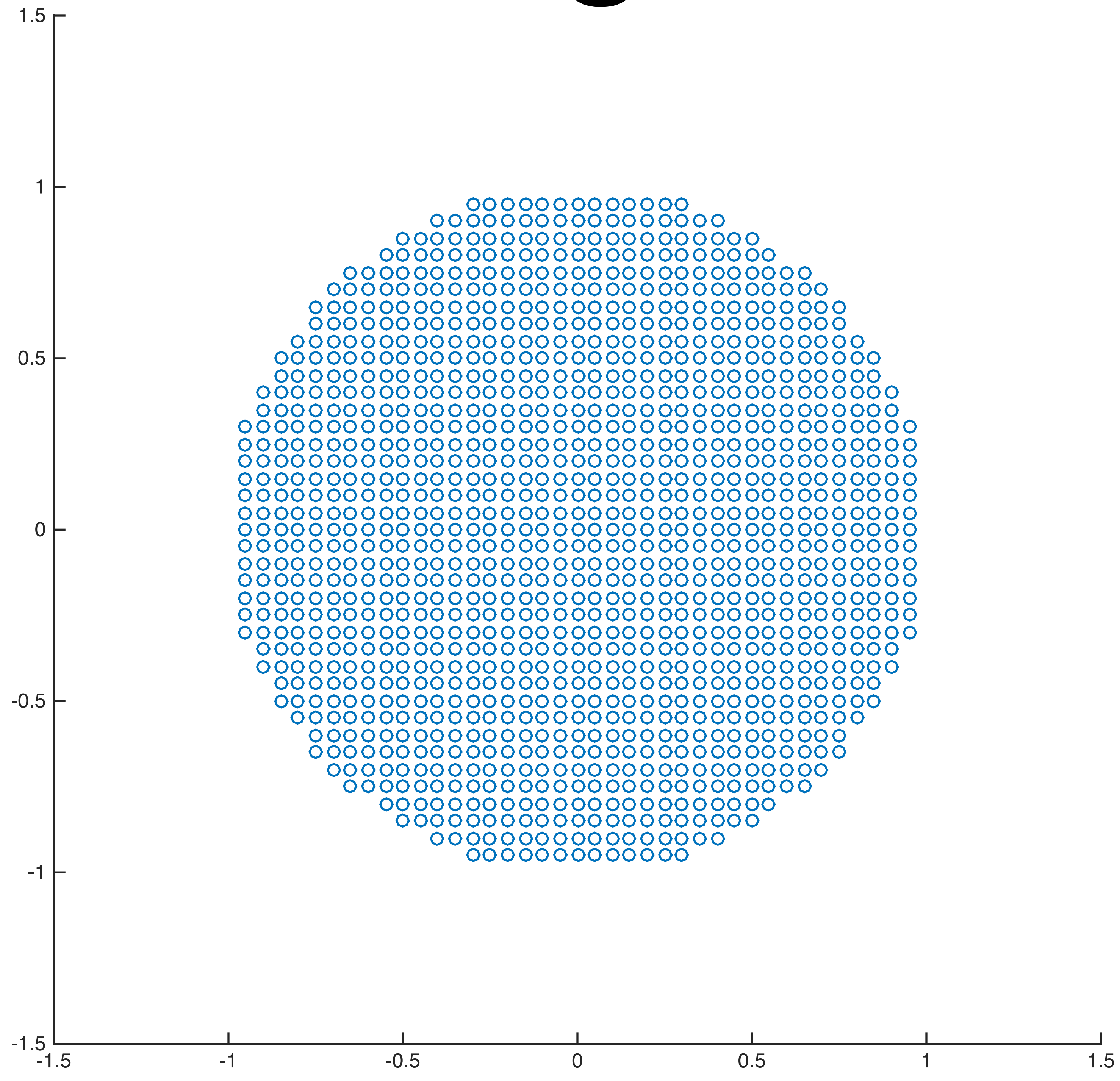
- Pick directions along which data varies the most
- First principal component:

$$\mathbf{w}_1 = \arg \max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \mathbf{w}^\top \Sigma \mathbf{w}$$

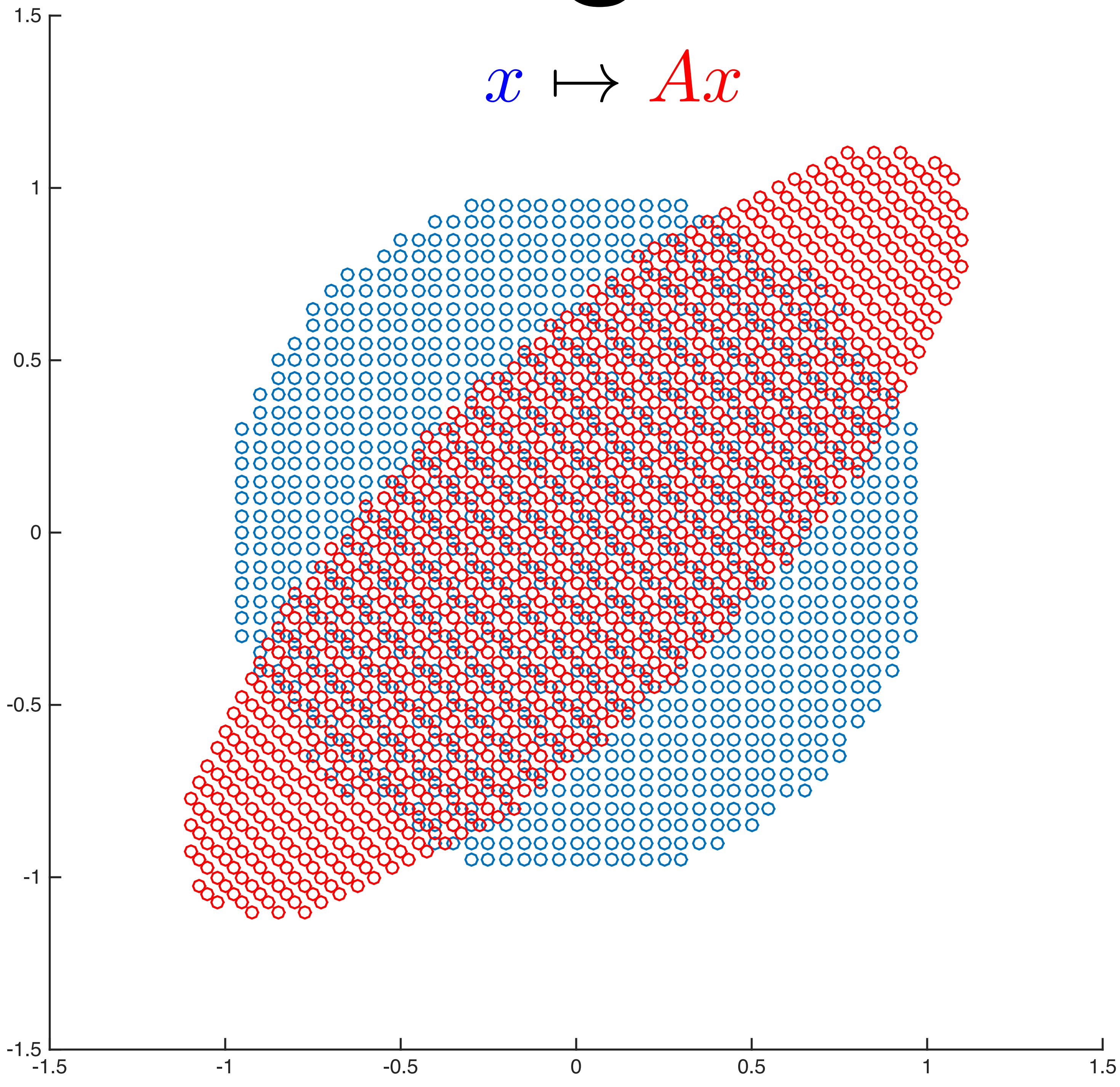
Σ is the covariance matrix

Solution: $\mathbf{w}_1 =$ Largest Eigenvector of Σ

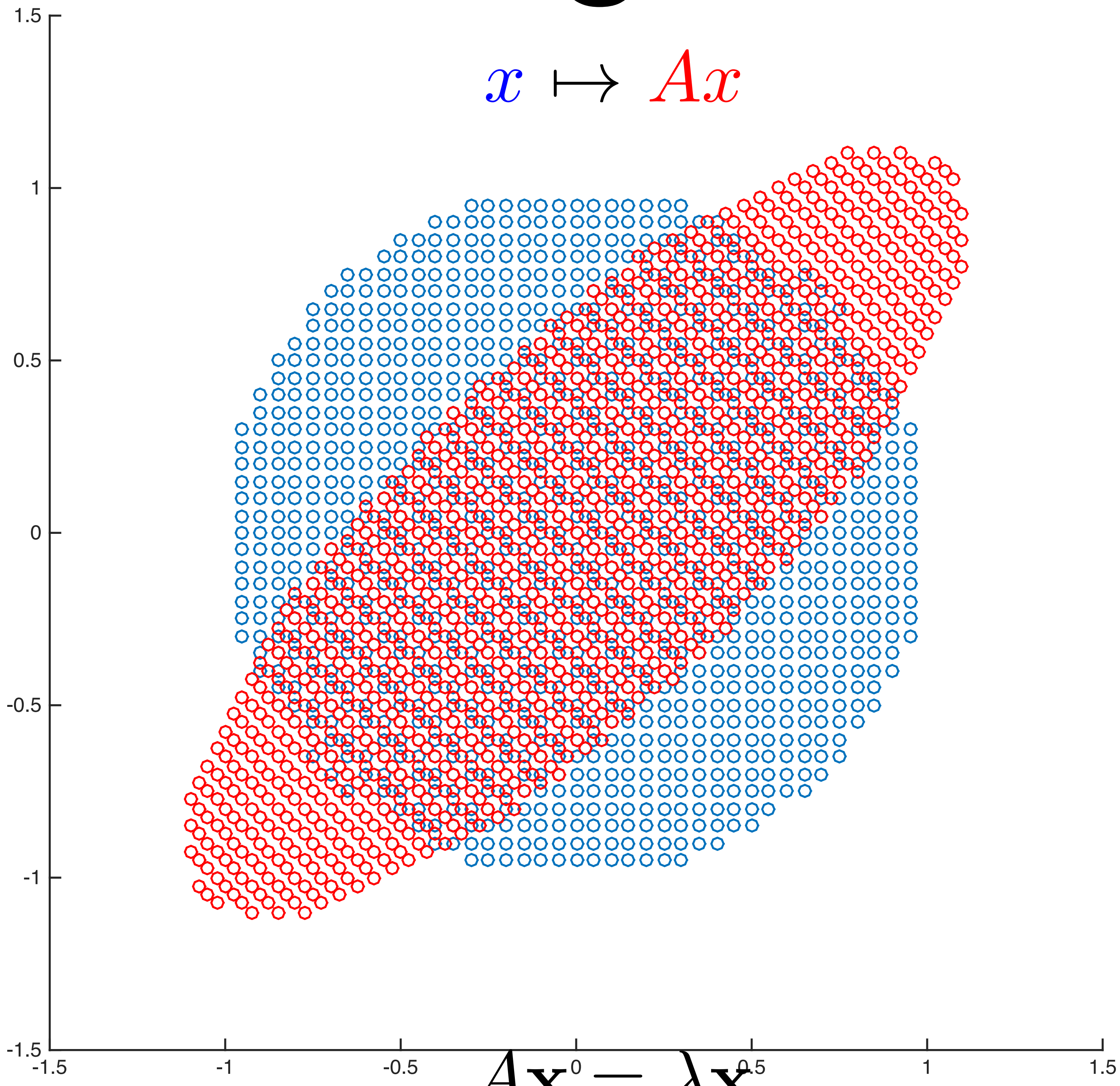
What are Eigen Vectors?



What are Eigen Vectors?

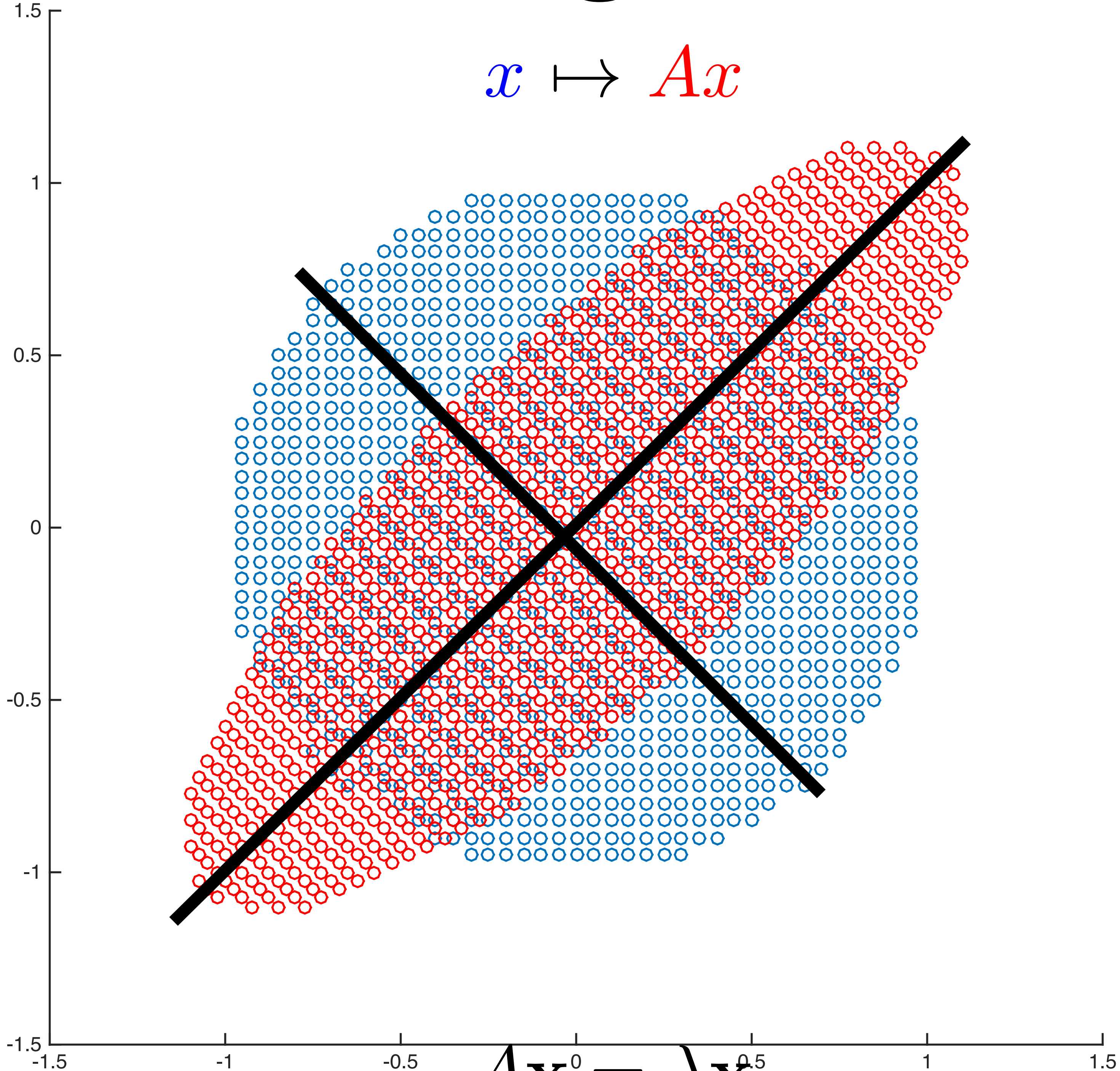


What are Eigen Vectors?



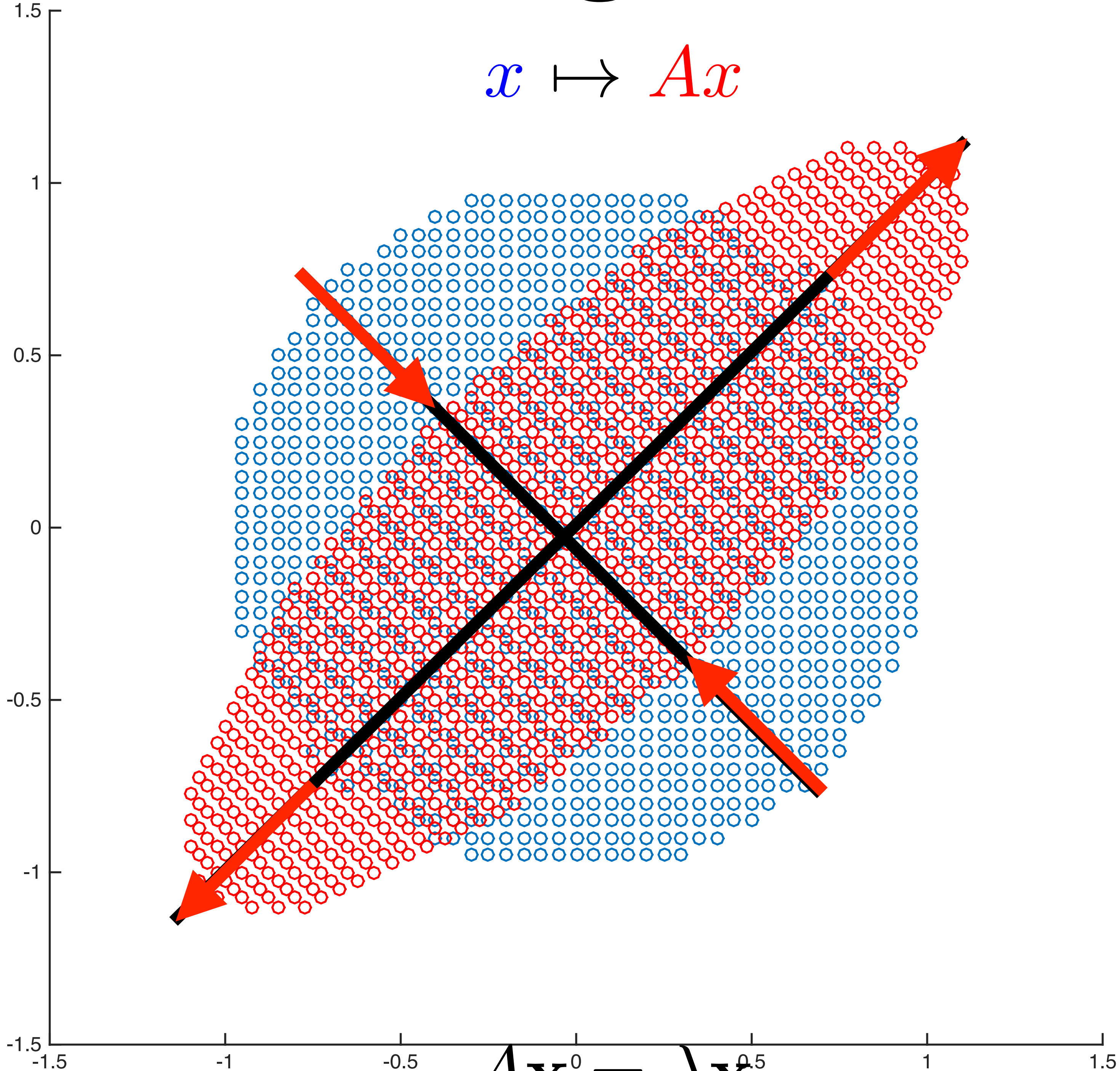
$$A\mathbf{x} = \lambda\mathbf{x}$$

What are Eigen Vectors?



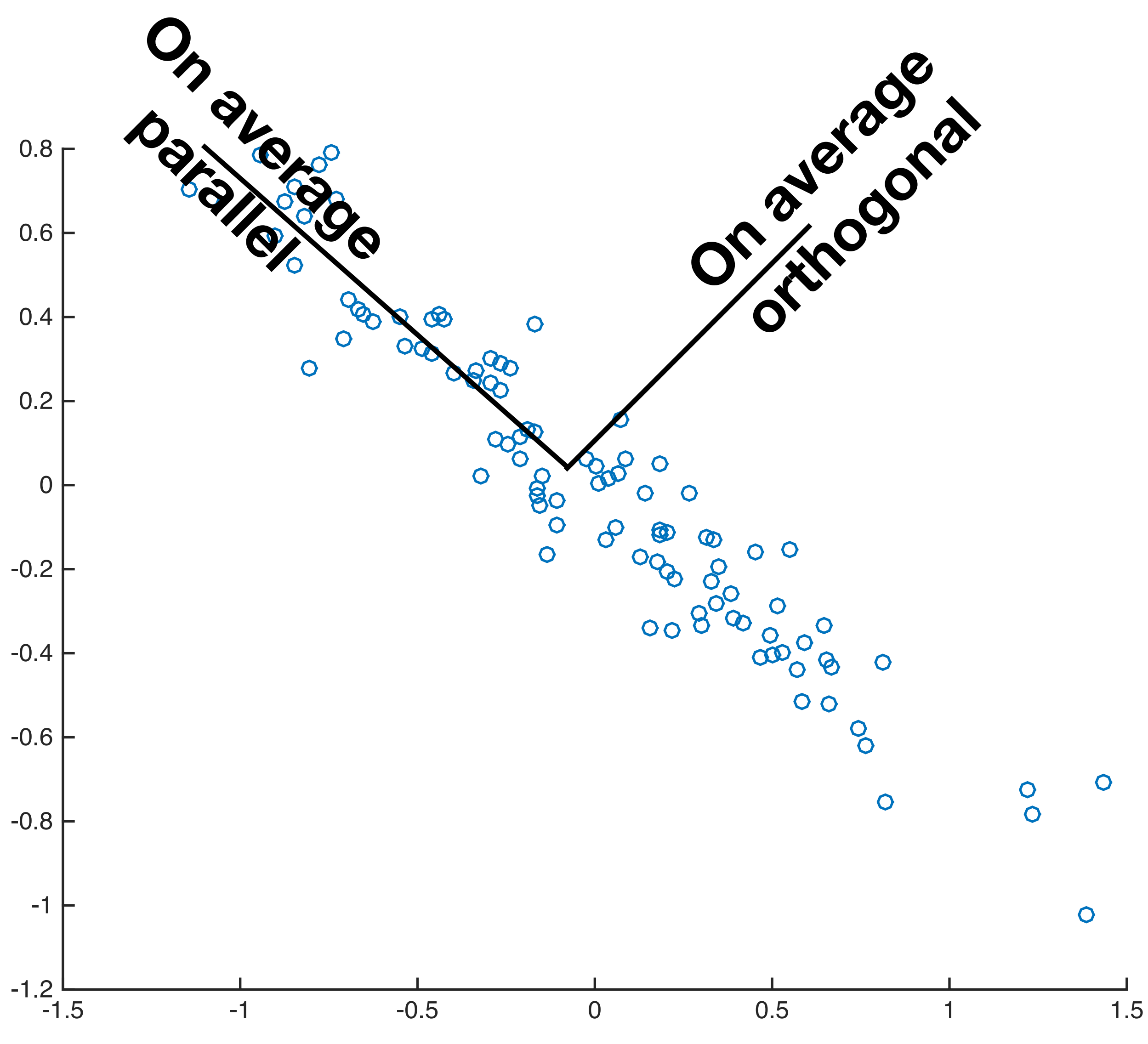
$$A\mathbf{x} = \lambda\mathbf{x}$$

What are Eigen Vectors?

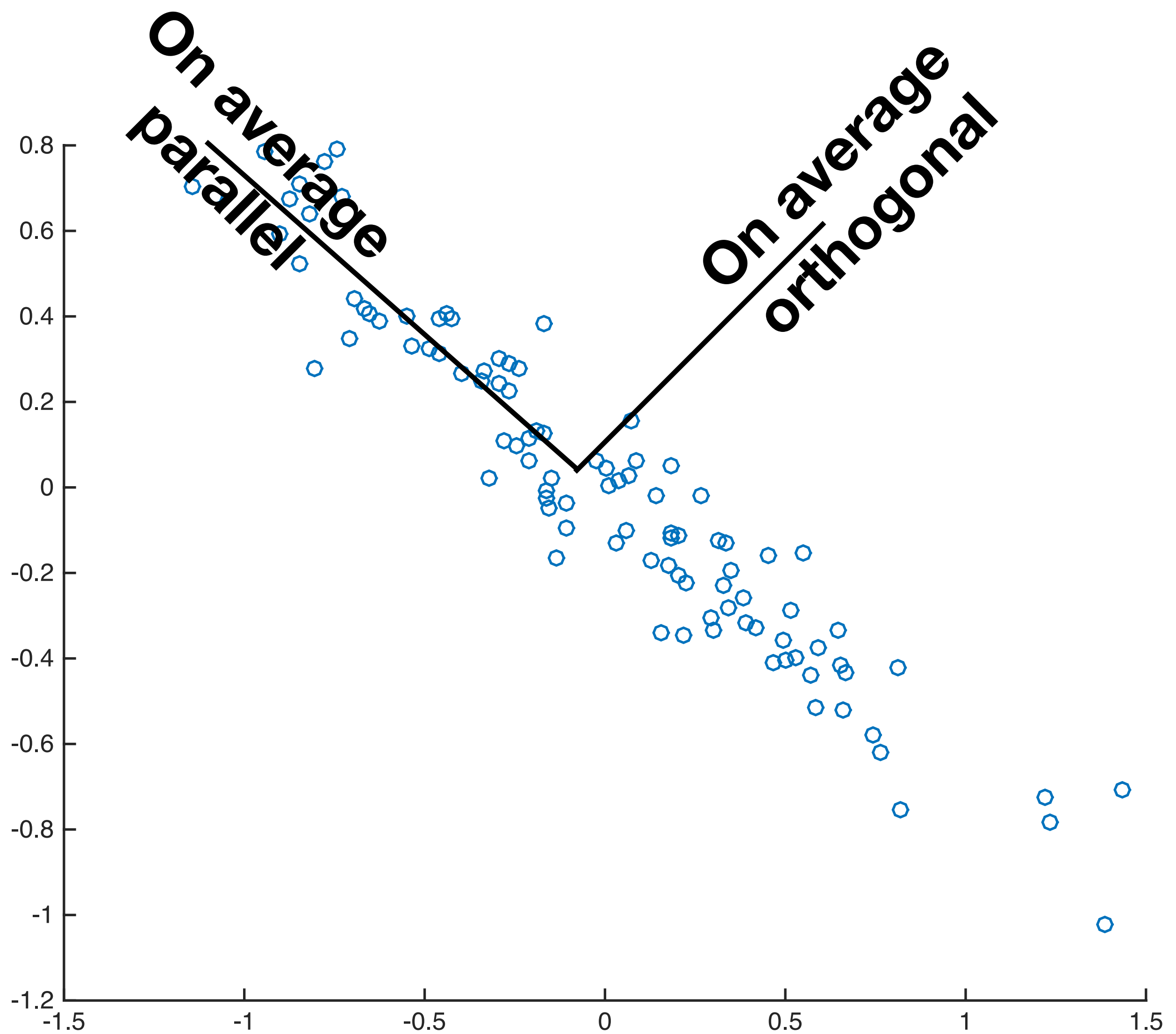


$$A\mathbf{x} = \lambda\mathbf{x}$$

Which Direction?



Which Direction?



Top Eigenvector of covariance matrix

- What if we want more than one number for each data point?
- That is we want to reduce to K > 1 dimensions?

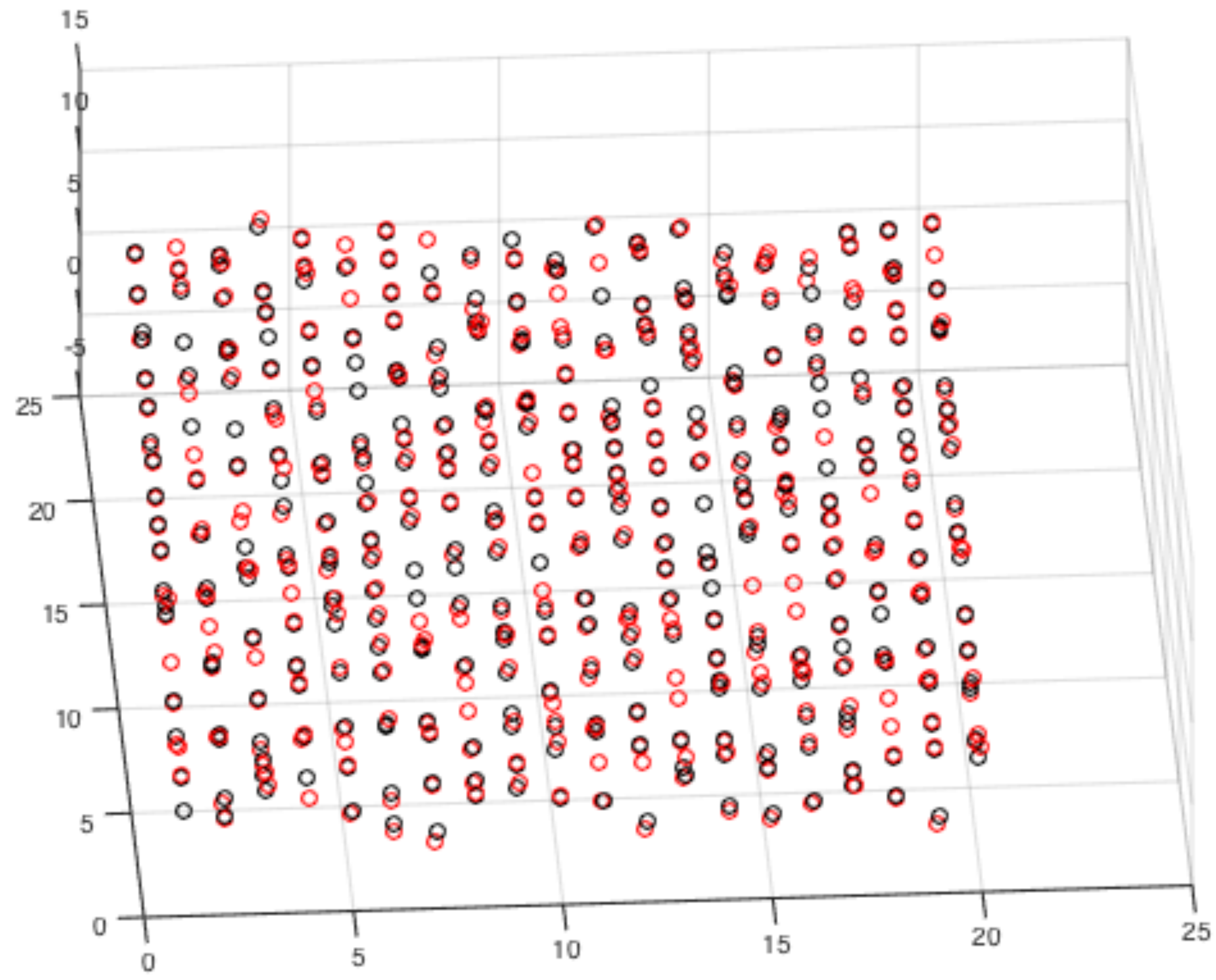
PCA: Variance Maximization

- How do we find the K components?

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Intuition: Remove top direction, now reduce dimension for remaining d-1 dimensions

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

Principal Component Analysis

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

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

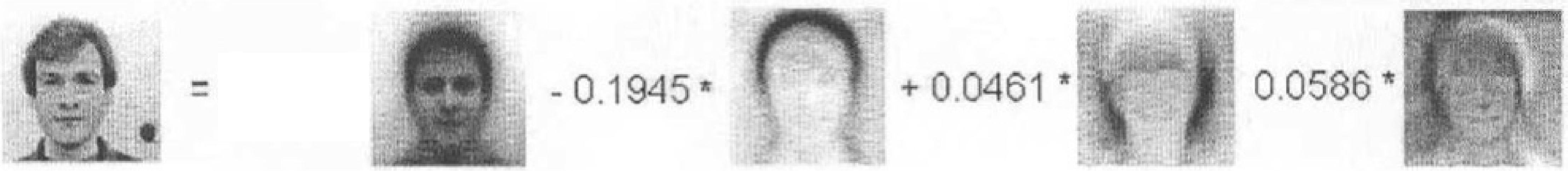
3. $Y = X \times W$

Demo

Principal Component Analysis (PCA)

Turk & Pentland'91

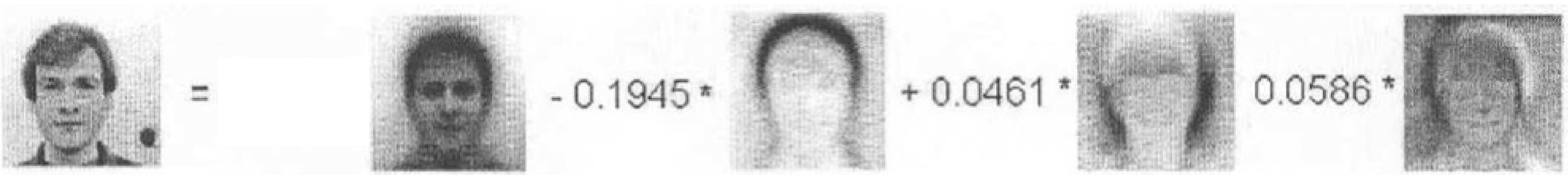
Eigen Face:



PRINCIPAL COMPONENT ANALYSIS (PCA)

Turk & Pentland'91

Eigen Face:

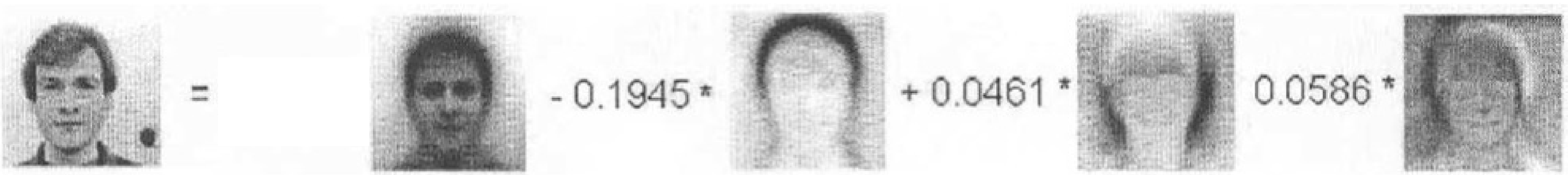


- Each x_t (each row of X) is a face image (vectorized version)

Principal Component Analysis (PCA)

Turk & Pentland'91

Eigen Face:

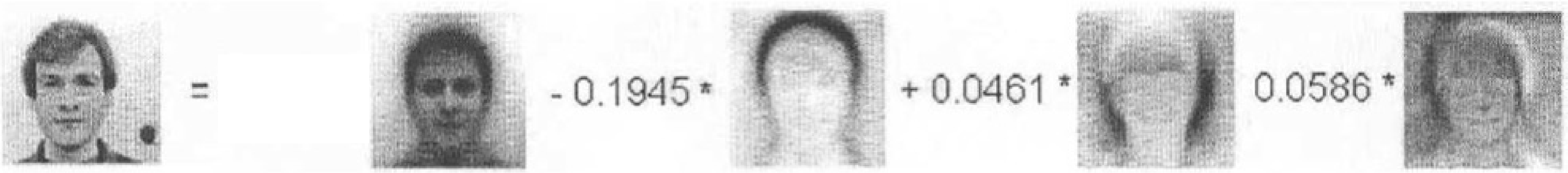


- Each x_t (each row of X) is a face image (vectorized version)
- Each y_t is the set of coefficients we multiply to the eigen face

PRINCIPAL COMPONENT ANALYSIS (PCA)

Turk & Pentland'91

Eigen Face:



- Each x_t (each row of X) is a face image (vectorized version)
- Each y_t is the set of coefficients we multiply to the eigen face
- Each column of W is an Eigenface

Orthonormal Projections

Orthonormal Projections

- Think of $\mathbf{w}_1, \ldots, \mathbf{w}_K$ as coordinate system for PCA (in a K dimensional subspace)

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

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. $\mathbf{w}_i \perp \mathbf{w}_j$ & $\|\mathbf{w}_i\| = 1$.

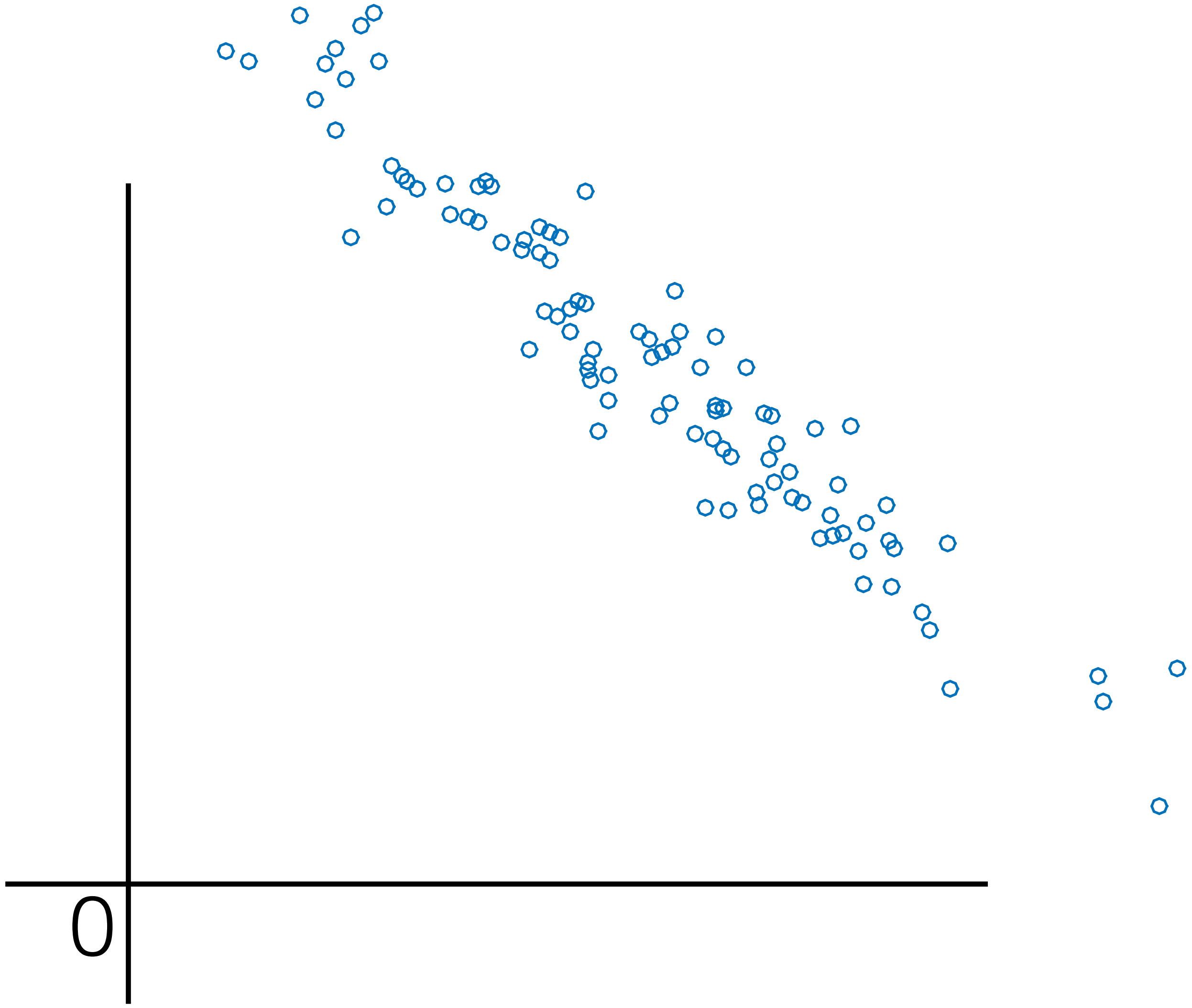
Orthonormal Projections

- 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. $\mathbf{w}_i \perp \mathbf{w}_j$ & $\|\mathbf{w}_i\| = 1$.

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

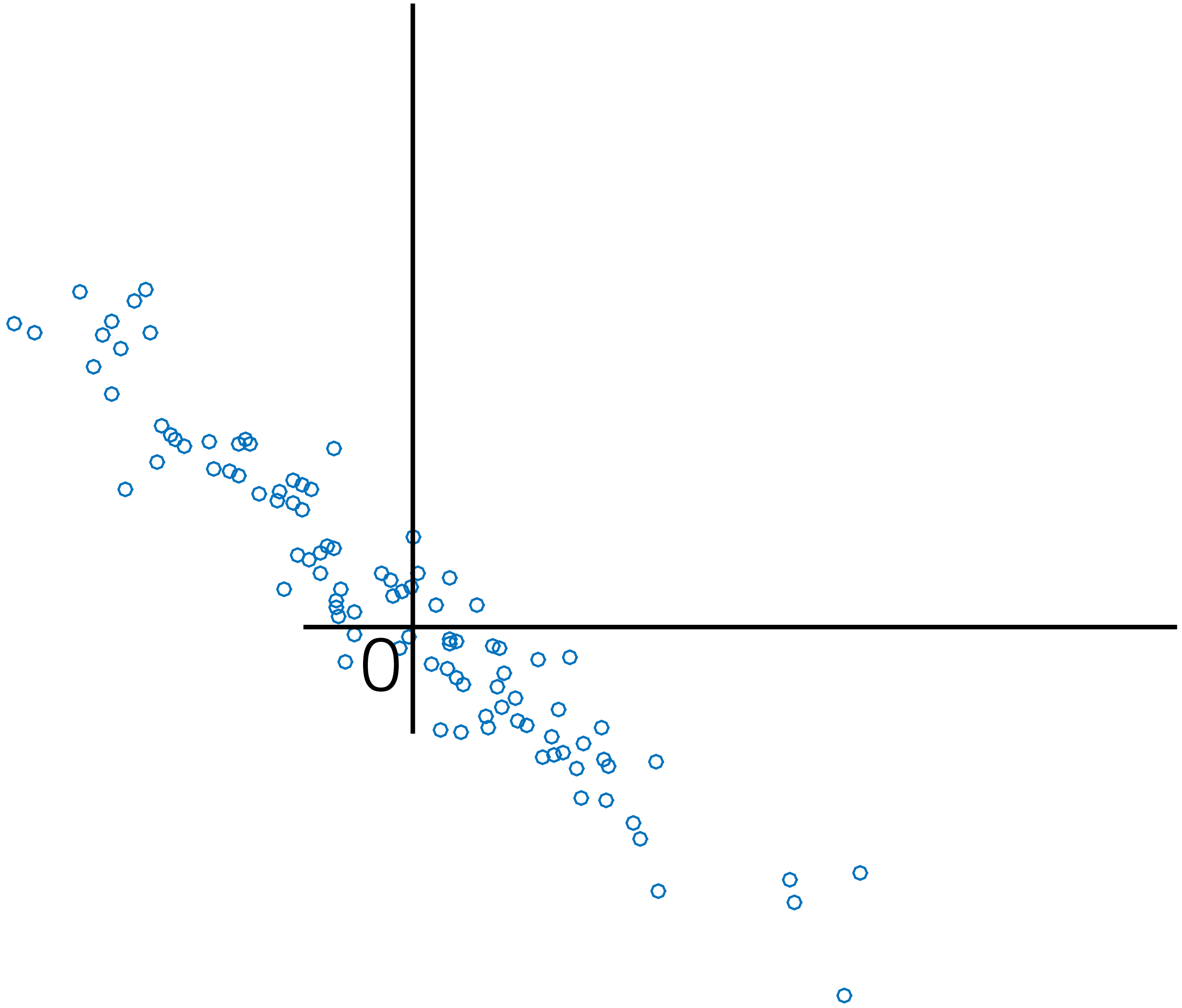
$$\mathbf{w}_i \perp \mathbf{w}_j \Rightarrow \sum_{k=1}^{d} \mathbf{w}_i[k]\mathbf{w}_j[k] = 0$$

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 $\mu = \frac{1}{n}\sum_{t=1}^{n} x_t$.

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,

$$\|\hat{\mathbf{x}}_t - \mathbf{x}_t\|_2^2$$

PCA: Minimizing Reconstruction Error

- Goal: find the basis that minimizes reconstruction error,

$$\|\hat{\mathbf{x}}_t - \mathbf{x}_t\|_2^2 = \left\| \sum_{j=1}^{K} \mathbf{y}_t[j]\mathbf{w}_j + \mu - \mathbf{x}_t \right\|_2^2$$

PCA: Minimizing Reconstruction Error

- Goal: find the basis that minimizes reconstruction error,

$$
\begin{aligned}
\|\hat{\mathbf{x}}_t - \mathbf{x}_t\|_2^2 &= \left\| \sum_{j=1}^{K} \mathbf{y}_t[j]\mathbf{w}_j + \mu - \mathbf{x}_t \right\|_2^2 \\
&= \left\| \sum_{j=1}^{K} \mathbf{y}_t[j]\mathbf{w}_j + \mu - \sum_{j=1}^{d} \mathbf{y}_t[j]\mathbf{w}_j - \mu \right\|_2^2
\end{aligned}
$$

PCA: Minimizing Reconstruction Error

- Goal: find the basis that minimizes reconstruction error,

$$
\begin{aligned}
\|\hat{\mathbf{x}}_t - \mathbf{x}_t\|_2^2 &= \left\| \sum_{j=1}^{K} \mathbf{y}_t[j]\mathbf{w}_j + \mu - \mathbf{x}_t \right\|_2^2 \\
&= \left\| \sum_{j=1}^{K} \mathbf{y}_t[j]\mathbf{w}_j + \mu - \sum_{j=1}^{d} \mathbf{y}_t[j]\mathbf{w}_j - \mu \right\|_2^2 \\
&= \left\| \sum_{j=K+1}^{d} \mathbf{y}_t[j]\mathbf{w}_j \right\|_2^2
\end{aligned}
$$

PCA: Minimizing Reconstruction Error

- Goal: find the basis that minimizes reconstruction error,

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

PCA: Minimizing Reconstruction Error

- Goal: find the basis that minimizes reconstruction error,

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

PCA: Minimizing Reconstruction Error

- Goal: find the basis that minimizes reconstruction error,

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

PCA: Minimizing Reconstruction Error

- Goal: find the basis that minimizes reconstruction error,

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

PCA: Minimizing Reconstruction Error

$$\frac{1}{n}\sum_{t=1}^{n}\|\hat{\mathbf{x}}_t - \mathbf{x}_t\|_2^2 = \frac{1}{n}\sum_{t=1}^{n}\sum_{j=k+1}^{d}\mathbf{y}_t[j]^2\|\mathbf{w}_j\|_2^2 \qquad (\text{but } \|\mathbf{w}_j\| = 1)$$

PCA: Minimizing Reconstruction Error

$$\frac{1}{n}\sum_{t=1}^{n}\|\hat{\mathbf{x}}_t - \mathbf{x}_t\|_2^2 = \frac{1}{n}\sum_{t=1}^{n}\sum_{j=k+1}^{d}\mathbf{y}_t[j]^2\|\mathbf{w}_j\|_2^2 \qquad (\text{but } \|\mathbf{w}_j\| = 1)$$

$$= \frac{1}{n}\sum_{t=1}^{n}\sum_{j=k+1}^{d}\mathbf{y}_t[j]^2$$

PCA: Minimizing Reconstruction Error

$$\frac{1}{n}\sum_{t=1}^{n}\|\hat{\mathbf{x}}_t - \mathbf{x}_t\|_2^2 = \frac{1}{n}\sum_{t=1}^{n}\sum_{j=k+1}^{d}\mathbf{y}_t[j]^2\|\mathbf{w}_j\|_2^2 \quad (\text{but } \|\mathbf{w}_j\| = 1)$$

$$= \frac{1}{n}\sum_{t=1}^{n}\sum_{j=k+1}^{d}\mathbf{y}_t[j]^2 \quad (\text{now } \mathbf{y}_j = \mathbf{w}_j^\top(\mathbf{x}_t - \mu))$$

$$= \frac{1}{n}\sum_{t=1}^{n}\sum_{j=k+1}^{d}(\mathbf{w}_j^\top(\mathbf{x}_t - \mu))^2$$

PCA: Minimizing Reconstruction Error

$$\frac{1}{n}\sum_{t=1}^{n}\|\hat{\mathbf{x}}_t - \mathbf{x}_t\|_2^2 = \frac{1}{n}\sum_{t=1}^{n}\sum_{j=k+1}^{d}\mathbf{y}_t[j]^2\|\mathbf{w}_j\|_2^2 \quad (\text{but } \|\mathbf{w}_j\| = 1)$$

$$= \frac{1}{n}\sum_{t=1}^{n}\sum_{j=k+1}^{d}\mathbf{y}_t[j]^2 \quad (\text{now } \mathbf{y}_j = \mathbf{w}_j^\top(\mathbf{x}_t - \mu))$$

$$= \frac{1}{n}\sum_{t=1}^{n}\sum_{j=k+1}^{d}\left(\mathbf{w}_j^\top(\mathbf{x}_t - \mu)\right)^2$$

$$= \frac{1}{n}\sum_{t=1}^{n}\sum_{j=k+1}^{d}\mathbf{w}_j^\top(\mathbf{x}_t - \mu)(\mathbf{x}_t - \mu)^\top\mathbf{w}_j$$

PCA: Minimizing Reconstruction Error

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

PCA: Minimizing Reconstruction Error

$$\text{Claim: } \sum_{j=1}^{d} \mathbf{w}_j^\top \Sigma \mathbf{w}_j = \text{ Constant } = \frac{1}{n}\sum_{t=1}^{n} \|x_t - \mu\|_2^2$$

PCA: Minimizing Reconstruction Error

Claim: $\sum_{j=1}^{d} \mathbf{w}_j^\top \Sigma \mathbf{w}_j = \text{ Constant } = \frac{1}{n}\sum_{t=1}^{n} \|x_t - \mu\|_2^2$

Recall that: $\frac{1}{n}\sum_{t=1}^{n} \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|_2^2 = \sum_{j=K+1}^{d} \mathbf{w}_j^\top \Sigma \mathbf{w}_j$

PCA: Minimizing Reconstruction Error

Claim: $\sum_{j=1}^{d} \mathbf{w}_j^\top \Sigma \mathbf{w}_j = \text{ Constant } = \frac{1}{n}\sum_{t=1}^{n} \|x_t - \mu\|_2^2$

Recall that: $\frac{1}{n}\sum_{t=1}^{n} \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|_2^2 = \sum_{j=K+1}^{d} \mathbf{w}_j^\top \Sigma \mathbf{w}_j$

Take K $= 0$ so that $\hat{\mathbf{x}}_t = \mu$

PCA: Minimizing Reconstruction Error

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

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

PCA: Minimizing Reconstruction Error

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

$$\begin{aligned}
&\underset{\forall j,\ \|\mathbf{w}_j\|_2=1}{\operatorname{argmin}} \sum_{j=k+1}^{d} \mathbf{w}_j^\top \Sigma \mathbf{w}_j \\
&= \underset{\forall j,\ \|\mathbf{w}_j\|_2=1}{\operatorname{argmin}} \left(\frac{1}{n} \sum_{t=1}^{n} \|\hat{\mathbf{x}}_t - \mu\|_2^2 - \sum_{j=1}^{K} \mathbf{w}_j^\top \Sigma \mathbf{w}_j \right)
\end{aligned}$$

PCA: Minimizing Reconstruction Error

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

$$\begin{aligned}
&\underset{\forall j,\ \|\mathbf{w}_j\|_2=1}{\operatorname{argmin}} \sum_{j=k+1}^{d} \mathbf{w}_j^\top \Sigma \mathbf{w}_j \\
&= \underset{\forall j,\ \|\mathbf{w}_j\|_2=1}{\operatorname{argmin}} \left(\frac{1}{n} \sum_{t=1}^{n} \|\hat{\mathbf{x}}_t - \mu\|_2^2 - \sum_{j=1}^{K} \mathbf{w}_j^\top \Sigma \mathbf{w}_j \right) \\
&= \underset{\|\mathbf{w}_j\|_2=1, \mathbf{w}_j \perp \mathbf{w}_k}{\operatorname{argmax}} \sum_{j=1}^{K} \mathbf{w}_j^\top \Sigma \mathbf{w}_j
\end{aligned}$$

Maximize Total Spread = Minimize Reconstruction Error

Principal Component Analysis

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

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

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

Reconstruction

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

Principal Component Analysis: Demo