Machine Learning for Data Science (CS4786)
Lecture 3

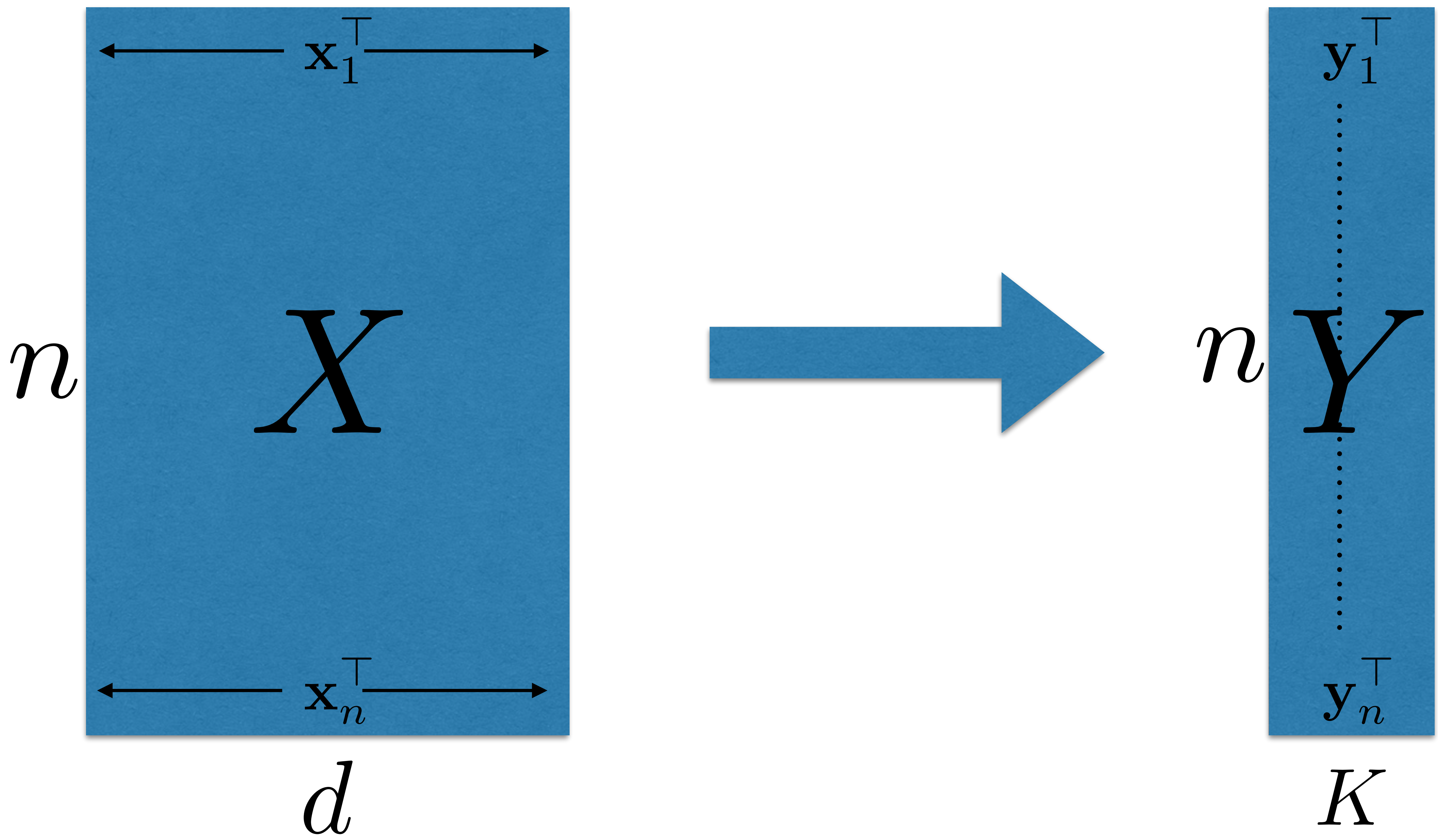
Principal Component Analysis

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

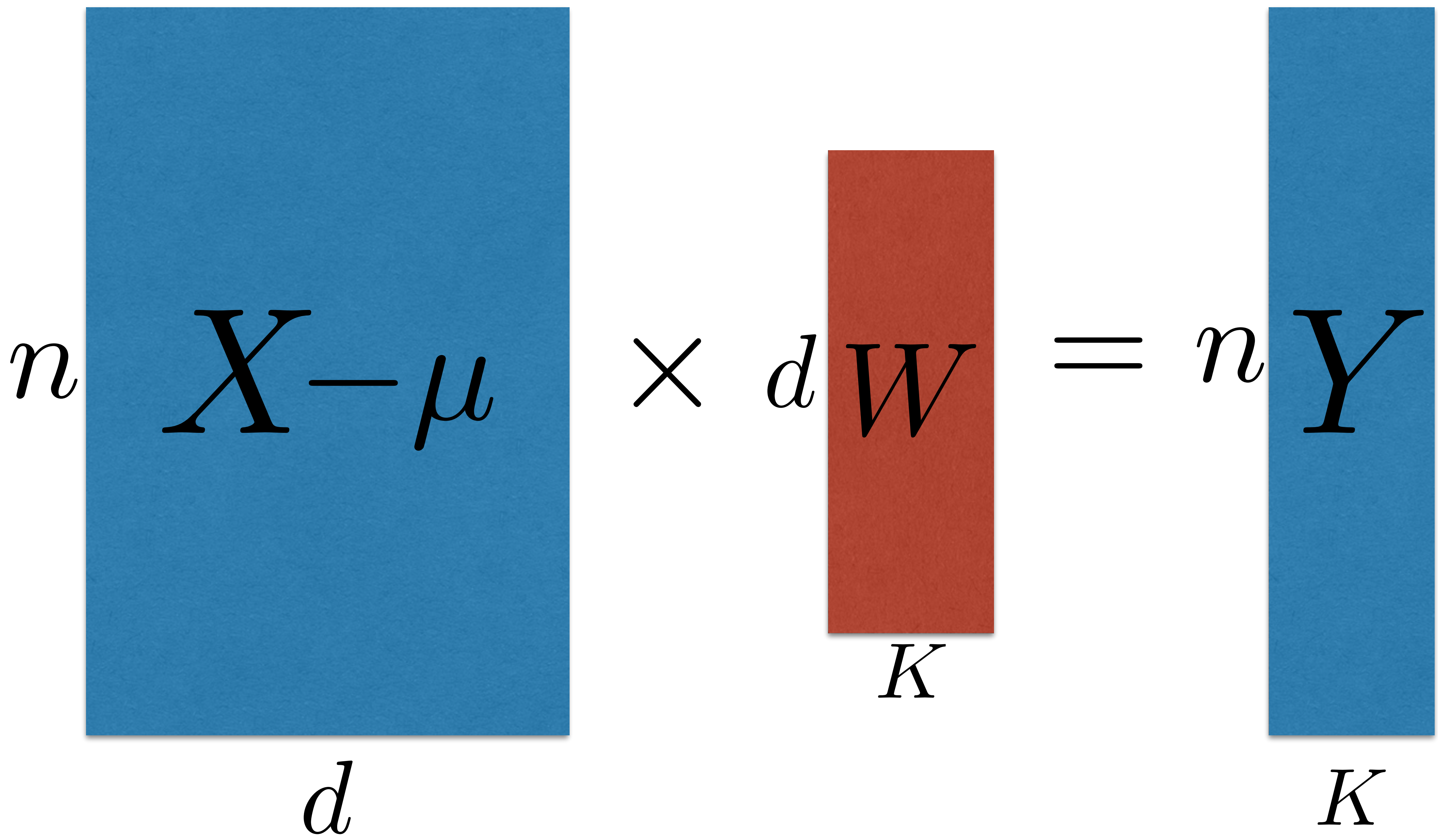
Announcements

- Waitlist size currently about 55 :(

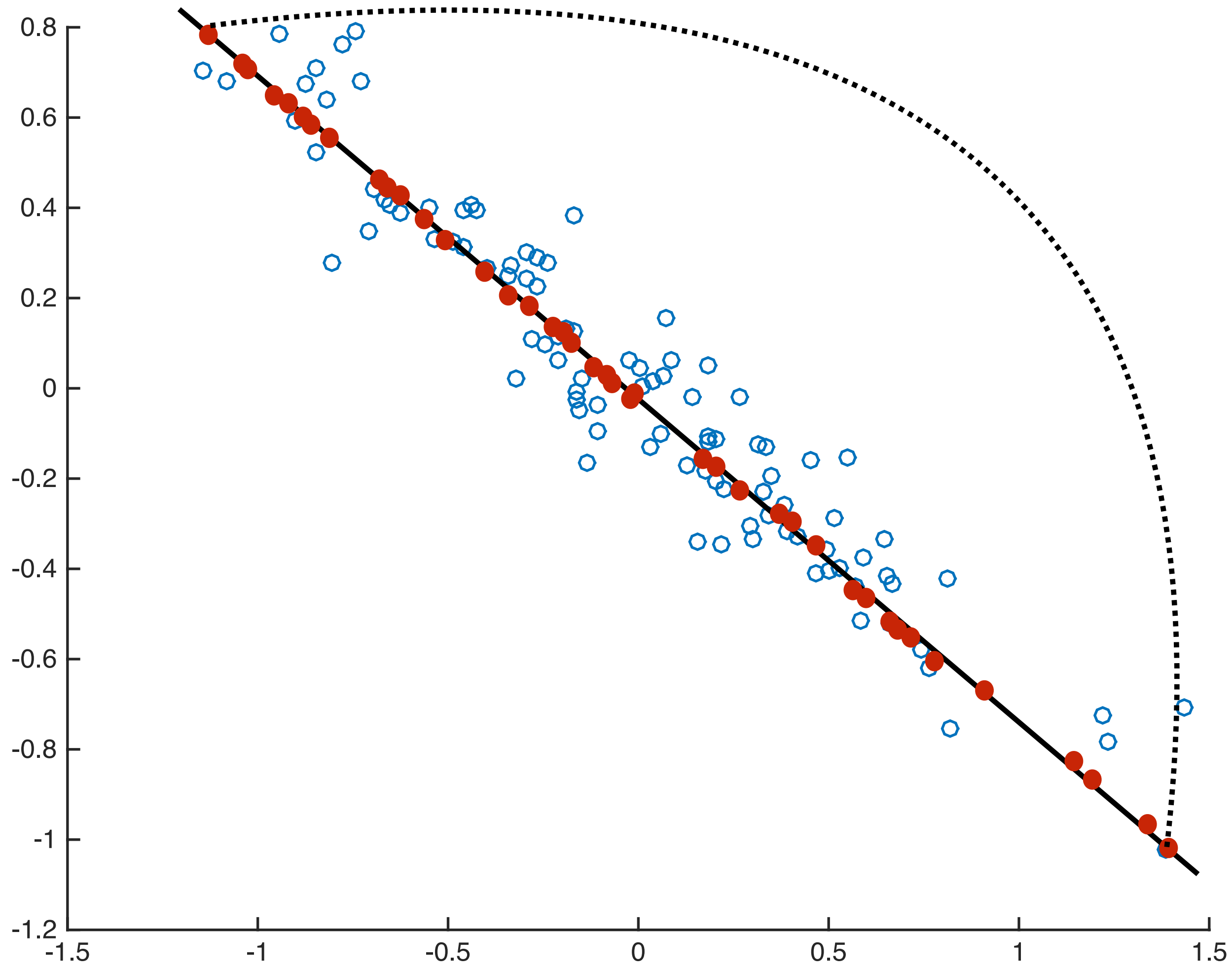
Dimensionality Reduction



Dimensionality Reduction



PCA: Variance Maximization



First principal direction = Top eigen vector

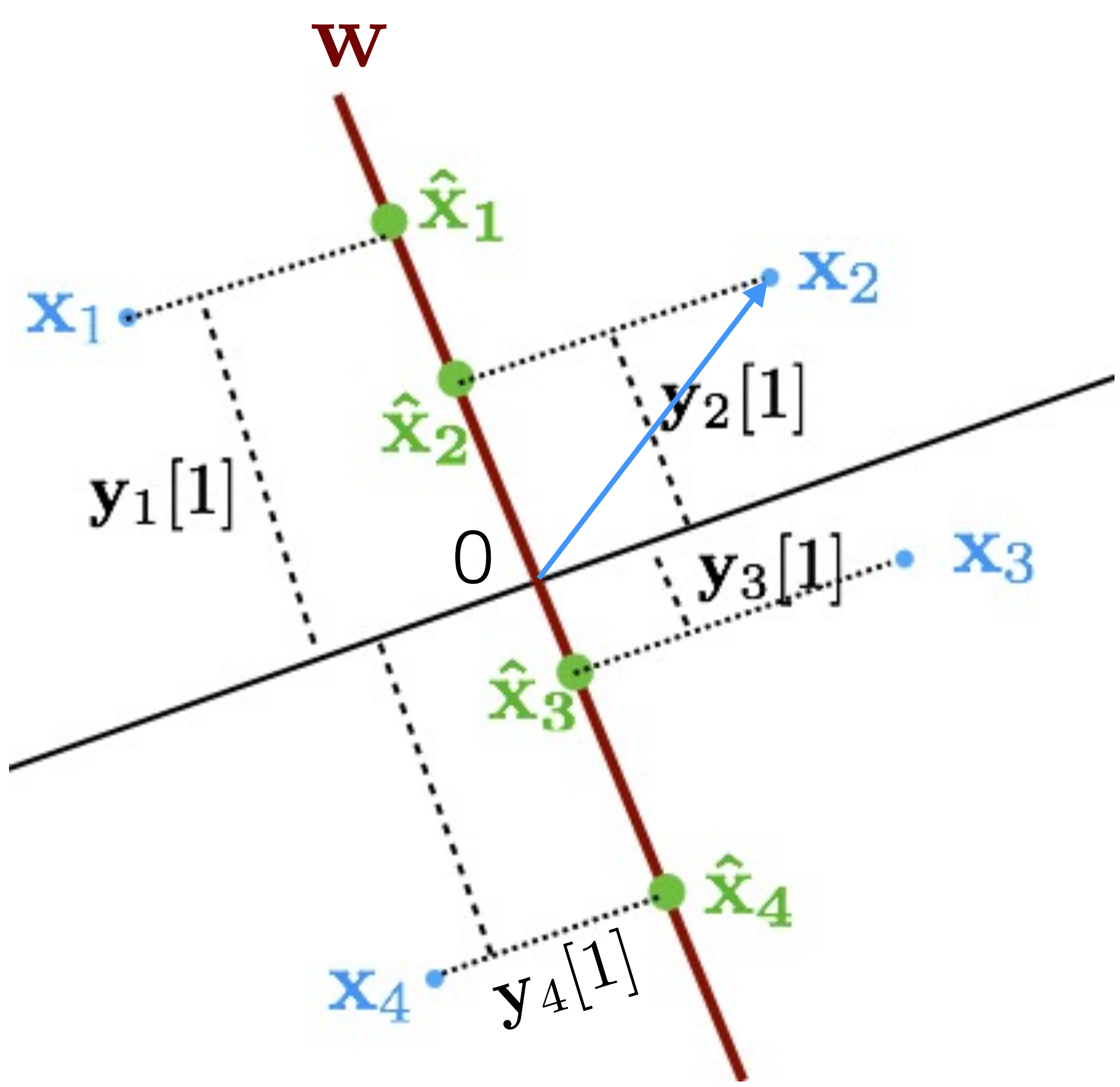
Principal Component Analysis

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

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

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

A Picture



$$\mathbf{y}_2[1] = \mathbf{x}_1^\top \mathbf{w} = \|\mathbf{x}_2\| \cos(\angle \mathbf{x}\mathbf{w})$$

Orthonormal Projections

- Think of $W_1, \ldots, W_K$ as coordinate system for PCA
- $\mathbf{y}$ values provide coefficients in this system
- Without loss of generality, $W_1, \ldots, W_K$ can be orthonormal, i.e. $W_i \perp W_j$ & $\|W_i\| = 1$.
- Reconstruction:

$$\hat{\mathbf{x}}_t = \mathbf{y}_t^\top W^\top + \mu$$

PCA: Variance Maximization

- How do we find the remaining components?

PCA: Variance Maximization

- How do we find the remaining components?
- We are looking for orthogonal directions.

PCA: Variance Maximization

- How do we find the remaining components?
- We are looking for orthogonal directions.
- Start with the d dimensional space
- While we haven't yet found K directions,
  - Find first principal component direction
  - Remove this direction and consider data points in the remaining subspace after projecting to first component

  End

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

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
- Recall $\mathrm{cov}(A, B) = \mathbb{E}[(A - \mathbb{E}[A])(B - \mathbb{E}[B])]$
- Alternatively,

$$\Sigma[i, j] = \frac{1}{n}\begin{bmatrix} \mathbf{x}_1[i] - \mu[i] \\ \vdots \\ \mathbf{x}_n[i] - \mu[i] \end{bmatrix}^\top \begin{bmatrix} \mathbf{x}_1[j] - \mu[j] \\ \vdots \\ \mathbf{x}_n[j] - \mu[j] \end{bmatrix}$$

Inner products measure similarity.

PCA: Minimizing Reconstruction Error

- Goal: find the basis that minimizes reconstruction error,

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

PCA: Minimizing Reconstruction Error

- Goal: find the basis that minimizes reconstruction error,

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

PCA: Minimizing Reconstruction Error

- Goal: find the basis that minimizes reconstruction error,

$$\frac{1}{n}\sum_{t=1}^{n}\|\hat{\mathbf{x}}_t - \mathbf{x}_t\|_2^2 = \frac{1}{n}\sum_{t=1}^{n}\sum_{j=k+1}^{d}\mathbf{w}_j^\top(\mathbf{x}_t - \mu)(\mathbf{x}_t - \mu)^\top\mathbf{w}_j = \sum_{j=k+1}^{d}\mathbf{w}_j^\top\Sigma\mathbf{w}_j$$

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

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

Using Lagrangian multipliers, there exists $\lambda_{k+1},\ldots,\lambda_d$ such that solution to above is given by:

$$\operatorname{minimize}\sum_{t=1}^{n}\sum_{j=k+1}^{d}\mathbf{w}_j^\top\Sigma\mathbf{w}_j + \sum_{j=k+1}^{d}\lambda_j\|\mathbf{w}_j\|_2^2$$

PCA: Minimizing Reconstruction Error

- Goal: find the basis that minimizes reconstruction error,

$$\frac{1}{n}\sum_{t=1}^{n}\|\hat{\mathbf{x}}_t - \mathbf{x}_t\|_2^2 = \frac{1}{n}\sum_{t=1}^{n}\sum_{j=k+1}^{d}\mathbf{w}_j^\top(\mathbf{x}_t-\mu)(\mathbf{x}_t-\mu)^\top\mathbf{w}_j = \sum_{j=k+1}^{d}\mathbf{w}_j^\top\Sigma\mathbf{w}_j$$

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

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

Using Lagrangian multipliers, there exists $\lambda_{k+1},\ldots,\lambda_d$ such that solution to above is given by:

$$\text{minimize} \sum_{t=1}^{n}\sum_{j=k+1}^{d}\mathbf{w}_j^\top\Sigma\mathbf{w}_j + \sum_{j=k+1}^{d}\lambda_j\|\mathbf{w}_j\|_2^2$$

Setting derivate to 0, $\Sigma\mathbf{w}_j = \lambda_j\mathbf{w}_j$. That is $\mathbf{w}_j$'s are eigenvectors and λ_j's are eigenvalues.

PCA: Minimizing Reconstruction Error

- Solution : $\mathbf{w}_j$'s are eigenvectors and λ_j's are corresponding eigenvalues
- Further, reconstruction error can be written as:

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

- Clearly to minimize reconstruction error, we need to minimize $\sum_{j=k+1}^{d} \lambda_j$. In other words we discard the $d-k$ directions that have the smallest eigenvalue

Principal Component Analysis

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

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

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

Reconstruction

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

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, $\Sigma = (X - \mu)(X - \mu)^\top = UD^2U$

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

$$W = \text{SVD}(X - \mu, K)$$

Principal Component Analysis: Demo