Machine Learning for Data Science (CS4786)
Lecture 4

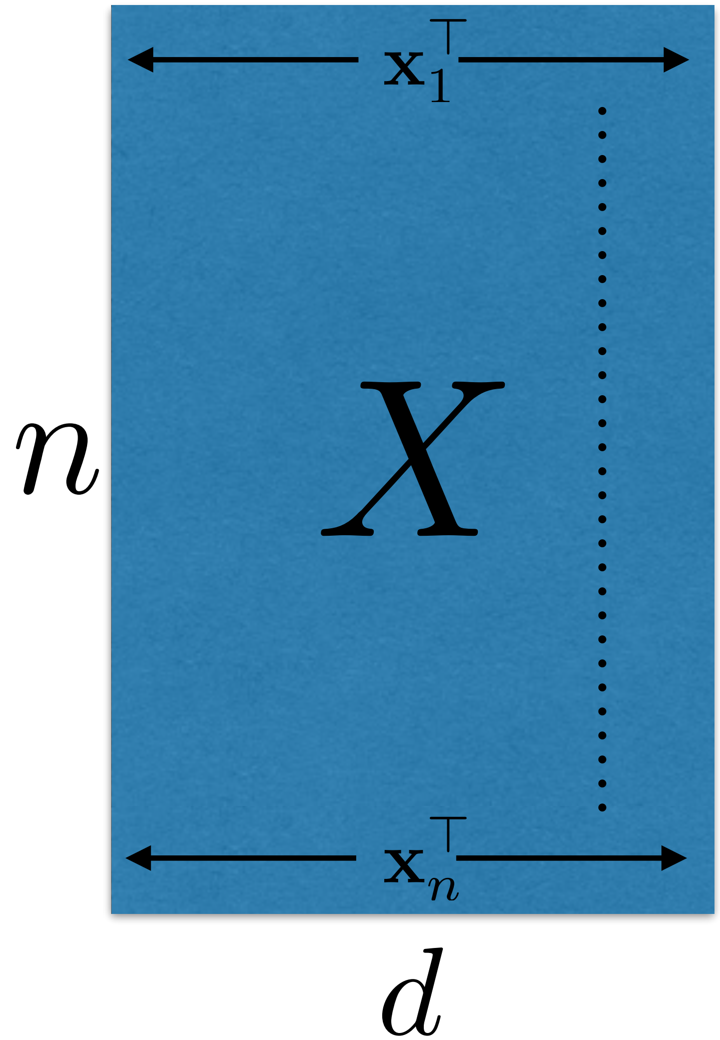
Canonical Correlation Analysis (CCA)

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

Announcement

- We are grading HW0 and you will be added to cms by monday
- HW1 will be posted tonight on webpage (homework tab)
- HW1 on CCA and PCA (due in a week)

Quiz



Assume points are centered. Which of the following are equal to the covariance matrix?

$A. \quad \Sigma = \frac{1}{n}\sum_{t=1}^{n} \mathbf{x}_t^\top \mathbf{x}_t$

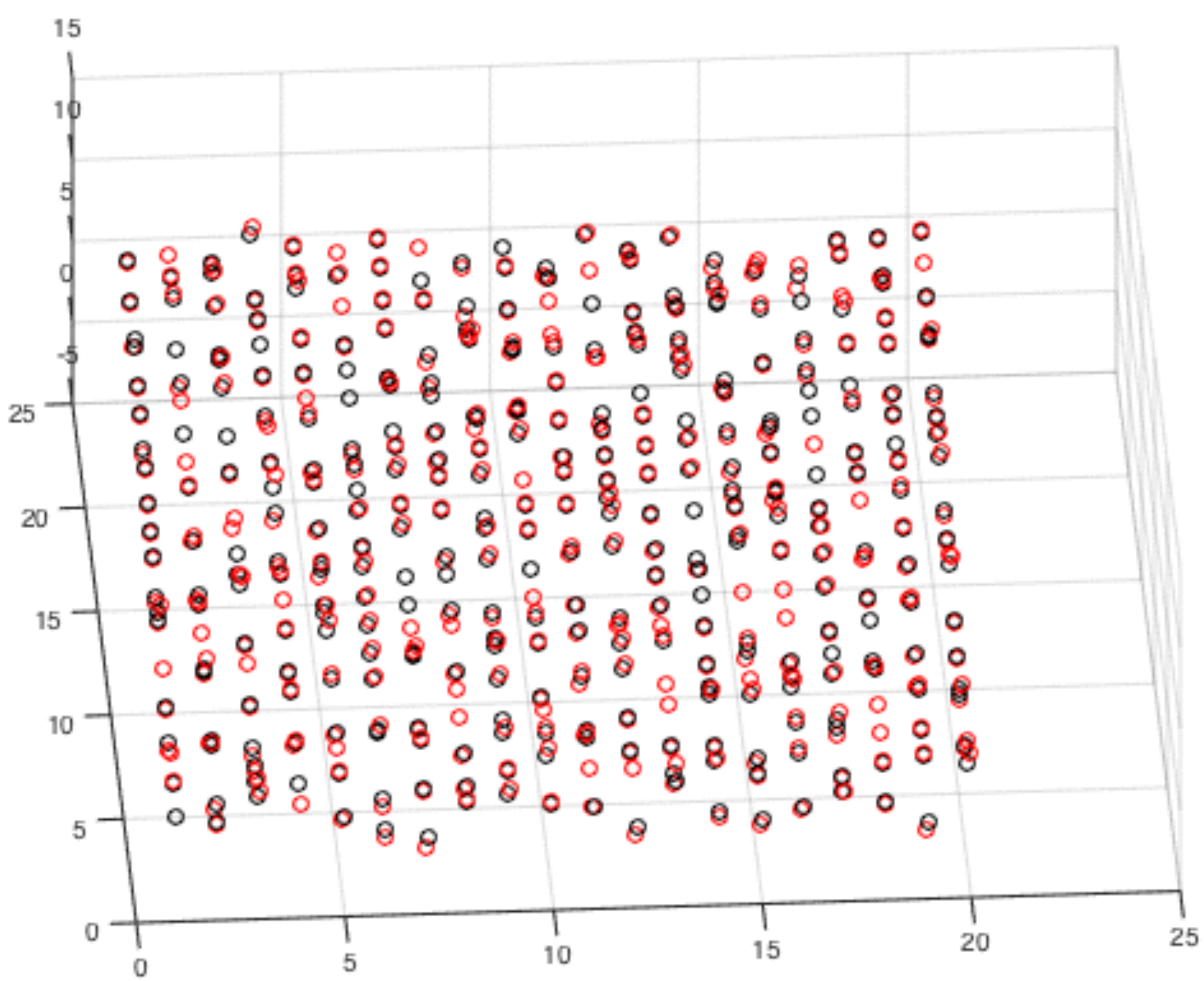
$B. \quad \Sigma = \frac{1}{n}\sum_{t=1}^{n} \mathbf{x}_t \mathbf{x}_t^\top$

$C. \quad \Sigma = XX^\top$

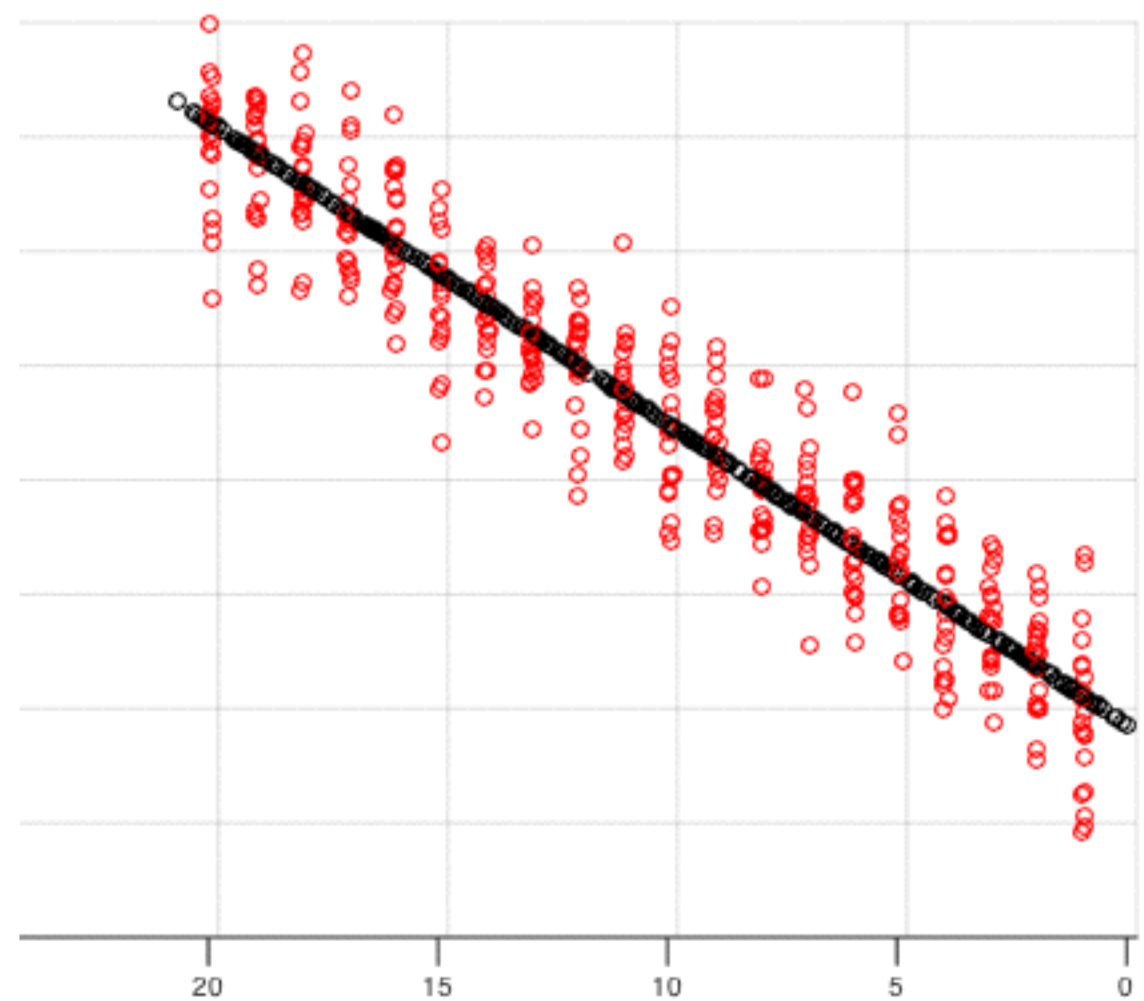
$D. \quad \Sigma = XX^\top$

Example:
Students in classroom

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

WHEN $d >> n$

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

$$V^{\top}V = I$$

$$X - \mu = UDV^{\top} \qquad U^{\top}U = I$$

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

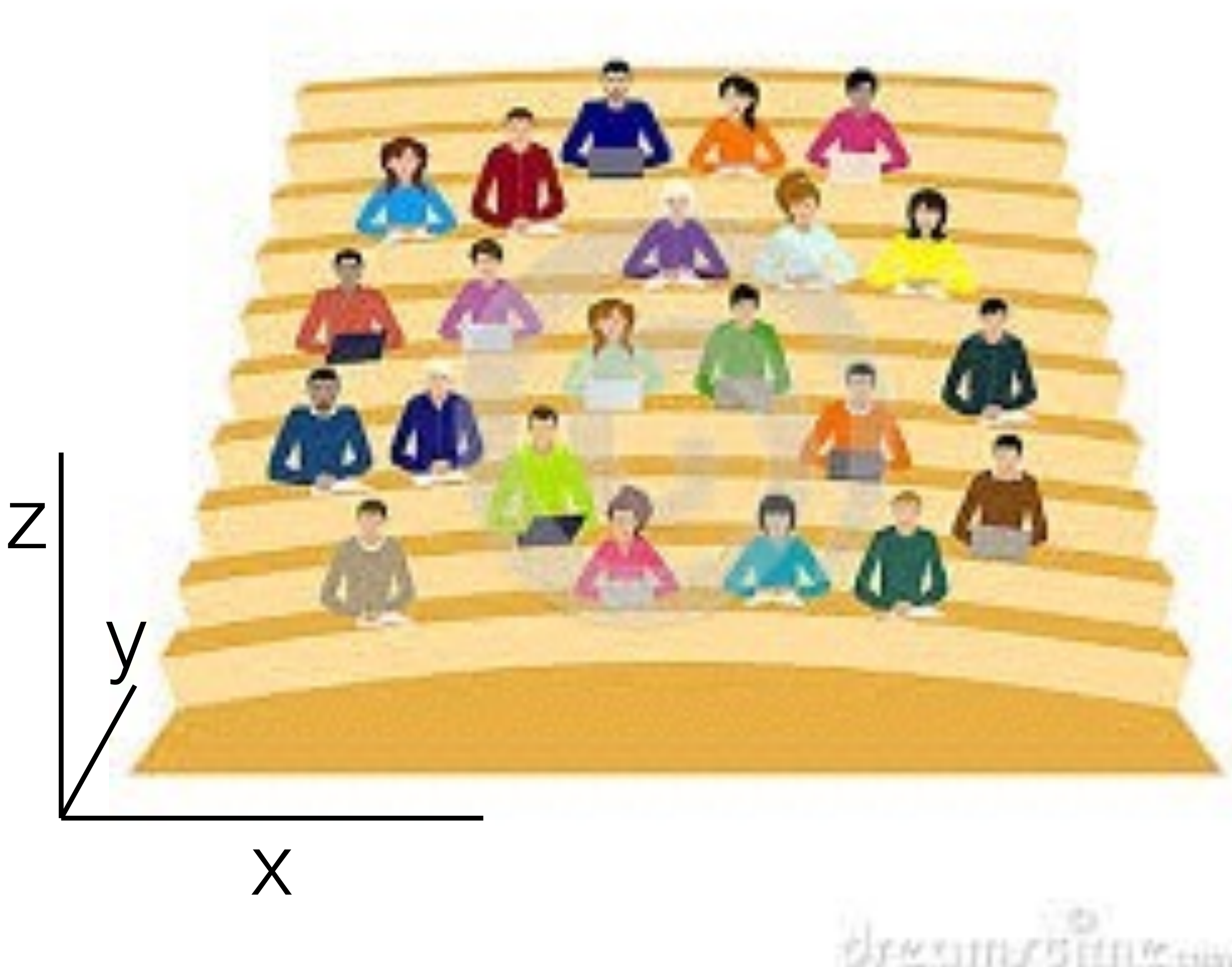
- Hence, matrix V is the same as matrix W got from eigen decomposition of Σ, eigenvalues are diagonal elements of D^2
- Alternative algorithm:

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

When to use PCA?

- When data naturally lies in a low dimensional linear subspace
- To minimize reconstruction error
- Find directions where data is maximally spread

Canonical Correlation Analysis



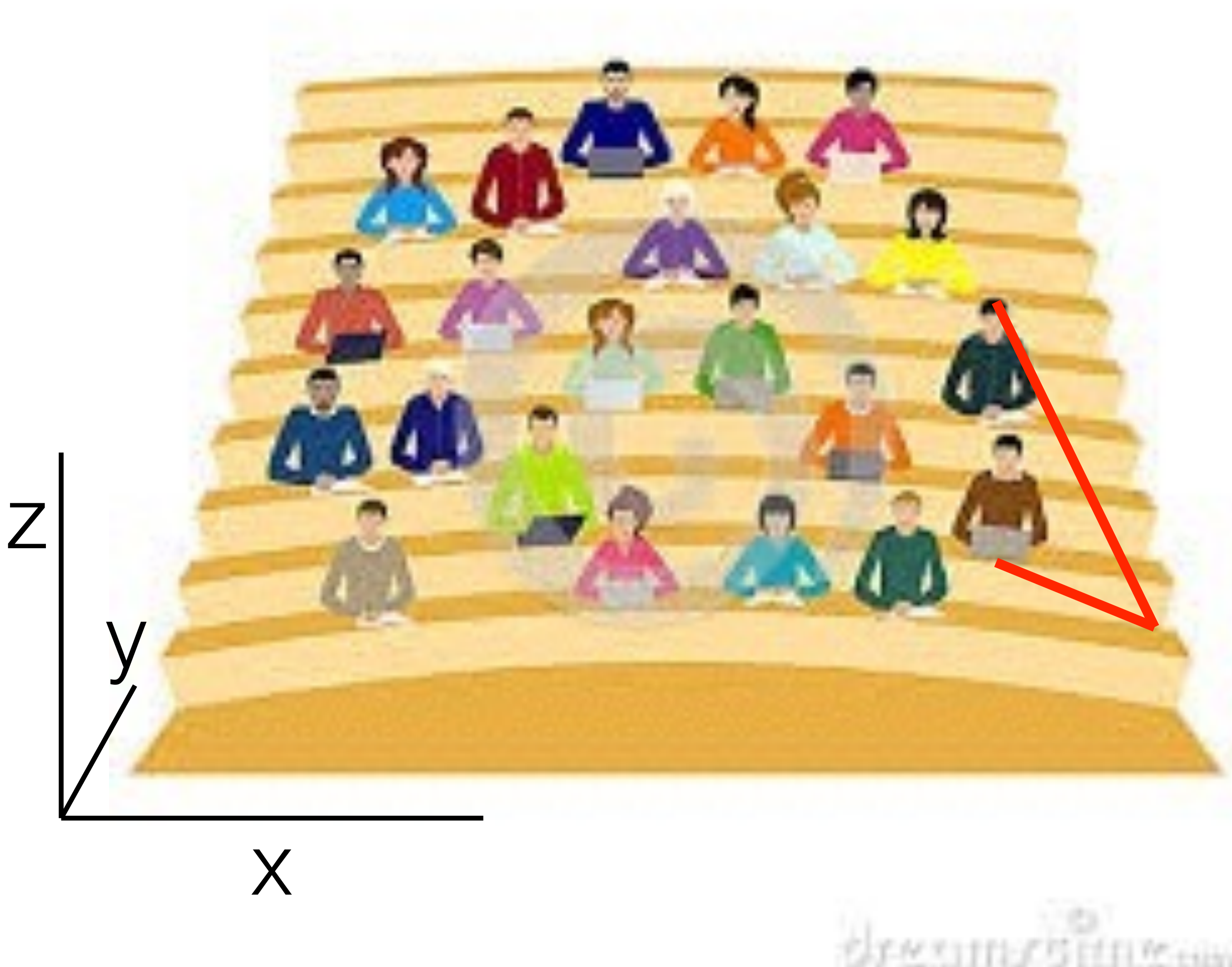
+

Age

Gender

Angle

Canonical Correlation Analysis



+

Age

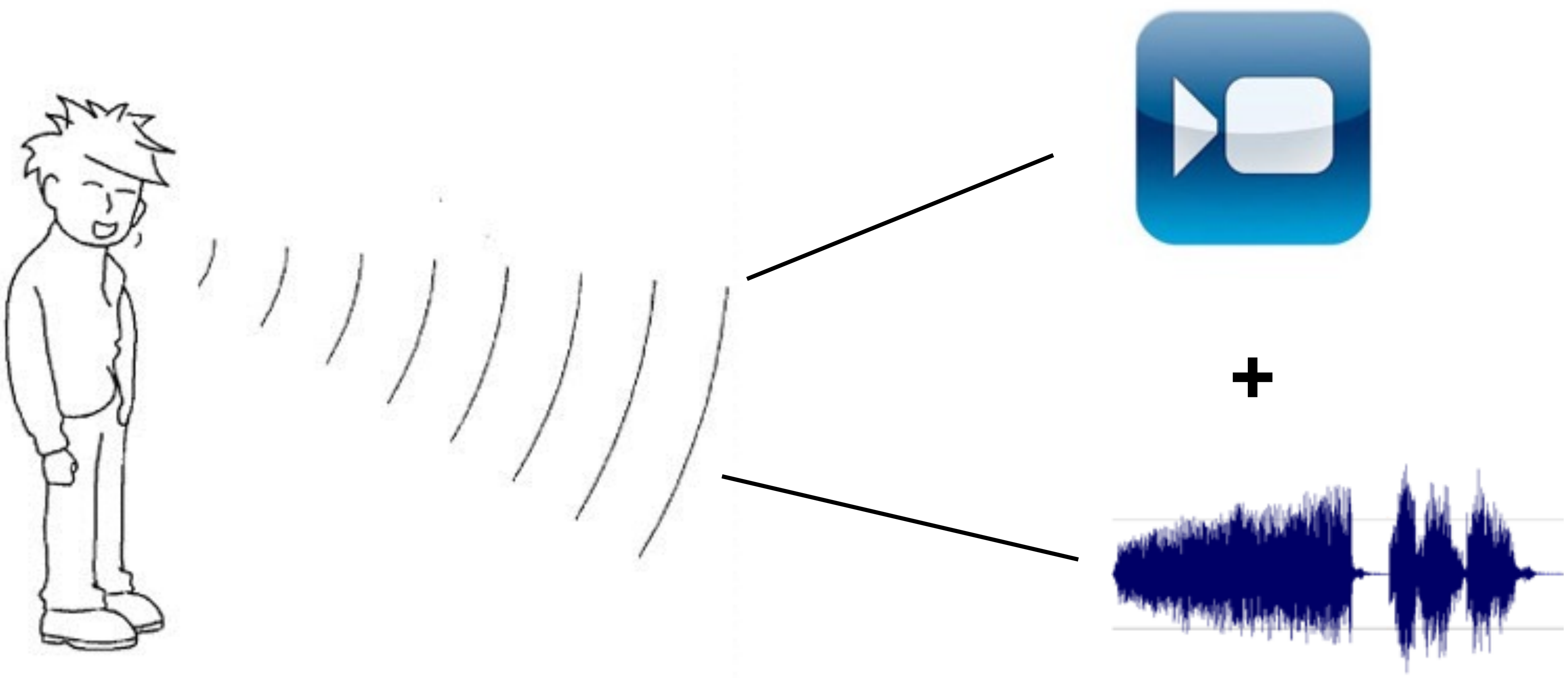
Gender

Angle

Two View Dimensionality Reduction

- Data comes in pairs $(\mathbf{x}_1, \mathbf{x}'_1), \ldots, (\mathbf{x}_n, \mathbf{x}'_n)$ where $\mathbf{x}_t$'s are d dimensional and $\mathbf{x}'_t$'s are d' dimensional
- Goal: Compress say view one into $\mathbf{y}_1, \ldots, \mathbf{y}_n$, that are K dimensional vectors
  - Retain information redundant between the two views
  - Eliminate "noise" specific to only one of the views

Example I: Speech Recognition



- Audio might have background sounds uncorrelated with video
- Video might have lighting changes uncorrelated with audio
- Redundant information between two views: the speech

EXAMPLE II: COMBINING FEATURE EXTRACTIONS

- Method A and Method B are both equally good feature extraction techniques
- Concatenating the two features blindly yields large dimensional feature vector with redundancy
- Applying techniques like CCA extracts the key information between the two methods
- Removes extra unwanted information

How do we get the right direction? (say K = 1)

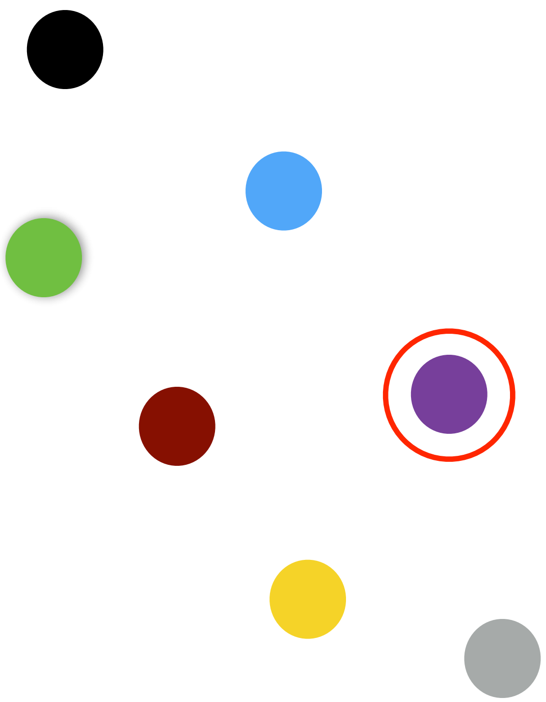
+

Age

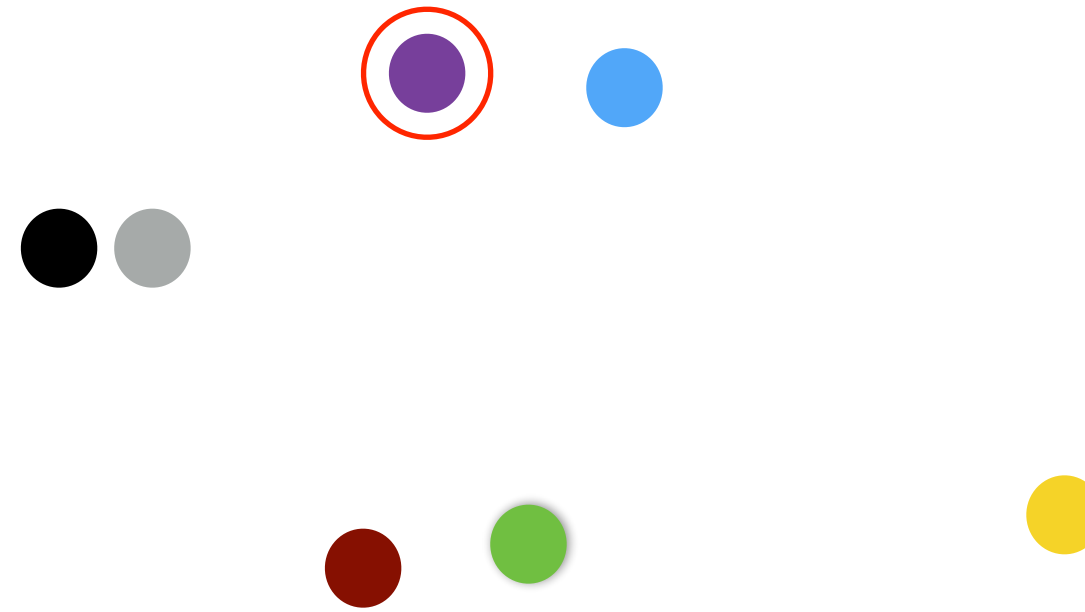
Gender

Angle

Which Direction to Pick?

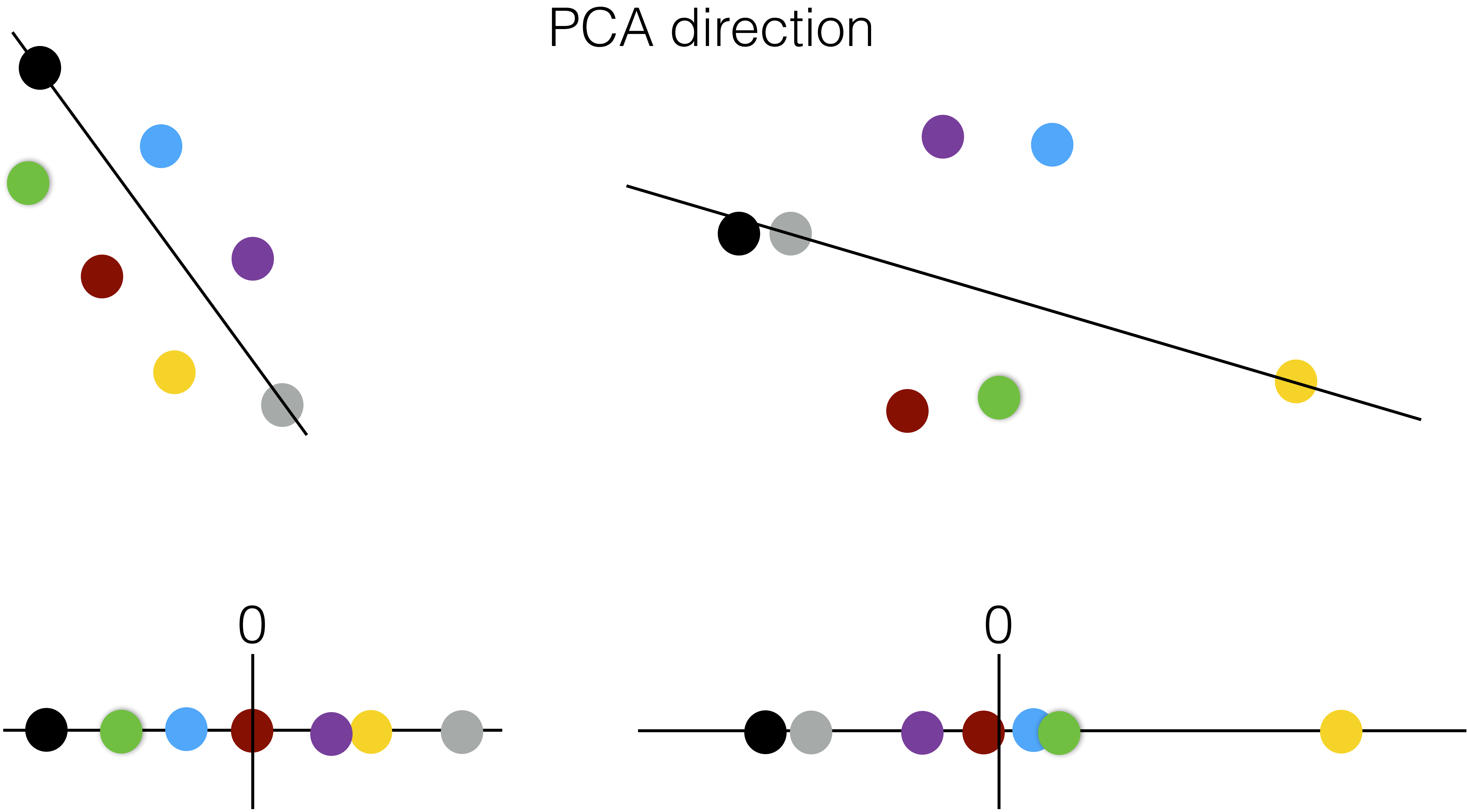


View I

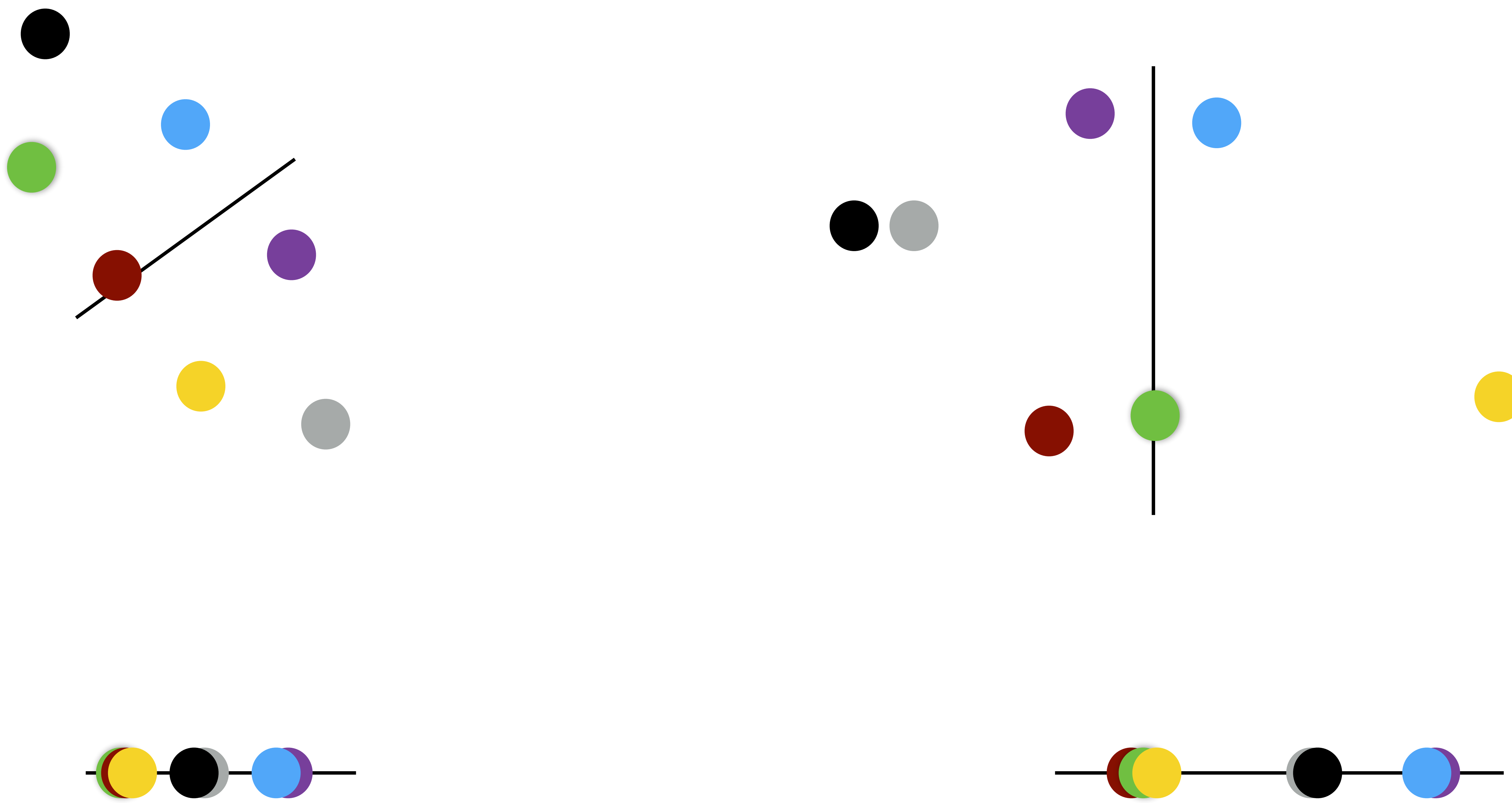


View II

Which Direction to Pick?



Which Direction to Pick?



Direction has large covariance

How do we pick the right direction to project to?

Maximizing Correlation Coefficient

- Say $\mathbf{w}_1$ and $\mathbf{v}_1$ are the directions we choose to project in views 1 and 2 respectively we want these directions to maximize,

$$\frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{y}_t[1]-\frac{1}{n}\sum_{t=1}^{n}\mathbf{y}_t[1]\right)\cdot\left(\mathbf{y}_t'[1]-\frac{1}{n}\sum_{t=1}^{n}\mathbf{y}_t'[1]\right)$$

where $\mathbf{y}_t[1] = \mathbf{w}_1^\top \mathbf{x}_t$ and $\mathbf{y}_t'[1] = \mathbf{v}_1^\top \mathbf{x}_t'$

Maximizing Correlation Coefficient

- Say $\mathbf{w}_1$ and $\mathbf{v}_1$ are the directions we choose to project in views 1 and 2 respectively we want these directions to maximize,

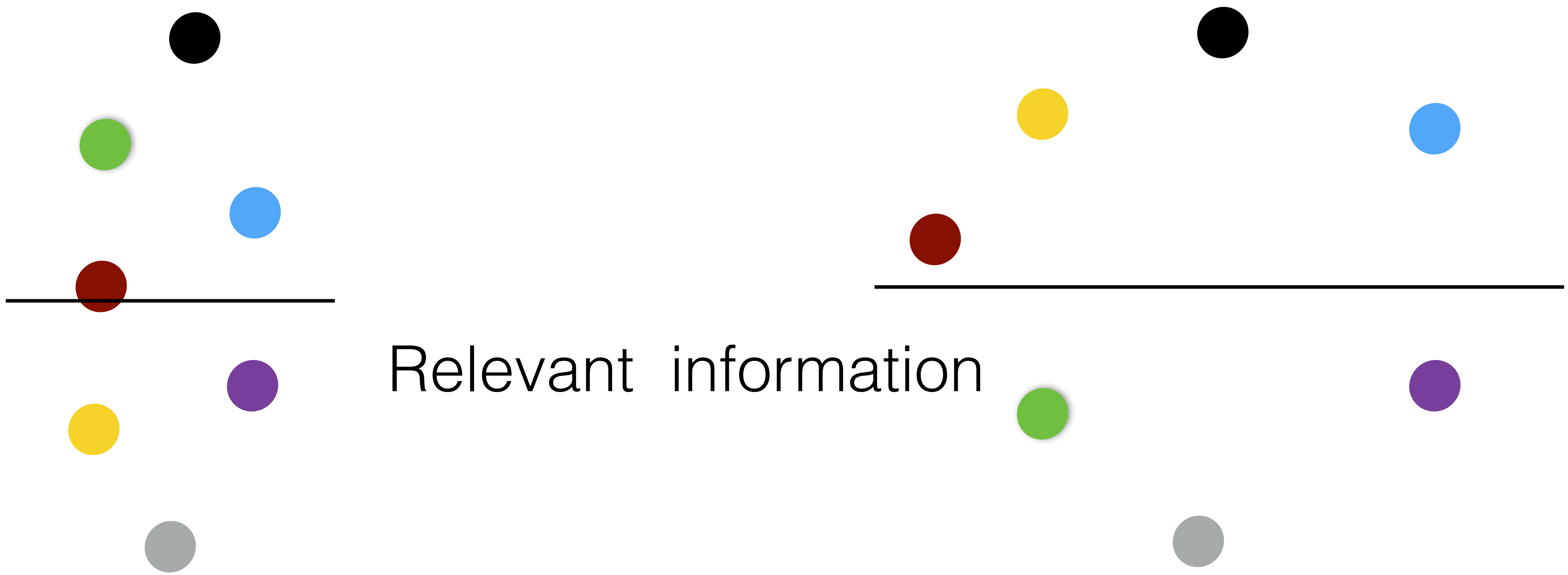
$$\frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{y}_t[1] - \frac{1}{n}\sum_{t=1}^{n}\mathbf{y}_t[1]\right)\cdot\left(\mathbf{y}'_t[1] - \frac{1}{n}\sum_{t=1}^{n}\mathbf{y}'_t[1]\right)$$

$$\text{s.t. } \tfrac{1}{n}\textstyle\sum_{t=1}^{n}\left(\mathbf{y}_t[1] - \tfrac{1}{n}\sum_{t=1}^{n}\mathbf{y}_t[1]\right)^2 = \tfrac{1}{n}\sum_{t=1}^{n}\left(\mathbf{y}'_t[1] - \tfrac{1}{n}\sum_{t=1}^{n}\mathbf{y}'_t[1]\right) = 1$$

where $\mathbf{y}_t[1] = \mathbf{w}_1^\top \mathbf{x}_t$ and $\mathbf{y}'_t[1] = \mathbf{v}_1^\top \mathbf{x}'_t$

What is the problem with the above?

Why not Maximize Covariance



$$\text{Say } \frac{1}{n}\sum_{t=1}^{n} \mathbf{x}_t[2] \cdot \mathbf{x'}_t[2] > 0$$

Scaling up this coordinate we can blow up covariance

Maximizing Correlation Coefficient

- Say $\mathbf{w}_1$ and $\mathbf{v}_1$ are the directions we choose to project in views 1 and 2 respectively we want these directions to maximize,

$$\frac{\frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{y}_t[1]-\frac{1}{n}\sum_{t=1}^{n}\mathbf{y}_t[1]\right)\cdot\left(\mathbf{y}'_t[1]-\frac{1}{n}\sum_{t=1}^{n}\mathbf{y}'_t[1]\right)}{\sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{y}_t[1]-\frac{1}{n}\sum_{t=1}^{n}\mathbf{y}_t[1]\right)^2}\sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{y}'_t[1]-\frac{1}{n}\sum_{t=1}^{n}\mathbf{y}'_t[1]\right)}}$$

Basic Idea of CCA

- Normalize variance in chosen direction to be constant (say 1)
- Then maximize covariance
- This is same as maximizing “correlation coefficient”

COVARIANCE VS CORRELATION

- $\text{Covariance}(A, B) = \mathbb{E}[(A - \mathbb{E}[A]) \cdot (B - \mathbb{E}[B])]$

  Depends on the scale of A and B. If B is rescaled, covariance shifts.

- $\text{Corelation}(A, B) = \frac{\mathbb{E}[(A-\mathbb{E}[A])\cdot(B-\mathbb{E}[B])]}{\sqrt{\text{Var}(A)}\sqrt{\text{Var}(B)}}$

  Scale free.

Maximizing Correlation Coefficient

- Say $\mathbf{w}_1$ and $\mathbf{v}_1$ are the directions we choose to project in views 1 and 2 respectively we want these directions to maximize,

$$\frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{y}_t[1] - \frac{1}{n}\sum_{t=1}^{n}\mathbf{y}_t[1]\right)\cdot\left(\mathbf{y}'_t[1] - \frac{1}{n}\sum_{t=1}^{n}\mathbf{y}'_t[1]\right)$$

$$\text{s.t. } \tfrac{1}{n}\textstyle\sum_{t=1}^{n}\left(\mathbf{y}_t[1] - \tfrac{1}{n}\sum_{t=1}^{n}\mathbf{y}_t[1]\right)^2 = \tfrac{1}{n}\sum_{t=1}^{n}\left(\mathbf{y}'_t[1] - \tfrac{1}{n}\sum_{t=1}^{n}\mathbf{y}'_t[1]\right) = 1$$

where $\mathbf{y}_t[1] = \mathbf{w}_1^\top \mathbf{x}_t$ and $\mathbf{y}'_t[1] = \mathbf{v}_1^\top \mathbf{x}'_t$

Canonical Correlation Analysis

- Hence we want to solve for projection vectors $\mathbf{w}_1$ and $\mathbf{v}_1$ that

$$\text{maximize } \frac{1}{n}\sum_{t=1}^{n} \mathbf{w}_1^\top(\mathbf{x}_t - \mu) \cdot \mathbf{v}_1^\top(\mathbf{x}_t' - \mu')$$
$$\text{subject to } \frac{1}{n}\sum_{t=1}^{n} (\mathbf{w}_1^\top(\mathbf{x}_t - \mu))^2 = \frac{1}{n}\sum_{t=1}^{n} (\mathbf{v}_1^\top(\mathbf{x}_t' - \mu'))^2 = 1$$

where $\mu = \frac{1}{n}\sum_{t=1}^{n} \mathbf{x}_t$ and $\mu' = \frac{1}{n}\sum_{t=1}^{n} \mathbf{x}_t'$

Canonical Correlation Analysis

- Hence we want to solve for projection vectors $\mathbf{w}_1$ and $\mathbf{v}_1$ that

$$\text{maximize } \frac{1}{n}\sum_{t=1}^{n} \mathbf{w}_1^\top(\mathbf{x}_t - \mu) \cdot \mathbf{v}_1^\top(\mathbf{x}_t' - \mu')$$
$$\text{subject to } \frac{1}{n}\sum_{t=1}^{n} (\mathbf{w}_1^\top(\mathbf{x}_t - \mu))^2 = \frac{1}{n}\sum_{t=1}^{n} (\mathbf{v}_1^\top(\mathbf{x}_t' - \mu'))^2 = 1$$

where $\mu = \frac{1}{n}\sum_{t=1}^{n} \mathbf{x}_t$ and $\mu' = \frac{1}{n}\sum_{t=1}^{n} \mathbf{x}_t'$

Canonical Correlation Analysis

- Hence we want to solve for projection vectors $\mathbf{w}_1$ and $\mathbf{v}_1$ that

$$\begin{aligned}&\text{maximize } \mathbf{w}_1^\top \Sigma_{1,2} \mathbf{v}_1\\ &\text{subject to } \mathbf{w}_1^\top \Sigma_{1,1} \mathbf{w}_1 = \mathbf{v}_1^\top \Sigma_{2,2} \mathbf{v}_1 = 1\end{aligned}$$

$$\Sigma = \begin{matrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{matrix} = \mathrm{cov}\left(X \; X' \right)$$

Solution

$$W_1 = \mathrm{eigs}\left(\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}, K\right)$$

$$W_2 = \mathrm{eigs}\left(\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}, K\right)$$

CCA Algorithm

1. $X = \left({}^{n}\underset{d_1}{X_1}, \underset{d_2}{X_2} \right)$

2. $\Sigma = \begin{matrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{matrix} = \mathrm{cov}\left(X \right)$

3. $W_1 = \mathrm{eigs}\left(\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}, K \right)$

4. $Y_1 = X_1 - \mu_1 \times W_1$