Machine Learning for Data Science (CS4786) Lecture 4

Canonical Correlation Analysis (CCA)

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

PCA: Variance Maximization

- 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

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

Covariance Vs Correlation

	Bread A	Bread B	Bread C	Bread D	Butter	Margarine	Soda
Store 1	944	896	1109	1074	11	79	6008
Store 2	953	950	1106	1071	12	77	9117
Store 3	967	976	1101	1054	17	70	6805
Store 4	969	1008	1079	1052	21	69	7306
Store 5	977	1024	1057	1020	27	63	5550
Store 6	1007	1038	1050	996	28	62	8907
Store 7	1019	1040	1043	996	30	61	7267
Store 8	1053	1055	962	973	34	59	6485
Store 9	1071	1096	922	967	35	56	6792
Store 10	1074	1097	896	935	36	54	7412

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.

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

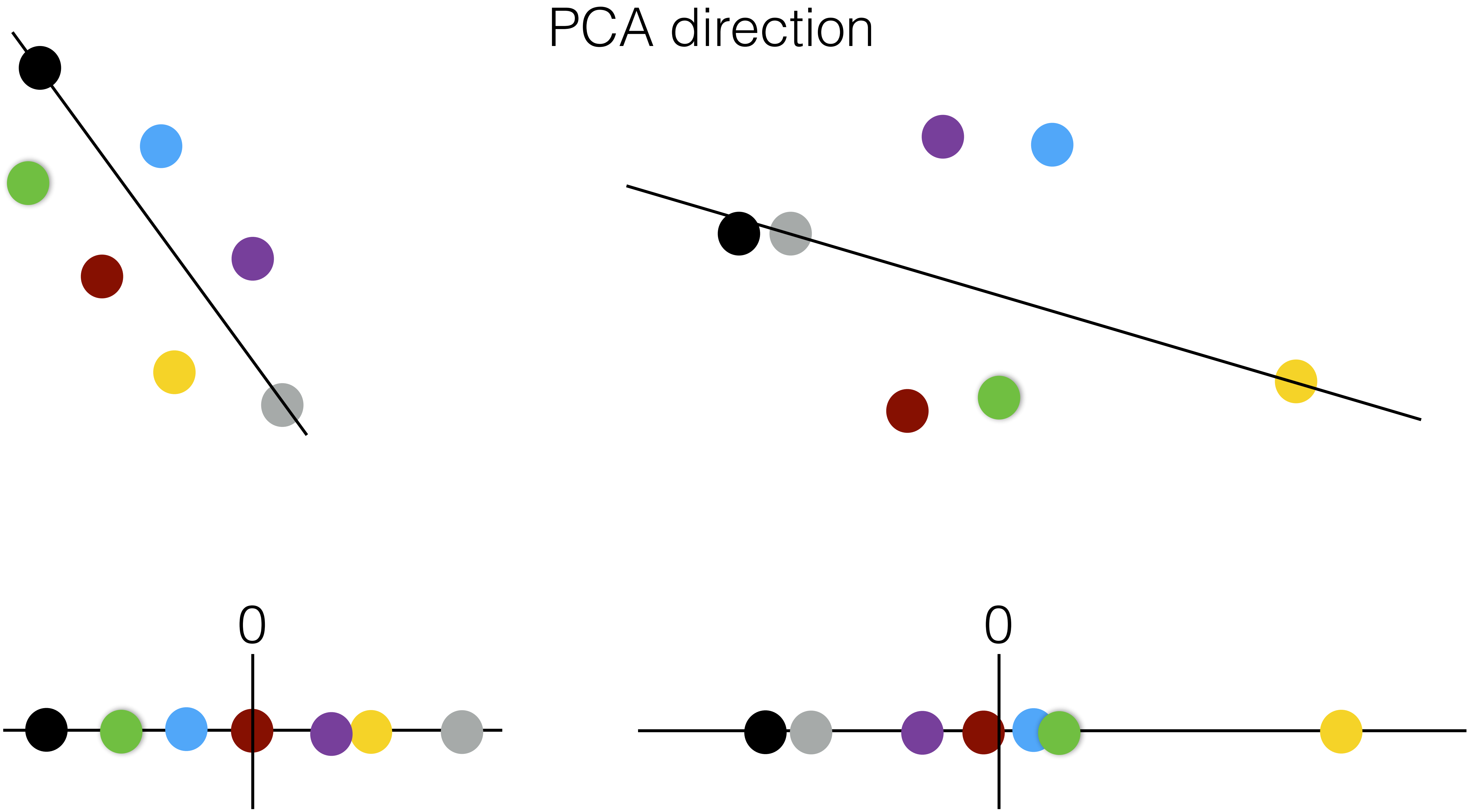
  Scale free.

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

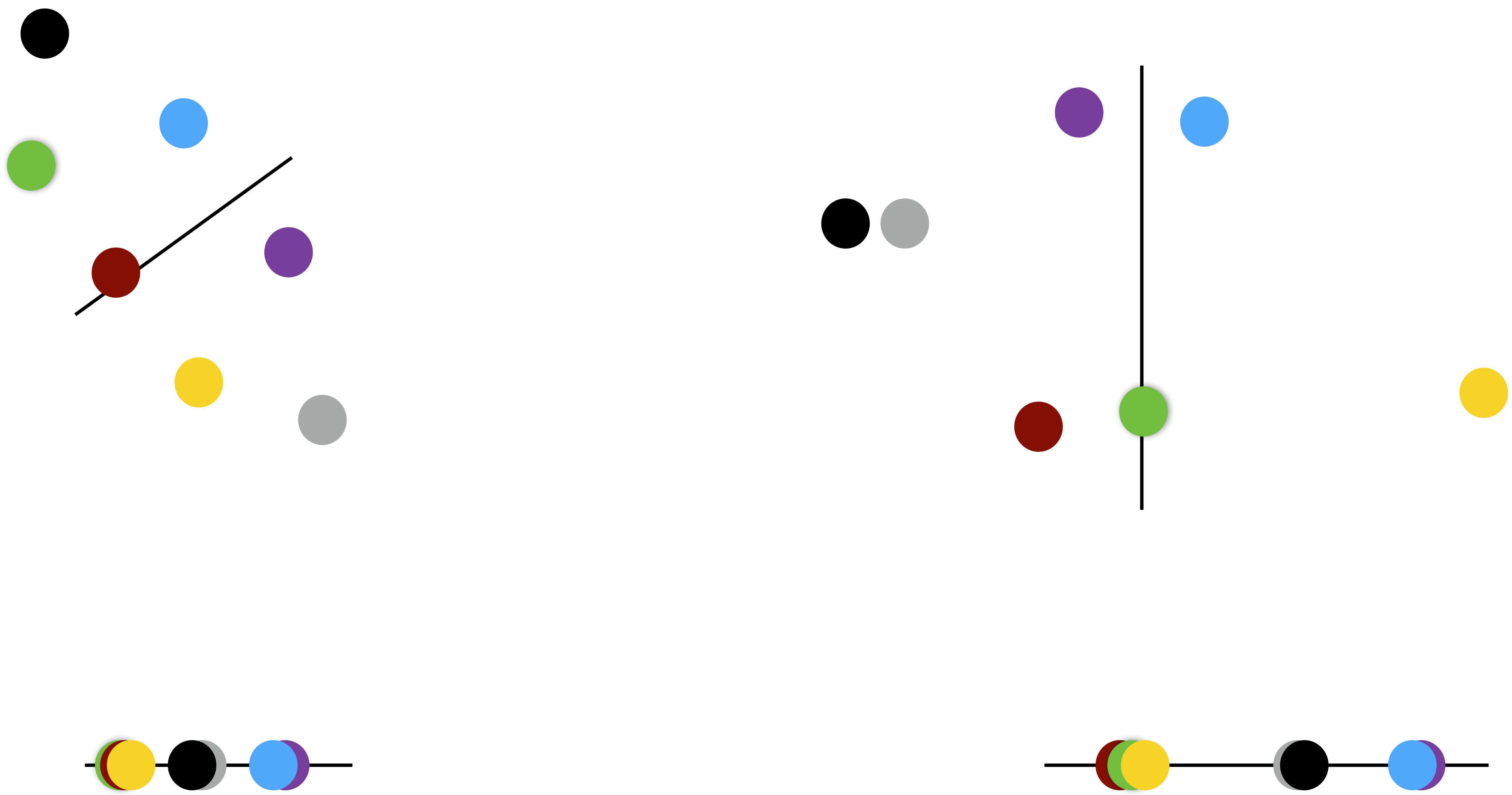
- Goal: Compress say view one into $\mathbf{y}_1, \ldots, \mathbf{y}_n$, that are K dimensional vectors
  - Retain information redundant between the two views
  - Eliminate "noise" specific to only one of the views

WHICH DIRECTION TO PICK?



$$\text{Average dot product} = \text{covariance small}$$

Which Direction to Pick?



Direction has large correlation

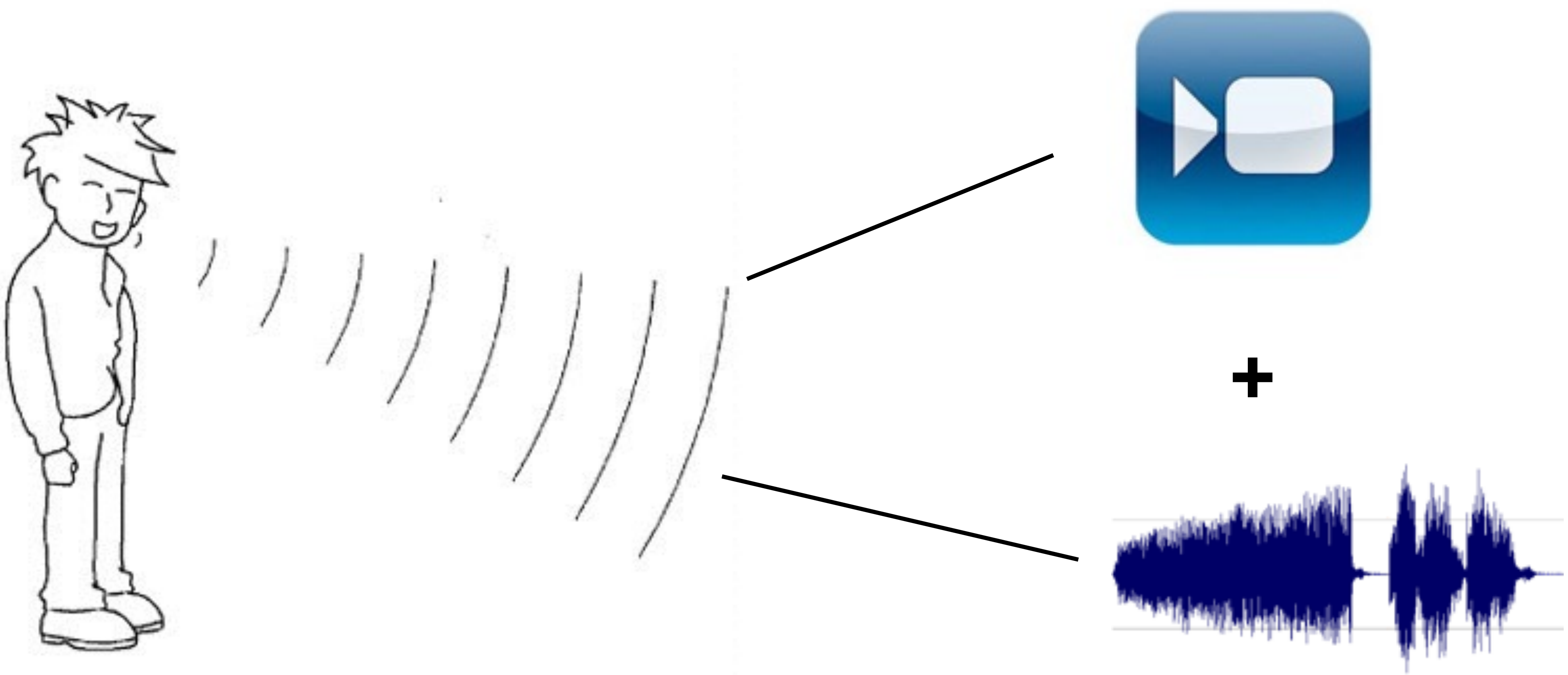
Basic Idea of CCA

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

When to use CCA?

- When we have redundancy in data.
- When the relevant information is part of the redundancy
- Same data point from two different view/sources

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

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

- Assume data in both views are centered : $\frac{1}{n}\sum_{t=1}^{n}\mathbf{x}_t = \mathbf{0}$, $\frac{1}{n}\sum_{t=1}^{n}\mathbf{x}'_t = \mathbf{0}$
Hence $\frac{1}{n}\sum_{t=1}^{n}\mathbf{y}'_t[1] = \frac{1}{n}\sum_{t=1}^{n}\mathbf{y}_t[1] = 0$

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

$$\text{maximize } \frac{1}{n}\sum_{t=1}^{n}\mathbf{y}_t[1]\cdot\mathbf{y}'_t[1]$$

$$\text{subject to } \frac{1}{n}\sum_{t=1}^{n}(\mathbf{y}_t[1])^2 = \frac{1}{n}\sum_{t=1}^{n}(\mathbf{y}'_t[1])^2 = 1$$

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

$$\text{maximize } \frac{1}{n}\sum_{t=1}^{n}\mathbf{w}_1^\top\mathbf{x}_t\cdot\mathbf{v}_1^\top\mathbf{x}_t'$$

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

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

$$\text{maximize } \frac{1}{n}\sum_{t=1}^{n} \mathbf{w}_1^\top \mathbf{x}_t \mathbf{x}_t'^\top \mathbf{v}_1$$

$$\text{subject to } \frac{1}{n}\sum_{t=1}^{n} \mathbf{w}_1^\top \mathbf{x}_t \mathbf{x}_t^\top \mathbf{w}_1 = \frac{1}{n}\sum_{t=1}^{n} \mathbf{v}_1^\top \mathbf{x}_t' \mathbf{x}_t'^\top \mathbf{v}_1 = 1$$

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

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

- Writing Lagrangian taking derivative equating to 0 we get

$$\Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1}\mathbf{w}_1 = \lambda^2 \Sigma_{1,1}\mathbf{w}_1 \quad \text{and} \quad \Sigma_{2,1}\Sigma_{1,1}^{-1}\Sigma_{1,2}\mathbf{v}_1 = \lambda^2 \Sigma_{2,2}\mathbf{v}_1$$

or equivalently

$$\left(\Sigma_{1,1}^{-1}\Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1}\right)\mathbf{w}_1 = \lambda^2 \mathbf{w}_1 \quad \text{and} \quad \left(\Sigma_{2,2}^{-1}\Sigma_{2,1}\Sigma_{1,1}^{-1}\Sigma_{1,2}\right)\mathbf{v}_1 = \lambda^2 \mathbf{v}_1$$

CCA Algorithm

1. $$X = \left(n \begin{array}{cc} \underbrace{X_1}_{d_1}, & \underbrace{X_2}_{d_2} \end{array} \right)$$

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

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

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

CCA Demo