

# Multi-view Learning

Vijaya Saradhi

**IIT Guwahati**

03<sup>rd</sup> July 2018

# Outline

- What are views?
- Examples
- Definitions
  - Supervised learning
  - Unsupervised learning
- Data
  - Single view
  - Multi view
- Dimensionality reduction
  - Single view
    - Principal Component Analysis (*unsupervised*)
    - Linear Discriminant Analysis (*supervised*)
  - Multi view
    - Canonical Correlation Analysis (*unsupervised*)
    - Linear Discriminant Analysis (*supervised*)
- An interesting application - *TM estimation and anomaly detection*

# What are views?

## Definition

### Oxford dictionary meanings

- Ability to see something or to be seen from a particular place
  - A particular way of considering or regarding something; an attitude or opinion
- 
- Data acquired from one sensor - forms a single view of *data*
  - Data acquired from **one sensor** with **multiple perspectives** - forms **multiple views** of *data*
  - Data **derived** by observing **single sensor** data from **multiple** perspectives
  - Data acquired using **multiple feature extractors** from **one sensor** - forms **multiple views** of *data*
  - Data acquired from random sampling from a given distribution - multiple views of *data*?
  - Data acquired from **multiple sensors** - forms **multiple views** of *data*

# Example 1 - One sensor multiple perspectives

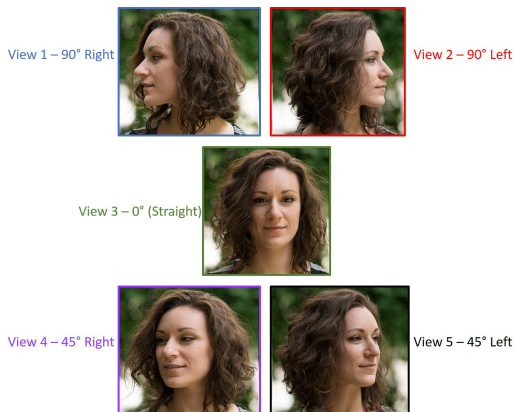


Figure 1: Pose variations. source: unknown

# Example 2 - One sensor and multiple derivations

source: google images

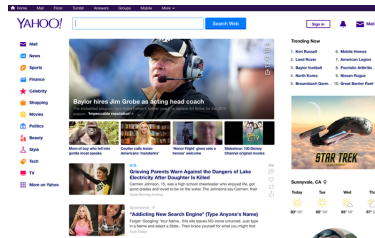
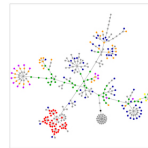
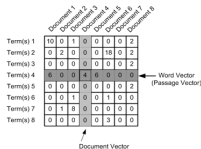


Figure 2: Single sensor data



## Example 3 - One sensor and multiple features

source: google images



Figure 5: Original Image

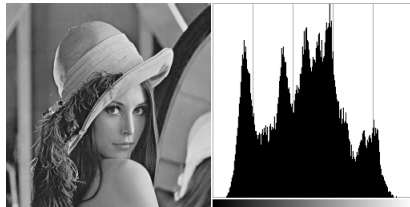


Figure 6: Histogram features

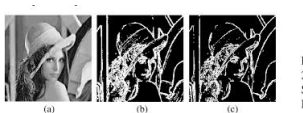


Figure 7: Edge features

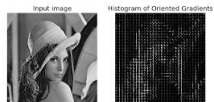


Figure 8: Histogram Oriented Gradient features

## Example 4 - Random Sampling

Six datasets

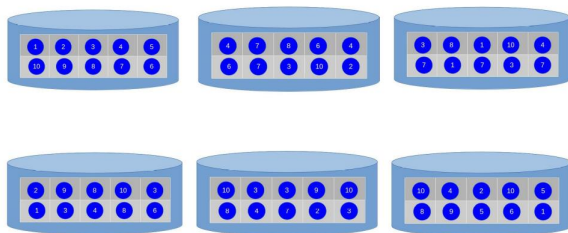


Figure 9: Six random samples of the original dataset

## Example 5a - Multiple Sensors

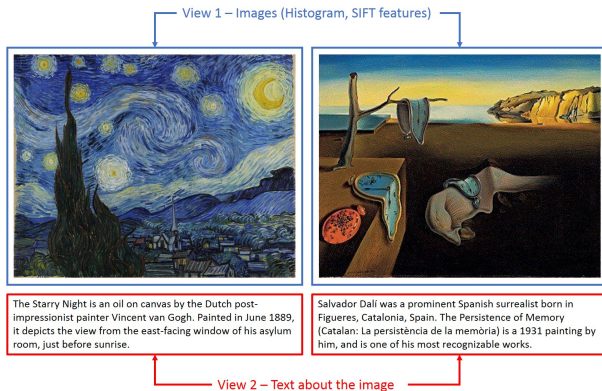


Figure 10: Image and Text. source: unknown



# Example 5b - Multiple Authors

<p>View 1 - English</p> <p><b>Chapter 1</b></p> <p><b>Mr. Sherlock Holmes</b></p> <p>Mr. Sherlock Holmes, who was usually very late in the mornings, save upon those not infrequent occasions when he was up all night, was seated at the breakfast table. I stood upon the hearth-rug and picked up the stick which our visitor had left behind him the night before. It was a fine, thick piece of wood, bulbous-headed, of the sort which is known as a "Penang lawyer." Just under the head was a broad silver band nearly an inch across. "To James Mortimer, M.R.C.S., from his friends of the C.C.H.," was engraved upon it, with the date "1884." It was just such a stick as the old-fashioned family practitioner used to carry—dignified, solid, and reassuring.</p>	<p>View 2 - German</p> <p><b>Kapitel 1</b></p> <p><b>Mr. Sherlock Holmes</b></p> <p>Mr. Sherlock Holmes, der morgens sehr spät aufzustehen pflegte – wenn man einmal von jenen nicht seltenen Gelegenheiten absah, da er die ganze Nacht auflebte –, saß am Frühstückstisch, während ich auf dem Kaminvorleger stand und den Spazierstock aufhob, den unser Besucher vergangene Nacht vergessen hatte. Es handelte sich um ein schönes, stabiles Stück Holz mit einem knollenförmigen Griff derjenigen Sorte, die bei uns unter dem Namen „Penang-Anwalt“ bekannt ist. Gleich unterhalb des Griffes war ein gut zwei Zentimeter breites Silberband befestigt, auf welchem die Widmung „Für James Mortimer, M. R. C. S., von seinen Freunden des C. C. H.“ eingraviert war, datiert auf das Jahr 1884; ein Stock, gerade wie ihn ein altmodischer Hausarzt getragen hätte: würdevoll, solide und zuverlässig.</p>
<p>View 3 - Spanish</p> <p><b>CAPÍTULO UNO</b></p> <p><b>EL SEÑOR SHERLOCK HOLMES</b></p> <p>El señor Sherlock Holmes, que de ordinario se levantaba muy tarde, excepto en las ocasiones nada infrecuentes en que no se acostaba en toda la noche, estaba desayunando. Yo, que me hallaba de pie junto a la chimenea, me agaché para recoger el bastón olvidado por nuestro visitante de la noche anterior. Sólido, de madera de buena calidad y con un abutamiento a modo de empuñadura, era del tipo que se conoce como «abogado de Penang». Inmediatamente debajo de la protuberancia el bastón llevaba una ancha tira de plata, de más de dos centímetros, en la que estaba grabado «A James Mortimer, M.R.C.S., de sus amigos de C.C.H.», y el año, «1884». Era exactamente la clase de bastón que solían llevar los médicos de cabecera a la antigua usanza: digno, sólido y que inspiraba confianza.</p>	<p>View 4 - French</p> <p><b>CHAPITRE I</b></p> <p><b>M. SHERLOCK HOLMES</b></p> <p>Ce matin-là, M. Sherlock Holmes qui, sauf les cas assez fréquents où il passait les nuits, se levait tard, était assis devant la table de la salle à manger. Je me tenais près de la cheminée, examinant la canne que notre visiteur de la veille avait oubliée. C'était un joli bâton, solide, terminé par une boule – ce qu'on est convenu d'appeler « une permission de minuit ». Immédiatement au-dessous de la pomme, un cercle d'or, large de deux centimètres, portait l'inscription et la date suivantes : « À M. James Mortimer, ses amis du C. C. H. — 1884 ». Cette canne, digne, grave, rassurante, ressemblait à celles dont se servent les médecins « vieux jeu ».</p>

Figure 11: Different languages. source: unknown

## Example 5c - Multiple Sensors and Multiple Authors

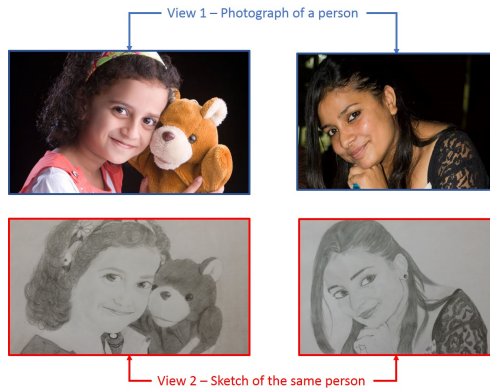


Figure 12: Photograph and Sketch. source: unknown

# Single view

Data must be generated from

- One sensor
- One source
- One perspective
- One distribution
- Holds the assumption *i.i.d*

Any one view of the above examples forms single view.

## Note on multiple views

- Given a data point in one view corresponding data point in other views is assumed to be known.

View	$\mathbf{x}_1$	$\mathbf{x}_2$	$\cdots$	$\mathbf{x}_n$
1	$\mathbf{x}_1^1$	$\mathbf{x}_2^1$	$\cdots$	$\mathbf{x}_n^1$
2	$\mathbf{x}_1^2$	$\mathbf{x}_2^2$	$\cdots$	$\mathbf{x}_n^2$
3	$\mathbf{x}_1^3$	$\mathbf{x}_2^3$	$\cdots$	$\mathbf{x}_n^3$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
$v$	$\mathbf{x}_1^v$	$\mathbf{x}_2^v$	$\cdots$	$\mathbf{x}_n^v$
label	$y_1$	$y_2$	$\cdots$	$y_n$

- The above assumption **does not hold** for random sampling method.
- Every  $\mathbf{x}_j$  must be present in all the vies. That is **there should not exist a view  $i$**  such that  $\mathbf{x}_j$  is present in every view except  $i^{th}$  view.
- Multi-view dataset is represented as:  $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$  where  $\mathbf{x}_i = (\mathbf{x}_i^1, \mathbf{x}_i^2, \cdots, \mathbf{x}_i^v)$

# What is Multi-view learning?

- Building *multiple learners* on *a dataset* and combining their decisions
- Building *multiple learners* on *multiple views of a dataset* and combining their decisions
- Underlying observation: The use of multiple learners in making decision.

# How to build multiple learners?

- Building one learning function on each view independently
- Formulate one objective function involving multiple views of data
- In this talk I will focus on the second bullet.

# Classification

## Given

$\{(\mathbf{x}_i, y_i)\}_{i=1}^n$  training set with  $(\mathbf{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y}$  drawn i.i.d from the distribution  $\mathcal{D}_{(\mathbf{x}, y)}$ .

## Find

a classification function  $f \in \mathcal{F} : \mathcal{X} \rightarrow \mathcal{Y}$  such that with high probability  $f(\mathbf{x}_{test}) = y_{test}$  for  $(\mathbf{x}_{test}, y_{test})$  independently drawn from  $\mathcal{D}_{(\mathbf{x}, y)}$ .

## ERM

The empirical risk minimization to find  $h$  solves

$$f = \operatorname{argmin}_{f \in \mathcal{F}} = \sum_{i=1}^n \{1 - \ell(f(\mathbf{x}_i), y_i)\} \quad (1)$$

where  $\ell(a, b) = 1$  if  $a \neq b$ ; 0 otherwise.

# Dimensionality Reduction

## Given

$X = \{\mathbf{x}_i\}_{i=1}^n$  where  $\mathbf{x}_i \in \mathbb{R}^p$  and a choice of dimensionality  $q < p$

## Optimize

An objective function  $f_X(\cdot)$  to produce a linear transformation  $P \in \mathbb{R}^{q \times p}$ .  
Transformed data points lie in low dimensional space as given by :  
 $PX \in \mathbb{R}^{q \times n}$ .

## Focus of this talk

- Single view
  - Principal component analysis
  - Linear discriminant analysis
- Multi view
  - Canonical correlation analysis
  - Linear discriminant analysis



# Single View - A Note

- One view of the data - that is  $\mathbf{x}_i$  - is utilized for learning  $f$  or  $P$ .
- One learning function is learned that is  $f$  or  $P$

# Multi-view

## View sufficiency

Each view is sufficient for the learning task at hand

## Compatibility

The target function of views predict the **same labels** with a high probability

# Multi-view

## Conditional Independence

Assume we have two views  $V_1$  and  $V_2$ . Associated learning functions are  $f^1$  and  $f^2$ . For any **fixed**  $(\hat{\mathbf{x}}^1, \hat{\mathbf{x}}^2)$ , the following holds with non-zero probability

$$Pr_{(\mathbf{x}^1, \mathbf{x}^2) \in \mathcal{D}} [\mathbf{x}^1 = \hat{\mathbf{x}}^1 | \mathbf{x}^2 = \hat{\mathbf{x}}^2] = Pr_{(\mathbf{x}^1, \mathbf{x}^2) \in \mathcal{D}} [\mathbf{x}^1 = \hat{\mathbf{x}}^1 | f^2(\mathbf{x}^2) = f^2(\hat{\mathbf{x}}^2)] \quad (2)$$

That is  $\mathbf{x}^1$  and  $\mathbf{x}^2$  are conditionally independent given the label.

# Single-view

- View sufficiency is met through feature engineering
- View effectiveness goes unquestioned
- Learning models are strengthened by varying
  - objective function
  - optimization methods
  - loss functions

# Multi-view Learning

## Consensus Principle

Aims to maximize the agreement on multiple distinct views. That is:

$$P(f^1(\mathbf{x}^1) \neq f^2(\mathbf{x}^2)) \geq \max \{P_{err}(f^1(\mathbf{x}^1)), P_{err}(f^2(\mathbf{x}^1))\} \quad (3)$$

and

$$\|f^1(\mathbf{x}_i^1) - f^2(\mathbf{x}_i^2)\| \leq \eta_i + \epsilon \quad (4)$$

## Complementary Principle

Each view of the data **may contain some knowledge that other views do not have**

# Important Definitions

## Mean vectors

Suppose  $\mathbf{X}' = [X_1, X_2, X_3, \dots, X_p]$  a  $p \times 1$ . random vector. Then each element of  $\mathbf{X}$  is a random variable. The marginal mean  $\mu_i$  is defined as  $\mu_i = E(X_i) \forall i = 1, 2, \dots, p$  is given as

$$\mu_i = \begin{cases} \int_{-\infty}^{\infty} x_i f_i(x_i) dx_i & \text{if } X_i \text{ is continuous RV} \\ \sum_{\text{all } x_i} x_i p_i(x_i) & \text{if } X_i \text{ is discrete RV} \end{cases}$$

# Important Definitions

## Mean vectors

Mean of a  $p \times 1$  random vector  $\mathbf{X}$  is given by.

$$E(\mathbf{X}) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} = \boldsymbol{\mu}$$

# Important Definitions

## Variance

The variance  $\sigma_i^2$  is defined as  $\sigma_i^2 = E(X_i - \mu_i)^2 \quad \forall i = 1, 2, \dots, p$  is given as

$$\sigma_i^2 = \begin{cases} \int_{-\infty}^{\infty} (x_i - \mu_i)^2 f_i(x_i) dx_i & \text{if } X_i \text{ is continuous RV} \\ \sum_{\forall x_i} (x_i - \mu_i)^2 p_i(x_i) & \text{if } X_i \text{ is discrete RV} \end{cases}$$

## Variance's Role

In principal component analysis variance play an important role. We elaborate on this soon.



# Important Definitions

## Covariance Matrix

The behavior of any pair of RVs,  $X_i, X_k$  is described by their joint probability function and measures linear association between them

$$\sigma_{ik} = \begin{cases} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_i - \mu_i)(x_k - \mu_k) f_{ik}(x_i, x_k) dx_i dx_k & \text{continuous RVs} \\ \sum_{\forall x_i} \sum_{\forall x_k} (x_i - \mu_i)(x_k - \mu_k) p_{ik}(x_i, x_k) & \text{discrete RVs} \end{cases}$$

# Important Definitions

## Covariance Matrix

Covariances of a  $p \times 1$  random vector  $\mathbf{X}$  is given by.

$$E(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})' = \begin{pmatrix} \begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \\ \vdots \\ X_p - \mu_p \end{bmatrix} & [X_1 - \mu_1, X_2 - \mu_2, \dots, X_p - \mu_p] \end{pmatrix}$$

# Important Definitions

## Covariance Matrix

Covariances of a  $p \times 1$  random vector  $\mathbf{X}$ ,  $\Sigma$  is given by.

$$E \begin{bmatrix} (X_1 - \mu_1)(X_1 - \mu_1) & (X_1 - \mu_1)(X_2 - \mu_2) & \cdots & (X_1 - \mu_1)(X_p - \mu_p) \\ (X_2 - \mu_2)(X_1 - \mu_1) & (X_2 - \mu_2)(X_2 - \mu_2) & \cdots & (X_2 - \mu_2)(X_p - \mu_p) \\ \vdots & \vdots & \ddots & \vdots \\ (X_p - \mu_p)(X_1 - \mu_1) & (X_p - \mu_p)(X_2 - \mu_2) & \cdots & (X_p - \mu_p)(X_p - \mu_p) \end{bmatrix}$$

# Important Definitions

## Covariance Matrix

Covariance matrix is given by

$$\begin{aligned}\Sigma &= E(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})' \\ &= \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}\end{aligned}$$

# Important Definitions

## Correlation Coefficient

Measures amount of **linear association** between two random variables  $X_i$  and  $X_k$ . Denoted by  $\rho_{ik}$  is defined in terms of covariance  $\sigma_{ik}$  and variances  $\sigma_{ii}$  and  $\sigma_{kk}$  is given as:

$$\rho_{ik} = \frac{\sigma_{ik}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{kk}}}$$

## Correlation Coefficient's Role

In canonical correlation analysis correlation coefficient play an important role. We elaborate on this soon.

# Linear Combination

- Let

$$\mathbf{c} = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_p \end{pmatrix}, \mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$$

- Linear combination is expressed as:  $\mathbf{c}'\mathbf{X} = c_1X_1 + c_2X_2 + \cdots + c_pX_p$ .
- Mean is given by:  $E(\mathbf{c}'\mathbf{X}) = c_1E(X_1) + c_2E(X_2) + \cdots + c_pE(X_p) = \mathbf{c}'\boldsymbol{\mu}$
- Variance is given by:  $\text{var}(\mathbf{c}'\mathbf{X}) = E(\mathbf{c}'\mathbf{X} - E(\mathbf{c}'\mathbf{X}))^2 = \mathbf{c}'\boldsymbol{\Sigma}\mathbf{c}$

# Linear Combination

$$\begin{aligned}
 Z_1 &= c_{11}X_1 + c_{12}X_2 + \cdots + c_{1p}X_p \\
 Z_2 &= c_{21}X_1 + c_{22}X_2 + \cdots + c_{2p}X_p \\
 &\vdots \\
 Z_q &= c_{q1}X_1 + c_{q2}X_2 + \cdots + c_{qp}X_p
 \end{aligned}$$

- Let

$$\mathbf{Z} = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_q \end{pmatrix}, \mathbf{C} = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1p} \\ c_{21} & c_{22} & \cdots & c_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ c_{q1} & c_{q2} & \cdots & c_{qp} \end{pmatrix}, \mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_q \end{pmatrix}$$

- Linear combination is expressed as:  $\mathbf{Z} = \mathbf{CX}$
- Mean is given by:  $E(\mathbf{Z}) = E(\mathbf{CX}) = \mathbf{C}\mu_{\mathbf{X}}$
- Covariance is given by:  $cov(\mathbf{Z}) = cov(\mathbf{CX}) = \mathbf{C}\Sigma_{\mathbf{X}}\mathbf{C}'$

# Intuition

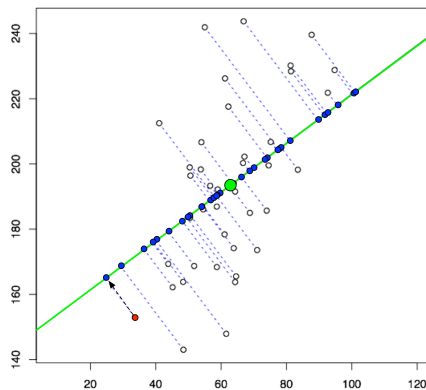


Figure 13: Example dataset. [source: google images](#)



# Intuition

- 1 Intuition is to find a linear combination
- 2 Find a  $\mathbf{c}_1 = [c_{11}, c_{12}, \dots, c_{1p}]'$  having **maximum variance**. That is  $\max \text{var}(\mathbf{c}'_1 \mathbf{X})$
- 3 Next look for a linear combination  $\mathbf{c}'_2 \mathbf{X}$  uncorrelated with  $\mathbf{c}'_1 \mathbf{X}$  and having maximum variance
- 4 Continue step 3 till the chosen  $q$  dimensions are obtained such that  $\mathbf{c}'_q \mathbf{X}$  has maximum variance and uncorrelated with  $\mathbf{c}'_1 \mathbf{X}, \mathbf{c}'_2 \mathbf{X}, \dots, \mathbf{c}'_{q-1} \mathbf{X}$

# Formulation

Finding first direction

$$\begin{array}{ll} \max_{\mathbf{c}_1} & \text{var}(\mathbf{c}_1' \mathbf{X}) \\ \text{s.t.} & \mathbf{c}_1' \mathbf{c}_1 = 1 \end{array}$$

Finding first direction

$$\begin{array}{ll} \max_{\mathbf{c}_1} & \mathbf{c}_1' \Sigma \mathbf{c}_1 \\ \text{s.t.} & \mathbf{c}_1' \mathbf{c}_1 = 1 \end{array}$$

# Formulation

## Formulate Lagrangian

$$L(\mathbf{c}_1, \lambda) = \mathbf{c}_1' \Sigma \mathbf{c}_1 - \lambda(\mathbf{c}_1' \mathbf{c}_1 - 1)$$

## First Order Necessary Condition

$$\frac{\partial L}{\partial \mathbf{c}_1} = \Sigma \mathbf{c}_1 - \lambda \mathbf{c}_1 = 0$$

$$\Sigma \mathbf{c}_1 = \lambda \mathbf{c}_1$$

$$\frac{\partial L}{\partial \lambda} = \mathbf{c}_1' \mathbf{c}_1 - 1 = 0$$

$$\mathbf{c}_1' \mathbf{c}_1 = 1$$

## Maximum variance

Substituting result of first order necessary condition into Lagrangian we have  $\lambda$  retains the maximum variance in the first direction.

$$\begin{aligned} L(\mathbf{c}_1, \lambda) &= \mathbf{c}_1' \lambda \mathbf{c}_1 - \lambda(0) \\ &= \lambda \end{aligned}$$

# Formulation

## Finding second direction

$$\begin{aligned} \max_{\mathbf{c}_2} \quad & \mathbf{c}_2' \Sigma \mathbf{c}_2 \\ \text{s.t.} \quad & \mathbf{c}_2' \mathbf{c}_2 = 1 \\ & \mathbf{c}_2' \mathbf{c}_1 = 0 \end{aligned}$$

## Formulate Lagrangian

$$L(\mathbf{c}_2, \lambda, \phi) = \mathbf{c}_2' \Sigma \mathbf{c}_2 - \lambda(\mathbf{c}_2' \mathbf{c}_2 - 1) - \phi \mathbf{c}_2' \mathbf{c}_1$$

# Optimization

## First Order Necessary Condition

$$\frac{\partial L}{\partial \mathbf{c}_2} = \Sigma \mathbf{c}_2 - \lambda \mathbf{c}_2 - \phi \mathbf{c}_1 = 0$$

multiplying this equation on the left by  $\mathbf{c}_1'$  we have:

$$\mathbf{c}_1' \Sigma \mathbf{c}_2 - \lambda \mathbf{c}_1' \mathbf{c}_2 - \phi \mathbf{c}_1' \mathbf{c}_1 = 0$$

yielding  $\phi = 0$ ; Therefore the first order necessary condition becomes:

$$\Sigma \mathbf{c}_2 - \lambda \mathbf{c}_2 = 0$$

# Solution

## Eigen vectors and Eigen values

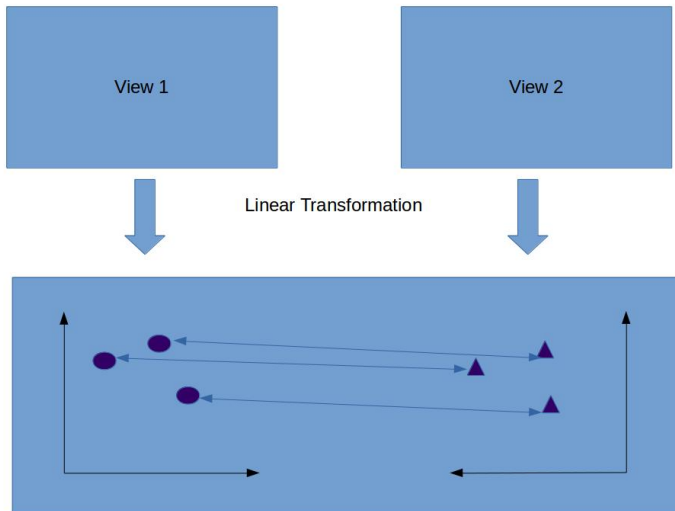
From the above, we note that  $\mathbf{c}_1$  is eigenvector corresponding to the largest Eigen value of  $\Sigma$ .  $\mathbf{c}_2$  correspond to second largest Eigen value.

Matrix  $\mathbf{C}$  (slide 31) correspond to the eigenvectors of  $\Sigma$ .

# Observations

- Single view of dataset is used.
- One subspace is learned through the optimization process.
- The learned subspace is used for further processing.

# Intuition





# Observations

- Find a linearly transformed subspace corresponding to view 1 and linear transformed subspace corresponding to view 2 such that covariance between the subspaces is maximized
- Two views are used
- One optimization is formulated
- Two subspaces - one for each view - are learned simultaneously using one optimization formulation
- The learned subspaces are used for further processing

## Two View Notations

Let  $\mathbf{X}^1$  and  $\mathbf{X}^2$  be two random vectors corresponding to two views with  $q \times 1$  and  $(p - q) \times 1$  dimensions respectively.

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_q \\ X_{q+1} \\ X_{q+2} \\ \vdots \\ X_p \end{bmatrix} = \begin{bmatrix} \mathbf{X}^1 \\ \mathbf{X}^2 \end{bmatrix}; \mu = E(\mathbf{X}) = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_q \\ \mu_{q+1} \\ \mu_{q+2} \\ \vdots \\ \mu_p \end{bmatrix} = \begin{bmatrix} \mu^1 \\ \mu^2 \end{bmatrix}$$

# Two View Notations

## Covariance Matrix

Covariance matrix is given by

$$\begin{aligned}
 \Sigma &= E(\mathbf{X} - \mu)(\mathbf{X} - \mu)' \\
 &= \begin{bmatrix} \underset{(q \times 1)}{(\mathbf{X}^1 - \mu^1)} \underset{(1 \times q)}{(\mathbf{X}^1 - \mu^1)'} & \underset{(q \times 1)}{(\mathbf{X}^1 - \mu^1)} \underset{(1 \times (p-q))}{(\mathbf{X}^2 - \mu^2)'} \\ \underset{((p-q) \times 1)}{(\mathbf{X}^2 - \mu^2)} \underset{(1 \times q)}{(\mathbf{X}^1 - \mu^1)'} & \underset{((p-q) \times 1)}{(\mathbf{X}^2 - \mu^2)} \underset{(1 \times (p-q))}{(\mathbf{X}^2 - \mu^2)'} \end{bmatrix} \\
 &= \left[ \begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array} \right]
 \end{aligned}$$

# Two View Notations

## Covariance Matrix

Covariance matrix is given by

$$\Sigma = E(\mathbf{X} - \mu)(\mathbf{X} - \mu)'$$

$$= \left[ \begin{array}{ccc|ccc} \sigma_{11} & \cdots & \sigma_{1q} & \sigma_{1,q+1} & \cdots & \sigma_{1p} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{q1} & \cdots & \sigma_{qq} & \sigma_{q,q+1} & \cdots & \sigma_{qp} \\ \hline \sigma_{q+q,1} & \cdots & \sigma_{q+1,q} & \sigma_{q+1,q+1} & \cdots & \sigma_{q+1,p} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \cdots & \sigma_{pq} & \sigma_{p,q+1} & \cdots & \sigma_{pp} \end{array} \right]$$

# Two View Notations

## Correlation Coefficient Matrix

$$\rho = \left[ \begin{array}{ccc|ccc} \frac{\sigma_{11}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{11}}} & \cdots & \frac{\sigma_{1q}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{qq}}} & \frac{\sigma_{1,q+1}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{q+1,q+1}}} & \cdots & \frac{\sigma_{1p}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{pp}}} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\sigma_{q1}}{\sqrt{\sigma_{qq}}\sqrt{\sigma_{11}}} & \cdots & \frac{\sigma_{qq}}{\sqrt{\sigma_{qq}}\sqrt{\sigma_{qq}}} & \frac{\sigma_{q,q+1}}{\sqrt{\sigma_{qq}}\sqrt{\sigma_{q+1,q+1}}} & \cdots & \frac{\sigma_{qp}}{\sqrt{\sigma_{qq}}\sqrt{\sigma_{pp}}} \\ \frac{\sigma_{q+1,1}}{\sqrt{\sigma_{q+1,q+1}}\sqrt{\sigma_{11}}} & \cdots & \frac{\sigma_{q+1,q}}{\sqrt{\sigma_{q+1,q+1}}\sqrt{\sigma_{qq}}} & \frac{\sigma_{q+1,q+1}}{\sqrt{\sigma_{q+1,q+1}}\sqrt{\sigma_{q+1,q+1}}} & \cdots & \frac{\sigma_{q+1,p}}{\sqrt{\sigma_{q+1,q+1}}\sqrt{\sigma_{pp}}} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\sigma_{p1}}{\sqrt{\sigma_{pp}}\sqrt{\sigma_{11}}} & \cdots & \frac{\sigma_{pq}}{\sqrt{\sigma_{pp}}\sqrt{\sigma_{qq}}} & \frac{\sigma_{p,q+1}}{\sqrt{\sigma_{pp}}\sqrt{\sigma_{q+1,q+1}}} & \cdots & \frac{\sigma_{pp}}{\sqrt{\sigma_{pp}}\sqrt{\sigma_{pp}}} \end{array} \right]$$

# Measuring Association

- Through maximizing the correlation coefficients
- Let  $U = \mathbf{x}'\mathbf{X}^1$  be the linear transformation for view 1
- Let  $V = \mathbf{y}'\mathbf{X}^2$  be the linear transformation for view 2
- View 1:  $\text{Var}(U) = \mathbf{x}' \text{Cov}(\mathbf{X}^1)\mathbf{x} = \mathbf{x}'\Sigma_{11}\mathbf{x}$
- View 2:  $\text{Var}(V) = \mathbf{y}' \text{Cov}(\mathbf{X}^2)\mathbf{y} = \mathbf{y}'\Sigma_{22}\mathbf{y}$
- $\text{Cov}(U, V) = \mathbf{x}' \text{Cov}(\mathbf{X}^1, \mathbf{X}^2)\mathbf{y} = \mathbf{x}'\Sigma_{12}\mathbf{y}$
- maximize  $\text{Corr}(U, V) = \frac{\mathbf{x}'\Sigma_{12}\mathbf{y}}{\sqrt{\mathbf{x}'\Sigma_{11}\mathbf{x}}\sqrt{\mathbf{y}'\Sigma_{22}\mathbf{y}}}$

# Formulation

## Maximizing Association

$$\begin{aligned} \max_{\mathbf{c}_1} \quad & \mathbf{y}' \Sigma_{21} \mathbf{x} \\ \text{s.t.} \quad & \mathbf{x}' \Sigma_{11} \mathbf{x} = 1 \\ & \mathbf{y}' \Sigma_{22} \mathbf{y} = 1 \end{aligned}$$

# Formulation

## Formulate Lagrangian

$$L(\mathbf{x}, \mathbf{y}, \rho_x, \rho_y) = \mathbf{y}' \Sigma_{21} \mathbf{x} - \frac{\rho_x}{2} \left( \mathbf{x}' \Sigma_{11} \mathbf{x} - 1 \right) - \frac{\rho_y}{2} \left( \mathbf{y}' \Sigma_{22} \mathbf{y} - 1 \right)$$

## First Order Necessary Condition

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{x}} &= \Sigma_{12} \mathbf{y} - \rho_x \Sigma_{11} \mathbf{x} = \mathbf{0} \\ \frac{\partial L}{\partial \mathbf{y}} &= \Sigma_{21} \mathbf{x} - \rho_y \Sigma_{22} \mathbf{y} = \mathbf{0} \end{aligned} \tag{5}$$



# Derivation

## Details

Subtract  $\mathbf{x}'$  times first equation from  $\mathbf{y}'$  times second and we have:

$$\mathbf{y}'\Sigma_{21}\mathbf{x} - \rho_y\mathbf{y}'\Sigma_{22}\mathbf{y} - \mathbf{x}'\Sigma_{12}\mathbf{y} + \rho_x\mathbf{x}'\Sigma_{11}\mathbf{x} = 0$$

$$\rho_x\mathbf{x}'\Sigma_{11}\mathbf{x} - \rho_y\mathbf{y}'\Sigma_{22}\mathbf{y} = 0$$

$$\rho_x = \rho_y = \rho$$

## Details

First sub equation of eq. (5) yields

$$\mathbf{x} = \frac{\Sigma_{11}^{-1}\Sigma_{12}\mathbf{y}}{\rho} \quad (6)$$

# Derivation

## Details

Substituting eq. (6) in second sub eq. (5) we have:

$$\Sigma_{21} \frac{\Sigma_{11}^{-1} \Sigma_{12} \mathbf{y}}{\rho} - \rho \Sigma_{22} \mathbf{y} = \mathbf{0}$$

$$(\Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} - \rho^2 \Sigma_{22}) \mathbf{y} = \mathbf{0}$$

In analogous way we can write:

$$(\Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} - \rho^2 \Sigma_{11}) \mathbf{x} = \mathbf{0}$$

# Derivation

## Generalized Eigenvalue Problem

Given a symmetric matrix  $A \in \mathbb{R}^{n \times n}$  and a symmetric positive definite matrix  $B \in \mathbb{R}^{n \times n}$  we consider the problem of finding a nonzero vector  $\mathbf{x}$  and a scalar  $\lambda$  so that  $A\mathbf{x} = \lambda B\mathbf{x}$  is the symmetric-definite generalized eigenvalue problem.

The scalar  $\lambda$  can be thought of as generalized eigenvalue. Determining  $\lambda(A, B) = \{\lambda | \det(A - \lambda B) = 0\}$

## Details

$$\begin{cases} (\Sigma'_{12} \Sigma_{11}^{-1} \Sigma_{12} - \rho^2 \Sigma_{22}) \mathbf{y} = \mathbf{0} \\ (\Sigma_{12} \Sigma_{22}^{-1} \Sigma'_{12} - \rho^2 \Sigma_{11}) \mathbf{x} = \mathbf{0} \end{cases} \quad (7)$$

# Intuition

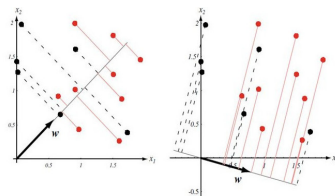


Figure 15: Example dataset. source: [duda hart textbook](#)

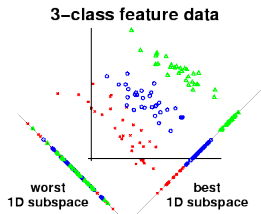


Figure 16: Example dataset. source: [google images](#)

# Intuition

- PCA seeks directions that are **efficient for representation**
- PCA does not need the class label information
- Discriminant analysis seeks directions that are **efficient for discrimination**
- Uses class label information

# Intuition

## Objective function

Find a direction  $\mathbf{w}$  that maximize the inter class distance and minimize intra class variance.

## Objective function

Maximize inter class distance is achieved by maximizing the difference between the projected means

Minimizing intra class variance is achieved by minimizing the sum of the class variances.

$$\max_{\mathbf{w}} \frac{|\hat{\mu}_1 - \hat{\mu}_2|^2}{\hat{\sigma}_1^2 + \hat{\sigma}_2^2} \quad (8)$$

# Intuition

## Within and Between Class Scatter

$$\hat{\mu}_1 = \frac{1}{n_1} \sum_{\mathbf{x} \in \mathcal{D}_1} \mathbf{w}' \mathbf{x} = \mathbf{w}' \mu_1$$

$$\begin{aligned}\hat{\sigma}_1^2 &= \sum_{\mathbf{x} \in \mathcal{D}_1} (\mathbf{w}' \mathbf{x} - \mathbf{w}' \mu_1)^2 \\ &= \sum_{\mathbf{x} \in \mathcal{D}_1} (\mathbf{w}' (\mathbf{x} - \mu_1) (\mathbf{x} - \mu_1)' \mathbf{w}) \\ &= \mathbf{w}' \mathbf{S}_1 \mathbf{w}\end{aligned}$$

# Intuition

## Within and Between Class Scatter

$$\hat{\mu}_2 = \frac{1}{n_2} \sum_{\mathbf{x} \in \mathcal{D}_2} \mathbf{w}' \mathbf{x} = \mathbf{w}' \mu_2$$

$$\begin{aligned}\hat{\sigma}_2^2 &= \sum_{\mathbf{x} \in \mathcal{D}_2} (\mathbf{w}' \mathbf{x} - \mathbf{w}' \mu_2)^2 \\ &= \sum_{\mathbf{x} \in \mathcal{D}_2} (\mathbf{w}' (\mathbf{x} - \mu_2) (\mathbf{x} - \mu_2)' \mathbf{w}) \\ &= \mathbf{w}' \mathbf{S}_2 \mathbf{w}\end{aligned}$$



# The numerator

## Between Class Scatter

$$\begin{aligned} |\hat{\mu}_1 - \hat{\mu}_2|^2 &= (\mathbf{w}'\mu_1 - \mathbf{w}'\mu_2)^2 \\ &= \mathbf{w}'(\mu_1 - \mu_2)(\mu_1 - \mu_2)'\mathbf{w} \\ &= \mathbf{w}'\mathbf{S}_B\mathbf{w} \end{aligned}$$

# The denominator

## Within Class Scatter

$$\begin{aligned}\hat{\sigma}_1^2 + \hat{\sigma}_2^2 &= \mathbf{w}' \mathbf{S}_1 \mathbf{w} + \mathbf{w}' \mathbf{S}_2 \mathbf{w} \\ &= \mathbf{w}' (\mathbf{S}_1 + \mathbf{S}_2) \mathbf{w} \\ &= \mathbf{w}' \mathbf{S}_W \mathbf{w}\end{aligned}$$

# The Objective

## Fisher's Linear Discriminant Analysis

$$\max_{\mathbf{w}} \frac{\mathbf{w}' \mathbf{S}_B \mathbf{w}}{\mathbf{w}' \mathbf{S}_W \mathbf{w}} \quad (9)$$

Equation (9) is a generalized Rayleigh quotient.  $\mathbf{w}$  that maximizes eq. (9) must satisfy

$\mathbf{S}_B \mathbf{w} = \lambda \mathbf{S}_W \mathbf{w}$  - a generalized eigenvalue problem.

# Intuition

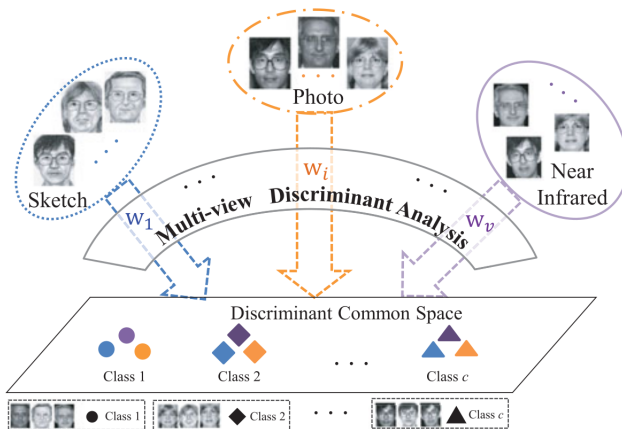


Figure 17: Example dataset. source: doi: 10.1109/TPAMI.2015.2435740

# Intuition

- Let  $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_v$  be  $v$  linear transforms to be found.
- Let  $\mathcal{X}^j = \{\mathbf{x}_{ijk} | i = 1, 2, \dots, c; k = 1, \dots, n_{ij}\}$ .  $j$  is the view index;  $i$  is the class label index;  $k$  is the data point index;
- Projection data is denoted as:  
$$\mathcal{Y} = \{\mathbf{w}'^j \mathbf{x}_{ijk} | i = 1, \dots, c; j = 1, \dots, v; k = 1, \dots, n_{ij}\}$$
- Objective 1: Between class variation  $\mathbf{S}_B^y$  from all views should be maximized.
- Objective 2: Within class variation  $\mathbf{S}_W^y$  from all views should be minimized.

# Intuition

## Objective function

$$\arg \max_{\mathbf{w}_1, \dots, \mathbf{w}_v} \frac{\text{tr}(\mathbf{S}_B^y)}{\text{tr}(\mathbf{S}_W^y)} \quad (10)$$

## Within Class Scatter

- $\mathbf{S}_W^y = \sum_{i=1}^c \sum_{j=1}^v \sum_{k=1}^{n_{ij}} (\mathbf{y}_{ijk} - \boldsymbol{\mu}_i)(\mathbf{y}_{ijk} - \boldsymbol{\mu}_i)'$
- where  $\boldsymbol{\mu}_i = \frac{1}{n_i} \sum_{j=1}^v \sum_{k=1}^{n_{ij}} \mathbf{y}_{ijk}$  is the mean of all sample of  $i^{th}$  class over all views in the common subspace.
- $n_i = \sum_{j=1}^v n_{ij}$
- $\boldsymbol{\mu} = \frac{1}{n} \sum_{i=1}^c \sum_{j=1}^v \sum_{k=1}^{n_{ij}} \mathbf{y}_{ijk}$  is the mean of all samples over all views.
- $n = \sum_{i=1}^c n_i$

# Intuition

## Between Class Scatter

$$\mathbf{S}_B^y = \sum_{i=1}^c n_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})'$$

# Matrix Form

## Within Class Scatter

$$\begin{aligned} \mathbf{S}_W^y &= [\mathbf{w}'_1 \cdots \mathbf{w}'_v] \begin{pmatrix} \mathbf{S}_{11} & \cdots & \mathbf{S}_{1v} \\ \vdots & \ddots & \vdots \\ \mathbf{S}_{v1} & \cdots & \mathbf{S}_{vv} \end{pmatrix} \begin{bmatrix} \mathbf{w}_1 \\ \vdots \\ \mathbf{w}_v \end{bmatrix} \\ &= \mathbf{W}' \mathbf{S} \mathbf{W} \end{aligned}$$

details: doi:10.1109/TPAMI.2015.2435740

## Between Class Scatter

$$\begin{aligned} \mathbf{S}_B^y &= [\mathbf{w}'_1 \cdots \mathbf{w}'_v] \begin{pmatrix} \mathbf{D}_{11} & \cdots & \mathbf{D}_{1v} \\ \vdots & \ddots & \vdots \\ \mathbf{D}_{v1} & \cdots & \mathbf{D}_{vv} \end{pmatrix} \begin{bmatrix} \mathbf{w}_1 \\ \vdots \\ \mathbf{w}_v \end{bmatrix} \\ &= \mathbf{W}' \mathbf{D} \mathbf{W} \end{aligned}$$

details: doi:10.1109/TPAMI.2015.2435740



# Formulation

Objective function

$$\arg \max_{\mathbf{w}_1, \dots, \mathbf{w}_v} \frac{\mathbf{W}' D \mathbf{W}}{\mathbf{W}' S \mathbf{W}} \quad (11)$$

# Introduction

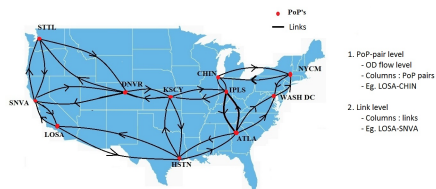


Figure 18: Abilene Network. source: google images

Time interval	P1-P1	P1-P2	...	Pn-Pn
30-June-2018 09:00-09:05	3475	4928	...	4983
30-June-2018 09:05-09:10	3474	4228	...	2746
.	.	.	.	.
.	.	.	.	.
06-Jul-2018 11:55-00:00	3347	4274	...	3748

# PCA subspace based method

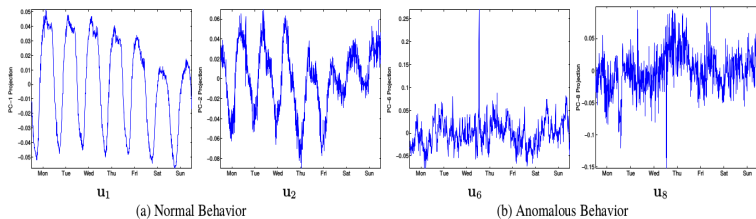


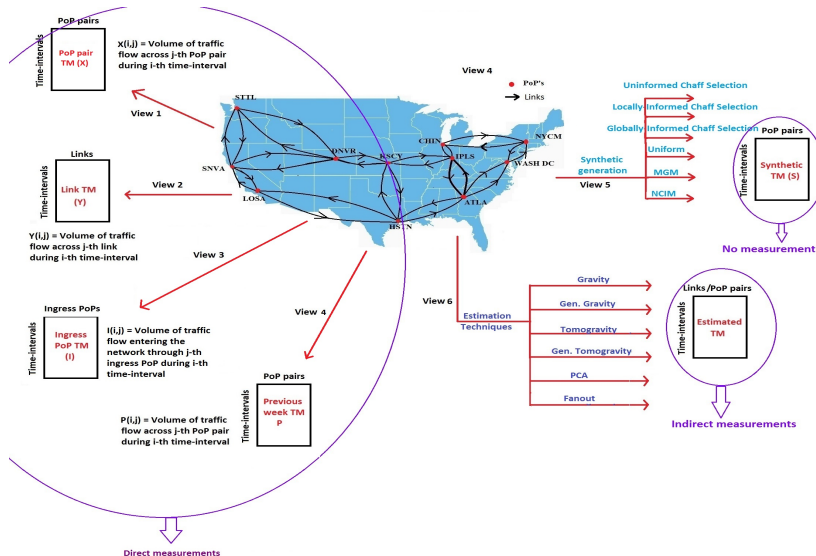
Figure 19: Anomaly detection. source: doi: 10.1145/1015467.1015492

# Limitations

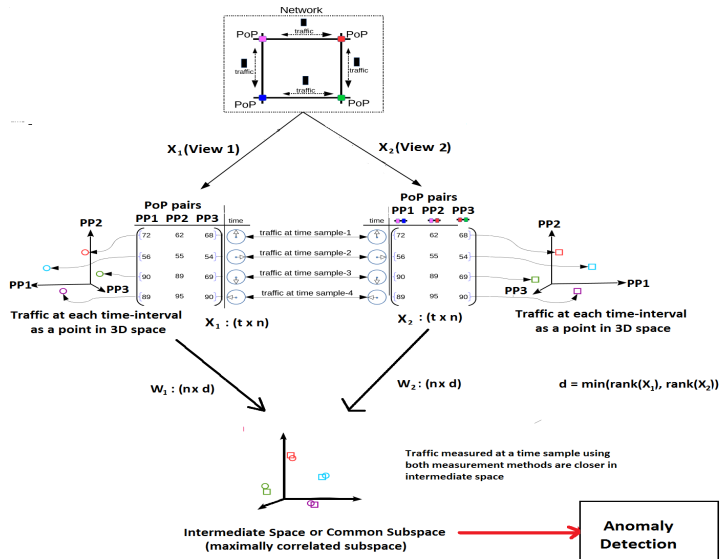
PCA subspace method highly popular for network wide anomaly detection. However, this method is shown to be sensitive to:

- **Small differences** in the number of principal components in the normal subspace.
- **Level of aggregation** of the traffic measurements.
- **Detection threshold**.
- Propose to alleviate sensitivities by the use of **multi-view subspace learning method**.
- Multiple views for traffic matrix is a **non-trivial task and is non-intuitive**.

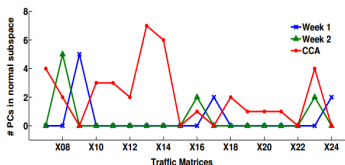
# Multiple View Of Traffic Matrix



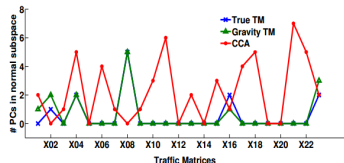
# Multiple View Subspace learning



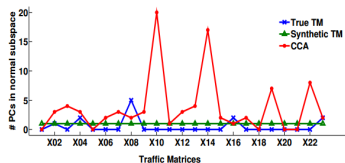
# Non-contamination of Normal Subspace



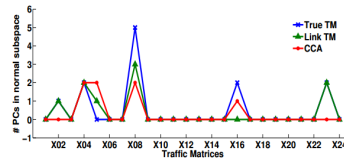
(a) Previous and Current Week TM



(b) Direct and Estimated TM



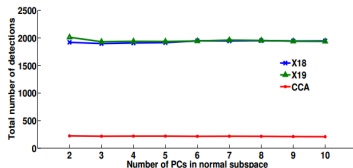
(c) Direct and Synthetic TM



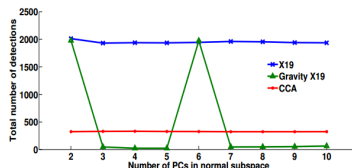
(d) TM of two different aggregation level

Figure 22: Number of PCs in the Normal Subspace TMs

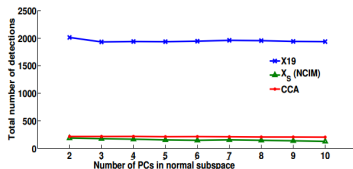
# Insensitive to Number of PCs in normal subspace



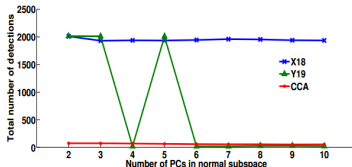
(a) Previous and Current Week TM



(b) Direct and Estimated TM



(c) Direct and Synthetic TM

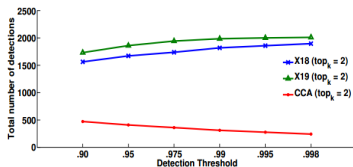


(d) TM of two different aggregation level

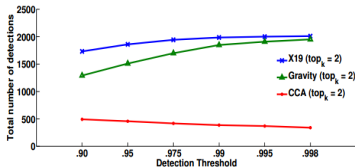
Figure 23: Impact of number of PCs in the Normal Subspace for X19



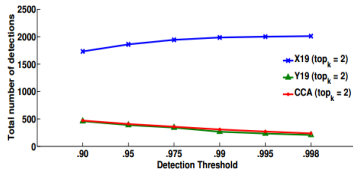
# Insensitive to Detection threshold



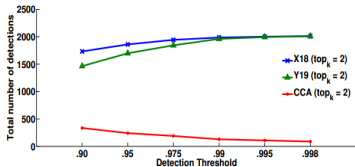
(a) Previous and Current Week TM



(b) Direct and Estimated TM



(c) Direct and Synthetic TM



(d) TM of two different aggregation level

# Summary

- (Unsupervised) Dimensionality reduction
  - **single view:** Principal component analysis
  - **multi view:** Canonical correlation analysis
- (Supervised) Dimensionality reduction
  - **single view:** linear discriminant analysis
  - **multi view:** multi-view linear discriminant analysis
- Presented an application.

# References

- <https://nsdl.co.in/about/why.php/>
- [http://deic.uab.es/~iblanes/beamer\\_gallery/](http://deic.uab.es/~iblanes/beamer_gallery/)
- <http://www.idbi.com/online-tax-payment.asp>

## E-Books

- Programming ASP.NET 3.5
- The Indian Financial System: Markets, Institutions and Services

- Thank You.