Dimensionality Reduction

Adepu Ravi Sankar
cs14resch11001@iith.ac.in

Course Instructor: Dr Vineeth N Balasubramanian
Indian Institute of Technology, Hyderabad

29th Oct 2016

Administravia

Kindly please mute your microphone

Outline of Today's lecture

- Overview
- Linear Methods
  - PCA
  - CCA
  - ICA
  - Fisher Discriminant Analysis
- Non-Linear Methods
  - Multi dimensionality Scaling
  - Manifold Learning

Why Dimensionality Reduction?

- Practical Reasons
- Theoretical Reasons

Practical Reasons for Dimensionality Reduction

- Redundancy reduction
- Intrinsic structure discovery
- Removal of irrelevant and noisy features
- Feature extraction (To avoid curse of dimensionality)
- Visualization purpose
- Computation and Machine learning perspective

Theoretical Reasons for Dimensionality Reduction

Strange and annoying phenomena: **"Curse of dimensionality"**

Notations may change for few models that we discuss

Principal Component Analysis

- PCA is the most basic yet important feature transform in feature extraction and dimensionality reduction researches

PCA Basic Idea

- How do we characterize the distribution in simple terms?
- To characterize the remaining distribution?
- How about multi dimensional case?
- Data cloud example

PCA Basic Idea

- How do we characterize the distribution in simple terms?
- To characterize the remaining distribution?
- How about multi dimensional case?
- Data cloud example
- Principal component analysis is the answer
- We deal with zero mean data
- Principal components are linear combinations $s = \sum_i w_i x_i$ that contain as much of the variance of the input data as possible
- The first principal component is intuitively defined as the linear combination of observed variables, which has the maximum variance.
- A constraint on the weights w_i , which we call the principal component weights, must be imposed as well. What if not?

PCA Cont...

- A natural constraint on weights?
- This definition gives only one principal component
- Typically, the way to obtain many principal components is by a "deflation" approach
- After estimating the first PC, we want to find the linear combination of maximum variance under the constraint that the new combination must be orthogonal to the first one
- This will then be called the second principal component
- This procedure can be repeated to obtain as many components as there are dimensions in the data space.

PCA Cont..

Formally

Assume that we have estimated k principal components, given by the weight vectors $\boldsymbol{w_1}, \boldsymbol{w_2}, \cdots, \boldsymbol{w_k}$. Then the $k+1$ th principal component weight vector is defined by

$$\max_{w} var\left(\sum_{i} w_i x_i\right)$$

with constraints

$$||\boldsymbol{w}|| = \sqrt{\sum_{i} w_i^2} = 1$$

$$\sum_{i} w_{ji} w_i = 0; \forall j = 1, .., k$$

Dimension reduction by PCA

- Consider the following general problem
- We have a very large number, say m, of random variables $x_1, ..., x_m$
- Computations that use all the variables would be too burdensome.
- We want to reduce the dimension of the data to n
- $z_i = \sum_{j=1}^{m} w_{ij} x_j$ for all $i = 1, .., n$
- What do we want preserve in projected data?

Dimension reduction by PCA

- Consider the following general problem
- We have a very large number, say m, of random variables $x_1, ..., x_m$
- Computations that use all the variables would be too burdensome.
- We want to reduce the dimension of the data to n
- $z_i = \sum_{j=1}^{m} w_{ij} x_j$ for all $i = 1, .., n$
- What do we want preserve in projected data?
- How to model this?

Dimension reduction by PCA

- Consider the following general problem
- We have a very large number, say m, of random variables $x_1, ..., x_m$
- Computations that use all the variables would be too burdensome.
- We want to reduce the dimension of the data to n
- $z_i = \sum_{j=1}^{m} w_{ij} x_j$ for all $i = 1, .., n$
- What do we want preserve in projected data?
- How to model this?
- Reconstruction error
- $$E\left\{ \sum_j \left(x_j - \sum_i a_{ji} z_i \right)^2 \right\} = E\left\{ ||\mathbf{x} - \sum_i \boldsymbol{a_i} z_i||^2 \right\}$$
- The solution is to take as the z_i the n first principal components (Proof follows)

Proof of how PCA is related to eigenvalues of the covariance matrix

Variance of a Random variable

$$var(x_1) = E\{x_1^2\} - (E\{x_1\})^2$$

also can be written as:

$$var(x_1) = E\{(x_1^2 - E\{x_1\})^2$$

Measures average deviation from the mean value

Covariance

$$cov(x_1, x_2) = E\{x_1 x_2\} - E\{x_1\}E\{x_2\}$$

When do you say the variables as *uncorrelated*?

Covariance matrix

$$\mathrm{C}(\mathrm{x}) = \begin{bmatrix} cov(x_1, x_1) & cov(x_1, x_2) & \dots & cov(x_1, x_n) \\ cov(x_2, x_1) & cov(x_2, x_2) & \dots & cov(x_2, x_n) \\ \vdots & & & \\ cov(x_n, x_1) & cov(x_n, x_1) & \dots & cov(x_n, x_n) \end{bmatrix}$$

- Covariance matrix is basically a generalization of variance to random vectors
- Matrix notation is $\mathbf{C}(x) = E\{\mathbf{x}\mathbf{x}^T\} - E\{x\}E\{x^T\}$.
- Symmetric matrix
- If the variables are uncorrelated, the covariance matrix is diagonal
- If normalized to unit variance, the covariance matrix equals the identity matrix

Eigenvalues of covariance matrix

The important point to note is that the variance of any linear combination (Principal component here) can be computed using the *covariance matrix* of the data

Consider any linear combination $\boldsymbol{w}^T\boldsymbol{x} = \sum_i w_i x_i$; we can compute its variance by:

$$E\{(\boldsymbol{w}^T\boldsymbol{x})^2\} = E\{(\boldsymbol{w}^T\boldsymbol{x})(\boldsymbol{w}^T\boldsymbol{x})\} = E\{\boldsymbol{w}^T(\boldsymbol{x}\boldsymbol{x}^T)\boldsymbol{w}\} = \boldsymbol{w}^T E\{\boldsymbol{x}\boldsymbol{x}^T\}\boldsymbol{w}$$

$$= \boldsymbol{w}^T\boldsymbol{C}\boldsymbol{w}$$

Where $\boldsymbol{C} = E\{\boldsymbol{x}\boldsymbol{x}^T\}$ is the covariance matrix

PCA formulation

The basic PCA can be formulated as:

$$\max_{w:||w||=1} \boldsymbol{w^T C w}$$

- Since **C** is a symmetric matrix $\boldsymbol{C = UDU^T}$
  - **U** is an orthogonal matrix
  - **D** $= \text{diag}(\lambda_1, .., \lambda_m)$ is diagonal
  - Columns of **U** are called *eigenvectors*
- Let $\boldsymbol{v = U^T w}$ PCA can be easily solved as :

$$\boldsymbol{w^T C w = w^T UDU^T w = v^T D v} = \sum_i v_i^2 \lambda_i$$

- Since **U** is orthogonal $||v|| = ||w||$; How?

PCA formulation

The basic PCA can be formulated as:

$$\max_{w:||w||=1} \boldsymbol{w^T C w}$$

- Since **C** is a symmetric matrix $\boldsymbol{C = UDU^T}$
  - **U** is an orthogonal matrix
  - **D** $= \text{diag}(\lambda_1, .., \lambda_m)$ is diagonal
  - Columns of **U** are called *eigenvectors*
- Let $\boldsymbol{v = U^T w}$ PCA can be easily solved as :

$$\boldsymbol{w^T C w = w^T UDU^T w = v^T D v} = \sum_i v_i^2 \lambda_i$$

- Since **U** is orthogonal $||v|| = ||w||$; How?
- Do $v^T v$
- So the constraint on v is same as for w

PCA Formulation cont...

- Let us make further change of variables to $m_i = v_i^2$
- The constraint of unit norm of v is now equivalent to the constraints that the sum of the m_i must equal one
- The problem is now transformed to

$$\max_{m_i \geq 0, \sum m_i = 1} \sum_i m_i \lambda_i$$

- It is obvious that max is found when the m_i corresponding to the largest λ_i is one and others are zero

- Let us denote by i^* the index of the maximum eigenvalue
- This corresponds to w begin equal to the i^*-th eigenvector, that is, the i^* -th column of U
- This is how first principal component is easily computed by the eigenvalue decomposition
- How is the second principal component calculated?
- The i-th principal component s_i is equal to $s_i = \boldsymbol{u_i^T x}$

Mathematical implication of PCA: Projected data is un*correlated*

We know $s = U^T x$; Now, $E\{ss^T\} = ?$

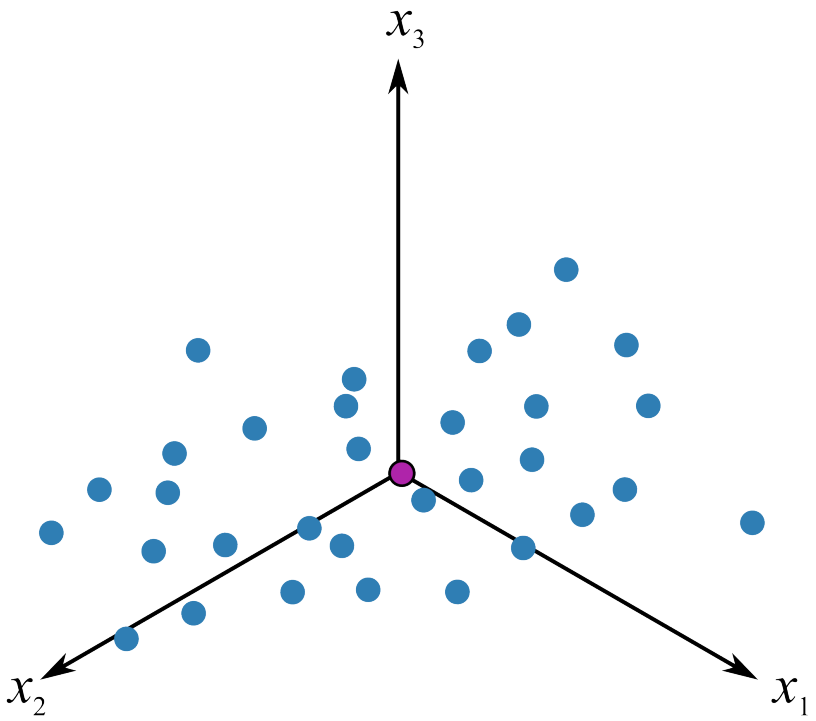
Proportion of variance explained by the components

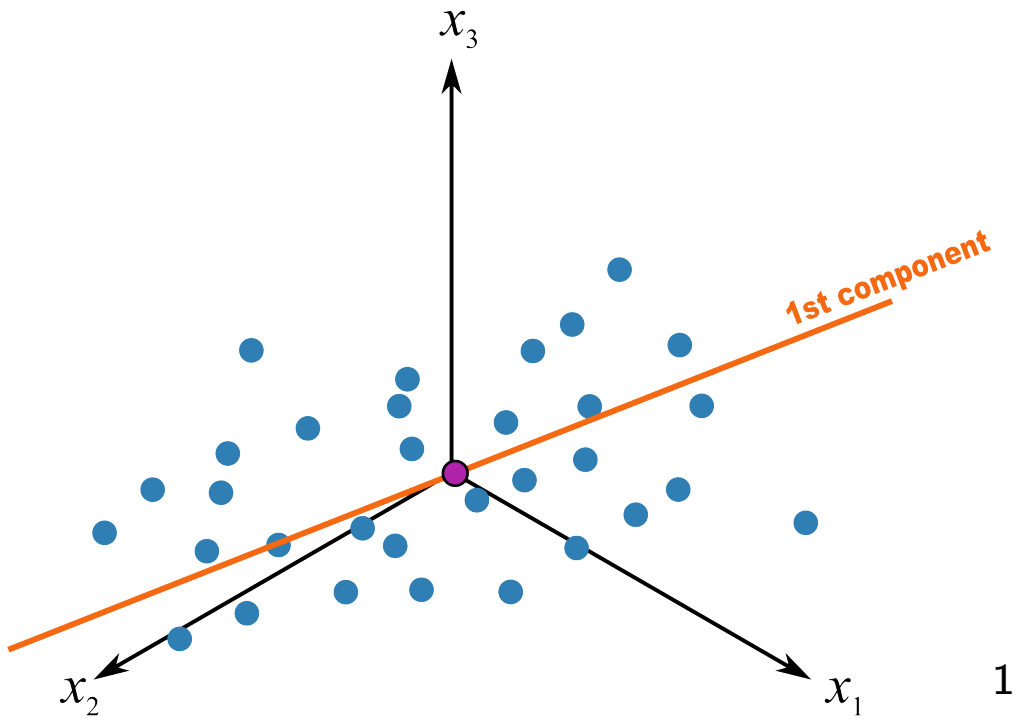
- We want to ask the question "How much variance do we need to capture?"
- We saw that the eigenvalues of the covariance matrix give the variance of each component
- This is typically interpreted as the amount of the total variance of the data which is "explained" by the component.
- If we take k first principal components, they together "explain" the variance of amount
- Proportion of variance explained is

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n} \lambda_i}$$

In summary PCA

- PCA projects the data along the directions where the data varies the most
- These directions are determined by the eigenvectors of the covariance matrix corresponding to the largest eigenvalues.





[1]

[1] http://learnche.org/pid/latent-variable-modelling/principal-component-analysis/geometric-explanation-of-pca

Independent Component Analysis

- PCA - Finding correlation that maximize variance through Reconstruction
- ICA - Maximize the independence
  - Tries to find a **linear transformation** such that each of individual features are mutually independent (in a statistical sense)
  - $Y_i \perp\!\!\!\perp Y_j$
  - $I(Y_i, Y_j) = 0$
  - $\max I(X, Y)$

ICA Motivating example[2]

- Blind source separation problem (cocktail party)
- Ex: Try to record a person's voice on a city street
- We assert that a measurement can be a combination of many distinct sources
- ICA has two related interpretations – filtering and dimensional reduction
- How Filtering?

[2] A Tutorial on Independent Component Analysis-Jonathon Shlens

ICA Motivating example[2]

- Blind source separation problem (cocktail party)
- Ex: Try to record a person's voice on a city street
- We assert that a measurement can be a combination of many distinct sources
- ICA has two related interpretations – filtering and dimensional reduction
- How Filtering?
- How Dimensional reduction?

[2] A Tutorial on Independent Component Analysis-Jonathon Shlens

ICA: Cocktail party problem

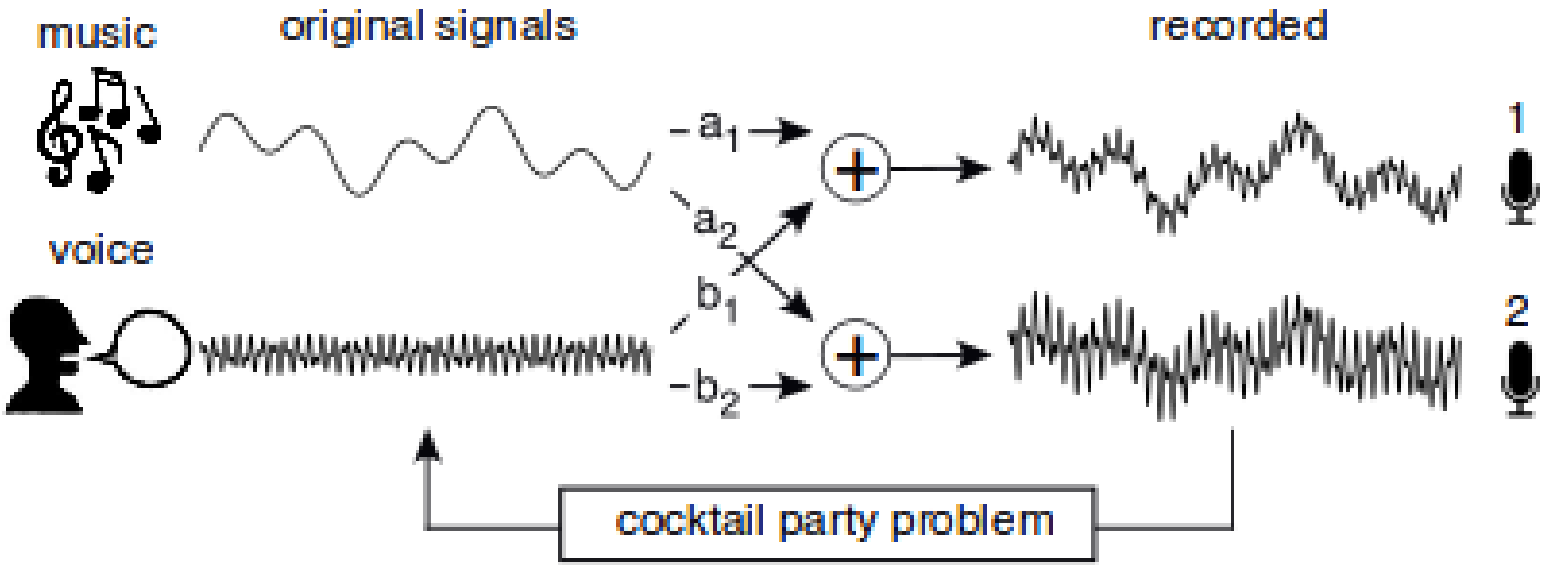


FIG. 1 Example of the cocktail party problem. Two sounds s_1, s_2 are generated by music and a voice and recorded simultaneously in two microphones. Sound adds linearly. Two microphones record a unique linear summation of the two sounds. The linear weights for each microphone (a_1, b_1 and a_2, b_2) reflect the proximity of each speaker to the respective microphones. The goal of the cocktail party problem is to recover the original sources (i.e. music and voice) solely using the microphone recordings (Bregman, 1994).

ICA: Image deblurring problem

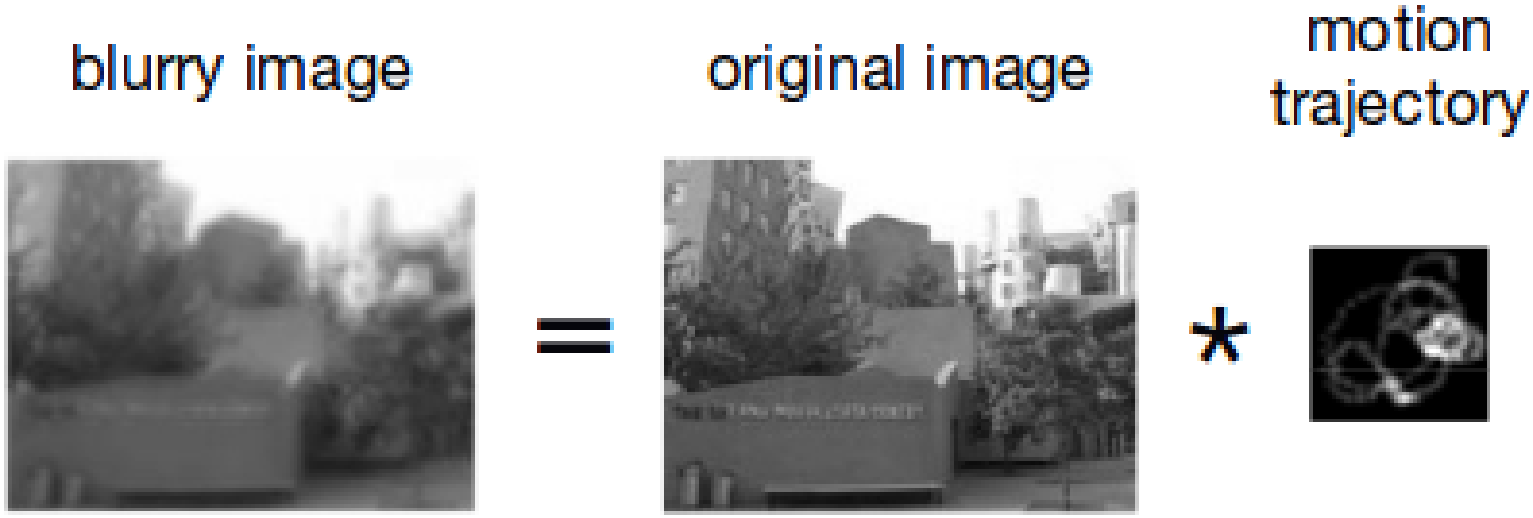


FIG. 2 Example of removing blur from an image due to camera motion. A blurry image (left panel) recorded on a camera sensory array is approximately equal to the convolution of the original image (middle panel) and the motion path of the camera (right panel). Each pixel in the blurry image is the weighted sum of pixels in the original image along the camera motion trajectory. De-blurring an image requires identifying the original image and the motion path from a single blurry image (reproduced from Fergus *et al.* (2006)).

ICA Setup

- We record some multi-dimensional data **x**
- Each sample is drawn from unknown distribution $P(\boldsymbol{x})$
- Assume there is some underlying sources **s** where each s_i is statistically independent
- Key assumption for ICA is that observed **x** is a *linear* mixture of underlying sources

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s}$$

- **A** is some unknown invertible, square matrix
- Goal of ICA: Find **A**, to recover original signal **s**
- Is this possible?

ICA Setup

- We record some multi-dimensional data **x**
- Each sample is drawn from unknown distribution $P(\boldsymbol{x})$
- Assume there is some underlying sources **s** where each s_i is statistically independent
- Key assumption for ICA is that observed **x** is a *linear* mixture of underlying sources

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s}$$

- **A** is some unknown invertible, square matrix
- Goal of ICA: Find **A**, to recover original signal **s**
- Is this possible?
- We will construct a new matrix **W** (*un-mixing matrix* which is an approximation of $\boldsymbol{A}^{-1}$)such that

$$\hat{\boldsymbol{s}} = \boldsymbol{W}\boldsymbol{x}$$

A STRATEGY FOR SOLVING ICA

- Divide-and-conquer provides a strategy to solve this problem
- Rather than trying to solve for **s** and **A** simultaneously, we focus on finding **A**
- Furthermore, rather than trying to solve for **A** all at once, we solve for **A** in a piece-meal fashion by cutting up **A** into simpler and more manageable parts
- Any ideas on how to split **A**?

A STRATEGY FOR SOLVING ICA

- Divide-and-conquer provides a strategy to solve this problem
- Rather than trying to solve for **s** and **A** simultaneously, we focus on finding **A**
- Furthermore, rather than trying to solve for **A** all at once, we solve for **A** in a piece-meal fashion by cutting up **A** into simpler and more manageable parts
- Any ideas on how to split **A**?
- SVD: $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$
- Decomposed into 3 simpler linear operations, a rotation **V**, a stretch along the axes Σ, and a second rotation **U**
- Each matrix has fewer entries, and each of them are easy to invert
- We estimate **A** and its inverse **W** by recovering each piece of the decomposition individually:

$$\boldsymbol{W} = \boldsymbol{A}^{-1} = \boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^T$$

How to calculate unmixing matrix W

- Examing the covariance of data $\mathbf{x}$ to calculate $\mathbf{U}$ and Σ
- Use the assumption of statistical independence of $\mathbf{s}$ to solve for $\mathbf{V}$

Examining the Covariance of the data

- Goal: Use covariance of data to recover two matrices of **W** i.e Σ and **U**
- Lets make one more assumption that covariance of sources is *whitented* i.e $\langle \boldsymbol{s}\boldsymbol{s}^T \rangle = \mathbf{I}$
- Calculate the covariance of the observed data:

$$\langle \boldsymbol{x}\boldsymbol{x}^T \rangle = \langle (\boldsymbol{A}\boldsymbol{s})(\boldsymbol{A}\boldsymbol{s})^T \rangle$$

$$= \langle (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\boldsymbol{T}}\boldsymbol{s})(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\boldsymbol{T}})^T \rangle$$

$$= \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\boldsymbol{T}} \langle \boldsymbol{s}\boldsymbol{s}^T \rangle \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^{\boldsymbol{T}}$$

$$= \boldsymbol{U}\boldsymbol{\Sigma}^2\boldsymbol{U}^{\boldsymbol{T}}$$

- The covariance of data is independent of sources **s** as well as **V** (Although with a shrewd choice of assumption)
- What is special about above Decomposition?

Examining the Covariance of the data cont...

- The above decomposition expresses the covariance of data in terms of diagonal matrix sandwiched between two orthogonal matrices

Examining the Covariance of the data cont...

- The above decomposition expresses the covariance of data in terms of diagonal matrix sandwiched between two orthogonal matrices
- We know any symmetric matrix (here our covariance matrix) can be orthogonally diagonalized by their eigen vectors i.e $\langle \boldsymbol{x}\boldsymbol{x}^T \rangle = \boldsymbol{E}\boldsymbol{D}\boldsymbol{E}^T$
- **E** is the orthogonal matrices whose columns are eigenvectors of covariance of **x**,
- **D** is the diagonal matrix of associated eigen values
- We have got $\langle \boldsymbol{x}\boldsymbol{x}^T \rangle = \boldsymbol{U}\boldsymbol{\Sigma}^2\boldsymbol{U}^T$ from ICA
- We have $\boldsymbol{W} = \boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^T$
- By mapping we have:

$$\boldsymbol{W} = \boldsymbol{V}\boldsymbol{D}^{\frac{-1}{2}}\boldsymbol{E}^T$$

- We have found latter two matrices as **D,E** who are eigenvalues and eigenvectors of covariance of data **x**
- Let's try to interpret the results before calculating **V**

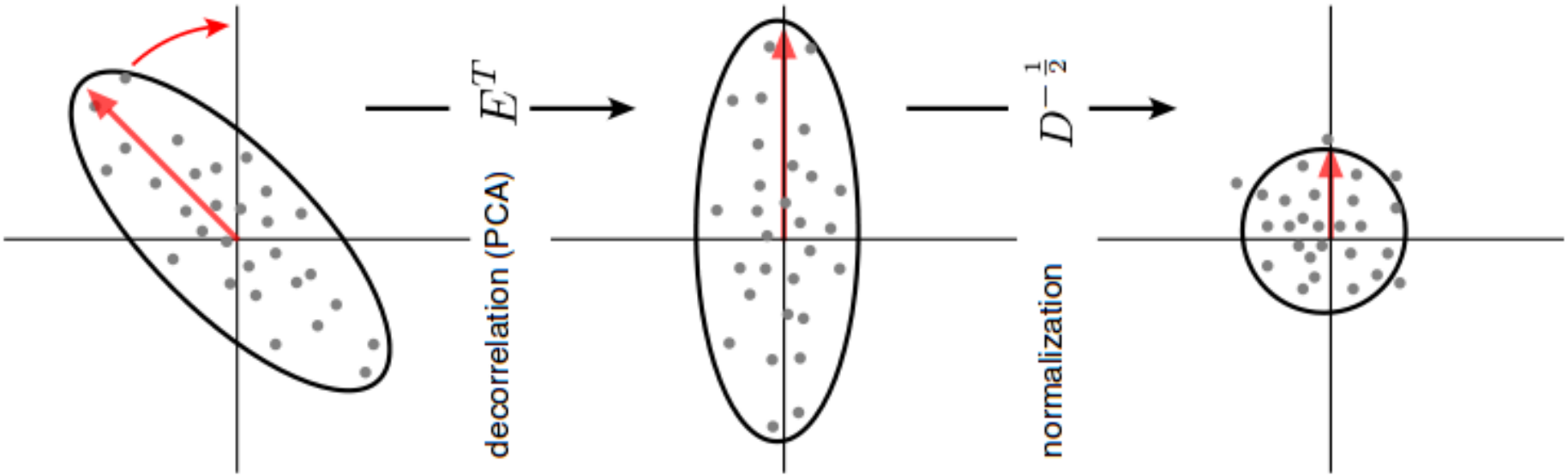
Revisiting Whitening

- Till now what we did in ICA is called *whitening*
- Whitening? Ever heard? What it does?

Revisiting Whitening

- Till now what we did in ICA is called *whitening*
- Whitening? Ever heard? What it does?
- Removes all linear dependencies in a data set (i.e. second-order correlations) and normalizes the variance along all dimensions
- Which is nothing but applying $\boldsymbol{D}^{\frac{-1}{2}}\boldsymbol{E}^{T}$
  - The first step performs the rotation in order to decorrelate the data, i.e. remove linear dependencies by multiplying with $\boldsymbol{E}^{T}$
  - The second operation normalizes the variance in each dimension by multiplying with $\boldsymbol{D}^{\frac{-1}{2}}$

ICA: Interpreting the results



Whitening a data set can be represented as a series of two linear operations. Data is projected on the principal components $\boldsymbol{E}^T\boldsymbol{x}$ Each axis is then scaled so that every direction has unit variance $\boldsymbol{D}^{\frac{-1}{2}}\boldsymbol{E}^T\boldsymbol{x}$

- This whitening simplifies the ICA problem to finding a single rotation matrix $\mathbf{V}$
- Let us define

$$\boldsymbol{x}_w = (\boldsymbol{D}^{\frac{-1}{2}}\boldsymbol{E}^T)\boldsymbol{x}$$

- Substituting the above in $\hat{\boldsymbol{s}} = \boldsymbol{W}\boldsymbol{x}$ and $\boldsymbol{W} = \boldsymbol{A}^{-1} = \boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^T$ simplifies ICA down to solving

$$\hat{\boldsymbol{s}} = \boldsymbol{V}\boldsymbol{x}_w$$

- How do we find $\mathbf{V}$?
- Small detour to *information theory*

Finding $\mathbf{V}$ using Information Theory

- Goal: Find the last rotation $\mathbf{V}$ such that estimate of $\hat{\boldsymbol{s}}$ is statistically independent
- Why mutual information?

Finding **V** using Information Theory

- Goal: Find the last rotation **V** such that estimate of $\hat{\boldsymbol{s}}$ is statistically independent
- Why mutual information?
- The mutual information measures the departure of two variables from statistical independence
- We use *multi-information*, a generalization of mutual information, which measures the statistical dependence between multiple variables:

$$I(y) = \int P(\boldsymbol{y}) \log_2 \frac{P(\boldsymbol{y})}{\prod_i P(y_i)} dy$$

  - Non-negative quantity
  - Reaches zero if and only if all variables are statistically independent. How?

Finding $\mathbf{V}$ using Information Theory

- Goal: Find the last rotation $\mathbf{V}$ such that estimate of $\hat{\boldsymbol{s}}$ is statistically independent
- Why mutual information?
- The mutual information measures the departure of two variables from statistical independence
- We use *multi-information*, a generalization of mutual information, which measures the statistical dependence between multiple variables:

$$I(y) = \int P(\boldsymbol{y}) \log_2 \frac{P(\boldsymbol{y})}{\prod_i P(y_i)} dy$$

  - Non-negative quantity
  - Reaches zero if and only if all variables are statistically independent. How?
  - Ex: if $P(\boldsymbol{y}) = \prod_i P(y_i)$, then $\log(1) = 0$ and $I(\boldsymbol{y}) = 0$)

Finding **V** using Information Theory

- Goal of ICA now becomes
  - Find a rotation matrix **V**, such that $I(\hat{\boldsymbol{s}}) = 0$ where $\hat{\boldsymbol{s}} = \boldsymbol{V}\boldsymbol{x}_w$
- Minimizing the multi-information is difficult in practice but can be simplified
- The multi-information is a function of the entropy H[·] of a distribution
- Entropy $H[\cdot] =$ - $\int P(\boldsymbol{y}) \log_2 P(\boldsymbol{y}) d\boldsymbol{y}$ measures the amount of uncertainty about a distribution P(y)
- The multi information is $I(\boldsymbol{y}) = \sum_i H[y_i] - H[\boldsymbol{y}]$
- Difference between the sum of entropies of the marginal distributions and the entropy of the joint distribution
- The multi information of $\hat{\boldsymbol{s}}$ is:

$$I(\hat{\boldsymbol{s}}) = \sum_i H[(\boldsymbol{V}\boldsymbol{x}_w)_i] - H[(\boldsymbol{V}\boldsymbol{x}_w)]$$

$$= \sum_i H[(\boldsymbol{V}\boldsymbol{x}_w)_i] - (H[\boldsymbol{x}_w] + \log_2 |\boldsymbol{V}|)$$

Finding $\mathbf{V}$ using Information Theory

- The optimization is simplified further by recognizing that we are only interested in finding the rotation matrix

$$\boldsymbol{V} = \underset{v}{\operatorname{argmin}} \sum_{i} H[(\boldsymbol{V}\boldsymbol{x}_w)_i]$$

- The optimization has simplified to finding a rotation matrix that minimizes the sum of the marginal entropies of $\hat{\boldsymbol{s}}$
- In summary,
  - We have identified an optimization that permits us to estimate $\mathbf{V}$
  - Reconstruct the original statistically independent source signals

$$\hat{\boldsymbol{s}} = \boldsymbol{V}\boldsymbol{x}_w$$

  - The columns of $\boldsymbol{W}^{-1}$ are the independent components of data

To summarize ICA in one slide

1. Subtract off the mean of the data in each dimension.
2. Whiten the data by calculating the eigenvectors of the covariance of the data
3. Identify final rotation matrix that optimizes statistical independence

PCA vs ICA

- Blind source separation: ICA(✓) PCA(N)
- What happens when you apply PCA on transpose of the data (Geometrically)?

PCA vs ICA

- Blind source separation: ICA(✓) PCA(N)
- What happens when you apply PCA on transpose of the data (Geometrically)?
- Highly Directional: ICA(✓) PCA(N)
- Example of Face detection
- What would PCA find when applied on face dataset?

PCA vs ICA

- Blind source separation: ICA(✓) PCA(N)
- What happens when you apply PCA on transpose of the data (Geometrically)?
- Highly Directional: ICA(✓) PCA(N)
- Example of Face detection
- What would PCA find when applied on face dataset?
- Brighness, Average Face (Things that matter to reconstruct image)
- What would ICA find when applied on face dataset?

PCA vs ICA

- Blind source separation: ICA(✓) PCA(N)
- What happens when you apply PCA on transpose of the data (Geometrically)?
- Highly Directional: ICA(✓) PCA(N)
- Example of Face detection
- What would PCA find when applied on face dataset?
- Brighness, Average Face (Things that matter to reconstruct image)
- What would ICA find when applied on face dataset?
- Independent components like eyes, nose hair etc

Canonical Correlation Analysis - A gentle introduction [3]

- PCA: Transformation of possible correlated variables into uncorrelated variables known as principal components without using any labels
- ICA: Not only decorrelates data but also attempts to make signals as statistically independent as possible
- CCA: Is a method of correlating linear relationships between two multidimensional variable
  - CCA makes use of two views of the same semantic object to extract the representation of the semantics
  - CCA can be seen as the problem of finding basis vectors for two sets of variables such that the correlation between the projections of the variables onto these basis vectors are mutually maximised
  - The main difference between CCA and the other three methods is that CCA is closely related to mutual information

[3] Canonical correlation analysis; An overview with application to learning methods - David R. Hardoon , Sandor Szedmak and John Shawe-Taylor

Canonical Correlation Analysis - Theoretical Foundations

- CCA: finding basis vectors for two sets of variables such that the correlation between the projections of the variables on to these basis vectors are mutually maximised
- CCA seeks a pair of linear transformations one for each of the sets of variables such that when the set of variables are transformed the corresponding co-ordinates are maximally correlated.
- Consider a multivariate random vector of from $(\boldsymbol{x}, \boldsymbol{y})$
- Let the given sample instances are

$$((\boldsymbol{x}_1, \boldsymbol{y}_1), ((\boldsymbol{x}_2, \boldsymbol{y}_2), \cdots, ((\boldsymbol{x}_n, \boldsymbol{y}_n))$$

- Use $\boldsymbol{S}_x$ to denote $(\boldsymbol{x}_1, \cdots \boldsymbol{x}_n)$ and similarly $\boldsymbol{S}_y$ to denote $(\boldsymbol{y}_1, \cdots \boldsymbol{y}_n)$
- We want to define a new coordinate for $\mathbf{x}$ by choosing a direction $\boldsymbol{w}_x$ and projecting $\mathbf{x}$ onto that direction

$$\boldsymbol{x} \rightarrow \langle \boldsymbol{w}_x, \boldsymbol{x} \rangle$$

Canonical Correlation Analysis - Theoretical Foundations cont..

- After mapping the new co-ordinate system becomes

$$S_{x,\boldsymbol{w}_x} = (\langle \boldsymbol{w}_x, \boldsymbol{x}_1 \rangle, \cdots, \langle \boldsymbol{w}_x, \boldsymbol{x}_n \rangle)$$

- The corresponding **y** coordinate becomes

$$S_{y,\boldsymbol{w}_y} = (\langle \boldsymbol{w}_y, \boldsymbol{y}_1 \rangle, \cdots, \langle \boldsymbol{w}_y, \boldsymbol{x}_y \rangle)$$

- We want CCA to choose $\boldsymbol{w}_x$ and $\boldsymbol{w}_y$ to maximise the correlation between two vectors

$$\rho = \max_{\boldsymbol{w}_x, \boldsymbol{w}_y} corr(S_x \boldsymbol{w}_x, S_y \boldsymbol{w}_y)$$

$$= \max_{\boldsymbol{w}_x, \boldsymbol{w}_y} \frac{\langle S_x \boldsymbol{w}_x, S_y \boldsymbol{w}_y \rangle}{||S_x \boldsymbol{w}_x|| ||S_y \boldsymbol{w}_y||}$$

Canonical Correlation Analysis - Theoretical Foundations cont..

- Rewriting the above function as empirical expectation:

$$\rho = \max_{\boldsymbol{w}_x, \boldsymbol{w}_y} \frac{\mathbb{E}[\langle \boldsymbol{w}_x, \boldsymbol{x} \rangle \langle \boldsymbol{w}_y, \boldsymbol{y} \rangle]}{\sqrt{\mathbb{E}[\langle \boldsymbol{w}_x, \boldsymbol{x} \rangle^2] \mathbb{E}[\langle \boldsymbol{w}_y, \boldsymbol{y} \rangle^2]}}$$

$$= \max_{\boldsymbol{w}_x, \boldsymbol{w}_y} \frac{\mathbb{E}[\boldsymbol{w}_x^T \boldsymbol{x} \boldsymbol{y}^T \boldsymbol{w}_y]}{\sqrt{\mathbb{E}[\boldsymbol{w}_x^T \boldsymbol{x} \boldsymbol{x}^T \boldsymbol{w}_x] \mathbb{E}[\boldsymbol{w}_y^T \boldsymbol{y} \boldsymbol{y}^T \boldsymbol{w}_y]}}$$

$$= \max_{\boldsymbol{w}_x, \boldsymbol{w}_y} \frac{\boldsymbol{w}_x^T \mathbb{E}[\boldsymbol{x} \boldsymbol{y}^T] \boldsymbol{w}_y}{\sqrt{\boldsymbol{w}_x^T \mathbb{E}[\boldsymbol{x} \boldsymbol{x}^T] \boldsymbol{w}_x \boldsymbol{w}_y^T \mathbb{E}[\boldsymbol{y} \boldsymbol{y}^T] \boldsymbol{w}_y}}$$

- Use the definition of Covariance matrix to re-write above $(\boldsymbol{x}, \boldsymbol{y})$ is:

$$C(\boldsymbol{x}, \boldsymbol{y}) = \mathbb{E}[xy^T] = \begin{bmatrix} C_{xx} & C_{xy} \\ C_{yx} & C_{yy} \end{bmatrix}$$

Canonical Correlation Analysis - Theoretical Foundations cont..

- C_{xx} is within-sets covariance matrices
- $C_{xy} = C_{yx}^T$ is between-sets covariance matrices
- Rewriting above ρ we have

$$\rho = \max_{\boldsymbol{w}_x, \boldsymbol{w}_y} \frac{\boldsymbol{w}_x^T C_{xy} \boldsymbol{w}_y}{\sqrt{\boldsymbol{w}_x^T C_{xx} \boldsymbol{w}_x \boldsymbol{w}_y^T C_{yy} \boldsymbol{w}_y}}$$

the maximum canonical correlation is the maximum of ρ with respect to $\boldsymbol{w}_x$ and $\boldsymbol{w}_y$

Canonical Correlation Analysis - Algorithm

- Observe that final optimizing problem is not scale invariant
- If $\boldsymbol{w}_x$ is a solution then $\alpha \cdot \boldsymbol{w}_x$
- How do we deal with this?
- Impose a restriction on solution
  - $\boldsymbol{w}_x^T C_{xx} \boldsymbol{w}_x = 1$
  - $\boldsymbol{w}_y^T C_{yy} \boldsymbol{w}_y = 1$
- Solving the constraint optimization using the Lagrangian

$$L(\lambda, \boldsymbol{w}_x, \boldsymbol{w}_y) = \boldsymbol{w}_x^T C_{xy} \boldsymbol{w}_x - \frac{\lambda_x}{2}(\boldsymbol{w}_x^T C_{xx} \boldsymbol{w}_x - 1) - \frac{\lambda_y}{2}(\boldsymbol{w}_y^T C_{yy} \boldsymbol{w}_y - 1)$$

- Taking the derivative w.r.t $\boldsymbol{w}_x$ and $\boldsymbol{w}_y$

$$\frac{\partial f}{\partial \boldsymbol{w}_x} = C_{xy} \boldsymbol{w}_y - \lambda_x C_{xx} \boldsymbol{w}_x = 0$$

$$\frac{\partial f}{\partial \boldsymbol{w}_y} = C_{yx} \boldsymbol{w}_x - \lambda_y C_{yy} \boldsymbol{w}_y = 0$$

Canonical Correlation Analysis - Algorithm - Cont..

- After doing slight trick i.e "Subtracting $\boldsymbol{w}_y^T$ times the second equation from $\boldsymbol{w}_x^T$ times the first we finally get
- $\boldsymbol{w}_y = \frac{C_{yy}^{-1} C_{yx} \boldsymbol{w}_x}{\lambda}$
  - Letting $\lambda_x = \lambda_y = \lambda$
  - Assuming C_{yy} is invertible
- Substituting $\boldsymbol{w}_y$ in $\frac{\partial f}{\partial \boldsymbol{w}_x}$, we get

$$\frac{C_{xy} C_{yy}^{-1} C_{yx} \boldsymbol{w}_x}{\lambda} - \lambda C_{xx} \boldsymbol{w}_x = 0$$

$$C_{xy} C_{yy}^{-1} C_{yx} \boldsymbol{w}_x = \lambda^2 C_{xx} \boldsymbol{w}_x$$

- Can you guess what's the above problem is called?
- Generalized Eigen value problem $Ax = \lambda Bx$

Canonical Correlation Analysis - Algorithm - Cont..

- After doing slight trick i.e "Subtracting $\boldsymbol{w}_y^T$ times the second equation from $\boldsymbol{w}_x^T$ times the first we finally get
- $\boldsymbol{w}_y = \frac{C_{yy}^{-1} C_{yx} \boldsymbol{w}_x}{\lambda}$
  - Letting $\lambda_x = \lambda_y = \lambda$
  - Assuming C_{yy} is invertible
- Substituting $\boldsymbol{w}_y$ in $\frac{\partial f}{\partial \boldsymbol{w}_x}$, we get

$$\frac{C_{xy} C_{yy}^{-1} C_{yx} \boldsymbol{w}_x}{\lambda} - \lambda C_{xx} \boldsymbol{w}_x = 0$$

$$C_{xy} C_{yy}^{-1} C_{yx} \boldsymbol{w}_x = \lambda^2 C_{xx} \boldsymbol{w}_x$$

- Can you guess what's the above problem is called?
- Generalized Eigen value problem $Ax = \lambda Bx$

Canonical Correlation Analysis - Algorithm -In summary

1. Find $\boldsymbol{w}_x$ by solving the generalized eigen value
2. The use this $\boldsymbol{w}_x$ to get $\boldsymbol{w}_y$

Fisher's Linear Discriminant [4]

- One way to view a linear classification model is in terms of dimensionality reduction
- Consider first the case of two classes, and suppose we take the D-dimensional input vector x and project it down to one dimension using

$$y = w^T x$$

- If we place a threshold on y and classify $y \geq -w_0$ as class C_1 , and otherwise class C_2
- In general, the projection onto one dimension leads to a considerable loss of information
- Classes that are well separated in the original D-dimensional space may become strongly overlapping in one dimension
- However, by adjusting the components of the weight vector w, we can select a projection that maximizes the class separation

[4] Pattern Recognition and Machine Learning - Christopher Bishop

Fisher cont..

- Consider a two-class problem in which there are N_1 points of class C_1 and N_2 points of class C_2
- The mean vectors of the two classes are given by

$$\boldsymbol{m_1} = \frac{1}{N_1} \sum_{n \in C_1} \boldsymbol{x}_n$$

$$\boldsymbol{m_2} = \frac{1}{N_2} \sum_{n \in C_2} \boldsymbol{x}_n$$

- What's simplest measure of the separation of the classes, when projected onto $\boldsymbol{w}$?
- Separation of the projected class means, choose w so as to maximize

$$m_2 - m_1 = \boldsymbol{w}^T(\boldsymbol{m}_2 - \boldsymbol{m}_1)$$

s.t

$$||\boldsymbol{w}|| = 1$$

Fisher cont..

- Consider a two-class problem in which there are N_1 points of class C_1 and N_2 points of class C_2
- The mean vectors of the two classes are given by

$$\boldsymbol{m_1} = \frac{1}{N_1} \sum_{n \in C_1} \boldsymbol{x}_n$$

$$\boldsymbol{m_2} = \frac{1}{N_2} \sum_{n \in C_2} \boldsymbol{x}_n$$

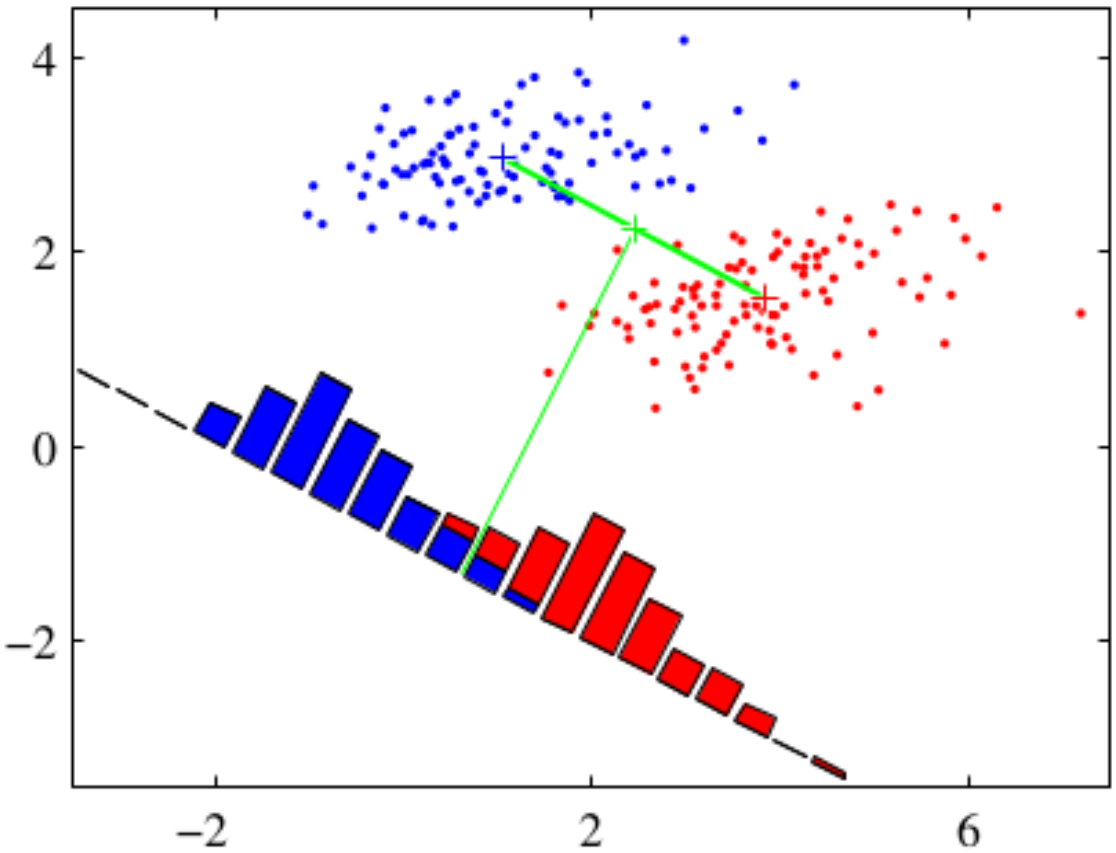
- What's simplest measure of the separation of the classes, when projected onto $\boldsymbol{w}$?
- Separation of the projected class means, choose w so as to maximize

$$m_2 - m_1 = \boldsymbol{w}^T(\boldsymbol{m}_2 - \boldsymbol{m}_1)$$

s.t

$$||\boldsymbol{w}|| = 1$$

Fisher cont..



The plot shows samples from two classes (depicted in red and blue) along with the histograms resulting from projection onto the line joining the class means. Note that there is considerable class overlap in the projected space.

Modelling Fisher Linear Discriminant problem

- How do we solve this

$$m_2 - m_1 = \boldsymbol{w}^T(\boldsymbol{m}_2 - \boldsymbol{m}_1)$$

- Use a Lagrangian to perform constrained Minimization
- We find $\boldsymbol{w} \propto (\boldsymbol{m_2} - \boldsymbol{m_1})$
- Whats the problem with this approach?
- Hint: Strongly non-diagonal covariances of the class distributions

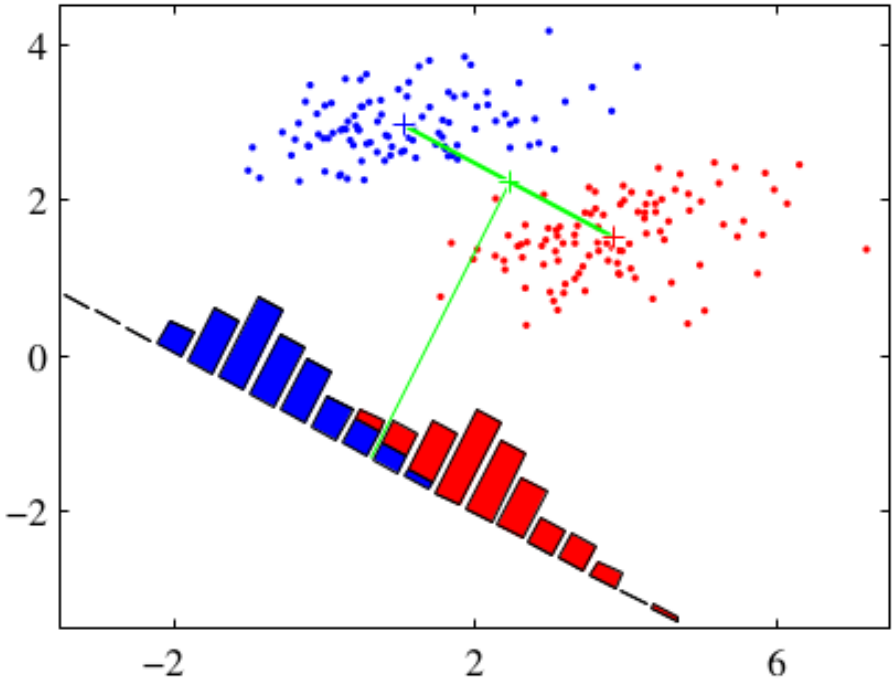
Modelling Fisher Linear Discriminant problem

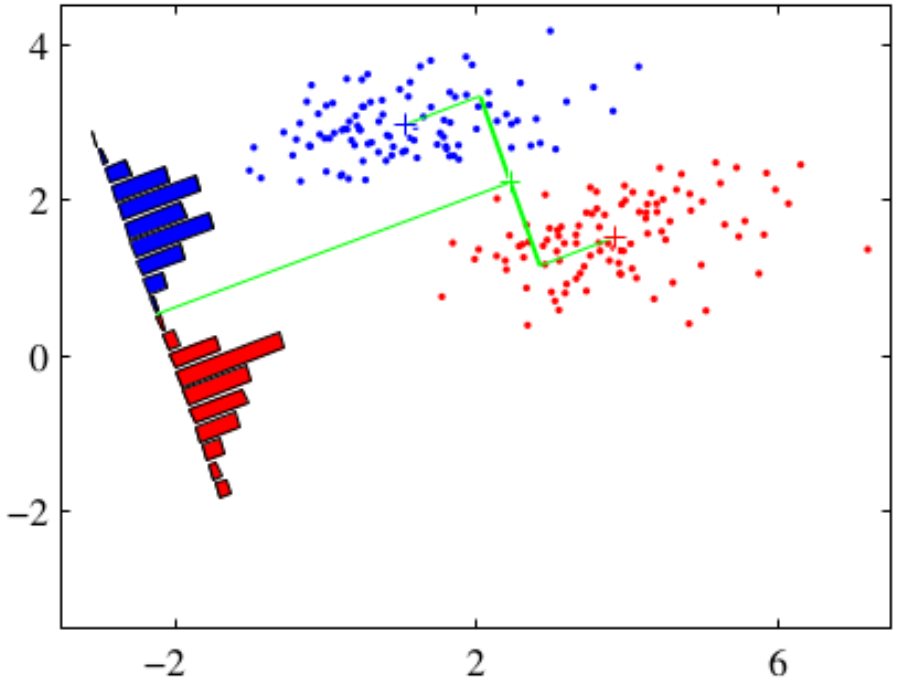
- How do we solve this

$$m_2 - m_1 = \boldsymbol{w}^T(\boldsymbol{m}_2 - \boldsymbol{m}_1)$$

- Use a Lagrangian to perform constrained Minimization
- We find $\boldsymbol{w} \propto (\boldsymbol{m_2} - \boldsymbol{m_1})$
- Whats the problem with this approach?
- Hint: Strongly non-diagonal covariances of the class distributions

Modelling Fisher Linear Discriminant problem

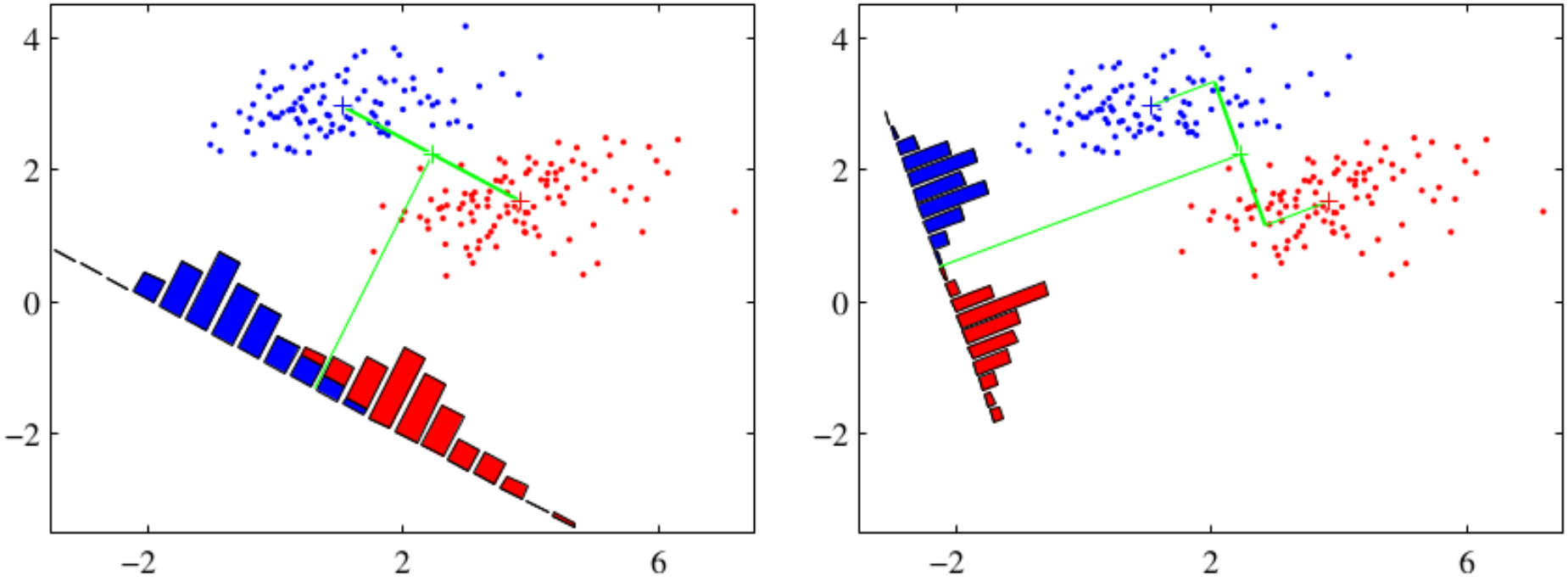




What is the difference and Key Idea?

Fisher Idea: maximize a function that will give a large separation between the projected class means while also giving a small variance within each class

Modelling Fisher Linear Discriminant problem



What is the difference and Key Idea?

Fisher Idea: maximize a function that will give a large separation between the projected class means while also giving a small variance within each class

Modelling Fisher Linear Discriminant problem

- The within-class variance of the transformed data from class C_k is:

$$s_k^2 = \sum_{n \in C_k} (y_n - m_k)^2$$

- where $y_n = \boldsymbol{w}^T \boldsymbol{x}_n$
- Define the total within-class variance for the whole data set as $s_1^2 + s_2^2$
- The Fisher criterion is defined as the ratio of the between-class variance to the within-class variance and is given by

$$J(\boldsymbol{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}$$

- By simple substitutions we get

$$J(\boldsymbol{w}) = \frac{\boldsymbol{w}^T \boldsymbol{S}_B \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S}_W \boldsymbol{w}}$$

- $\boldsymbol{S}_B$ is the *between-class* covariance matrix
- $\boldsymbol{S}_W$ is the *within-class* covariance matrix

Modelling Fisher Linear Discriminant problem

- $J(\boldsymbol{w}) = \frac{\boldsymbol{w}^T \boldsymbol{S}_B \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S}_W \boldsymbol{w}}$
- $\boldsymbol{S}_B = (\boldsymbol{m}_2 - \boldsymbol{m}_1)(\boldsymbol{m}_2 - \boldsymbol{m}_1)^T$
- $\boldsymbol{S}_w = \sum_{n \in C_1} (\boldsymbol{x}_n - \boldsymbol{m}_1)(\boldsymbol{x}_n - \boldsymbol{m}_1)^T + \sum_{n \in C_2} (\boldsymbol{x}_n - \boldsymbol{m}_2)(\boldsymbol{x}_n - \boldsymbol{m}_2)^T$
- Differentiating $J(\boldsymbol{w})$ w.r.t $\boldsymbol{w}$, we can see J is maximized when

$$(\boldsymbol{w}^T \boldsymbol{S}_B \boldsymbol{w}) \boldsymbol{S}_W \boldsymbol{w} = (\boldsymbol{w}^T \boldsymbol{S}_W \boldsymbol{w}) \boldsymbol{S}_B \boldsymbol{w}$$

- We see that $\boldsymbol{S}_B w$ is always in the direction of $(\boldsymbol{m}_2 - \boldsymbol{m}_1)$
- Ignoring the magnitude of $\mathbf{w}$ we can drop the scalar factors $(\boldsymbol{w}^T \boldsymbol{S}_B \boldsymbol{w})$ and $((\boldsymbol{w}^T \boldsymbol{S}_W \boldsymbol{w}))$
- Multiplying both sides of (1) $\boldsymbol{S}_W^{-1}$ we obtain

$$\boldsymbol{w} \propto \boldsymbol{S}_W^{-1} (\boldsymbol{m}_2 - \boldsymbol{m}_1)$$

- What if the *within-class* covariance is isotropic?

Non Linear Methods

- Multidimensionality scaling
  - Given the distances or similarities can you calculate the original locations of data points
  - Airport distance chart example
- Manifold learning

Thank you
&
Wish you a safe and Happy Diwali !!