Module 8

Part 2 – Unsupervised Learning

Dimensionality Reduction

José Manuel García Nieto – University of Málaga

Table of contents

- Principal Component Analysis (PCA)
- Non-Negative Matrix Factorization (NMF)
- Linear Discriminant Analysis (LDA)
- Practical Notes

Principal Component Analysis (PCA)

- Purposes of the principal component analysis (PCA)
  - Transform a set of correlated variables into another set of uncorrelated variables.
  - Get a new set of orthogonal vectors (principal components or PCs).
  - Order the new variables from the largest to the smallest variance.
- Terminology

a) Loading

- Normalized principal component (PC)
- There are as many loading vectors as the number of variables.

b) Variance σ^2 (or standard deviation σ)

- The principal components have associated variances.

c) Transformed scores

- Raw scores (original observations) represented using the PCs as new coordinate axes.

Principal Component Analysis (PCA)

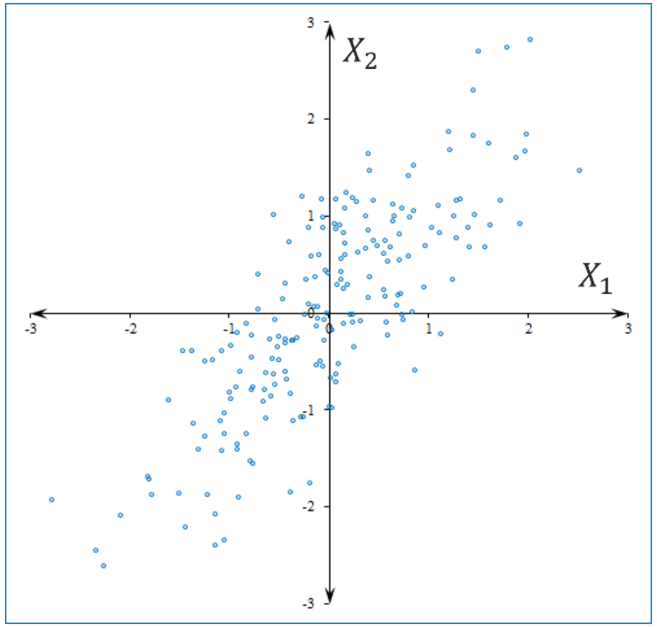
- Principal Components
  - Let's suppose that there are k variables or features $X1$, $X2$, …, Xk.
  - Then, the $PC1$, $PC2$, …, PCk are linear combinations of the original features:

    $PCi{=}\alpha 1,\ iX1{+}\alpha 2,\ iX2{+}\ldots{+}\alpha k,\ iXk$
  - Conversely, the original features can be expressed in terms of the PCs:

    $Xi = \beta 1,\ iPC1 + \beta 2,\ iPC2{+}\ldots{+}\beta k,\ iPCk$

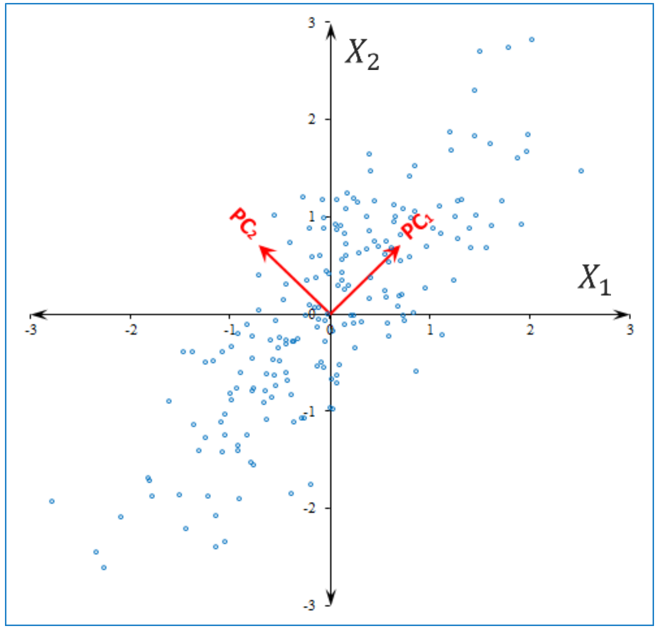
Principal Component Analysis (PCA)

- Principal Component and variance
  - Suppose a dataset with two variables that can be conveniently visualized on a plane:



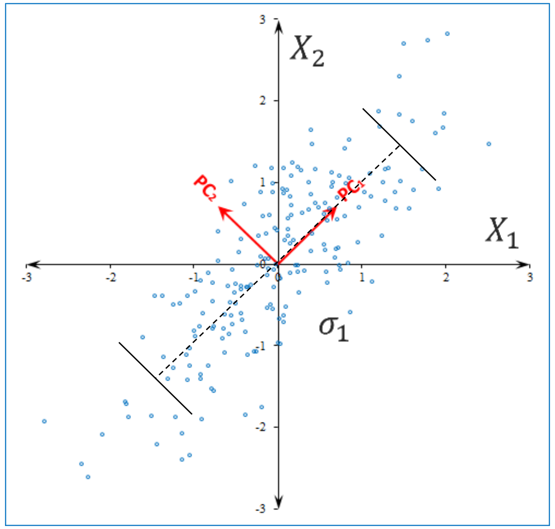
Principal Component Analysis (PCA)

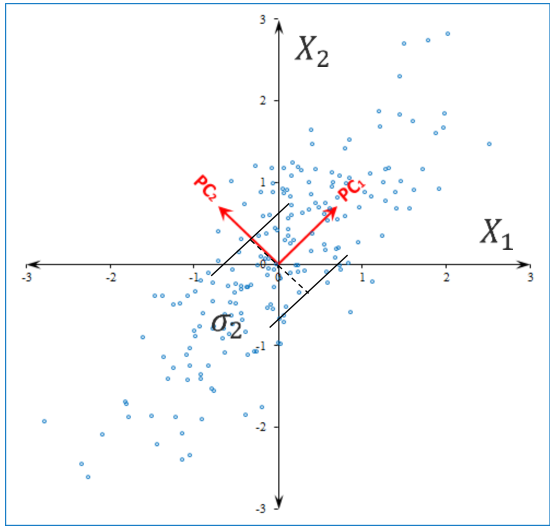
- Principal Component and variance
  - We can find the $PC1$ and $PC2$ that are orthogonal to each other



Principal Component Analysis (PCA)

- Principal Component and variance
  - $PC1$ is the direction of the largest variance, while $PC2$ is the direction of the next largest variance

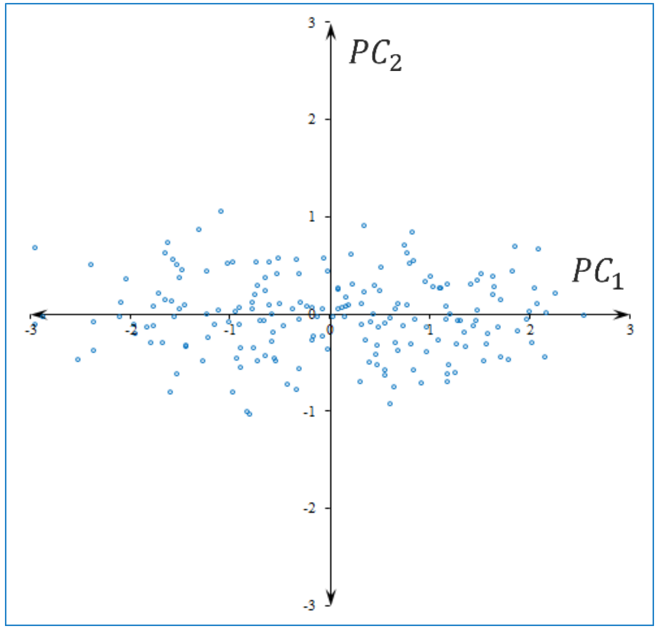




  - So, we have $\sigma 1 > \sigma 2$

Principal Component Analysis (PCA)

- Transformed scores
  - The observations can be represented using the $PC1$ and $PC2$ as new coordinate axes



Principal Component Analysis (PCA)

- Cumulative variance
  - As the principal components can be regarded as independent variables, the total variance is:

$$\sigma_{total}^2 = \sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \cdots$$

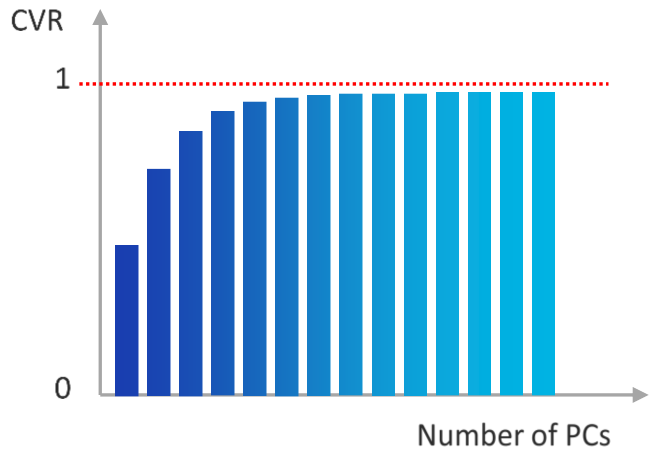
  - So, we can calculate the cumulative variance ratios (CVRs):

$$CVR_1 = \frac{\sigma_1^2}{\sigma_{total}^2}$$

$$CVR_2 = \frac{\sigma_1^2 + \sigma_2^2}{\sigma_{total}^2}$$

$$CVR_3 = \frac{\sigma_1^2 + \sigma_2^2 + \sigma_3^2}{\sigma_{total}^2}$$

$$\vdots$$



Principal Component Analysis (PCA)

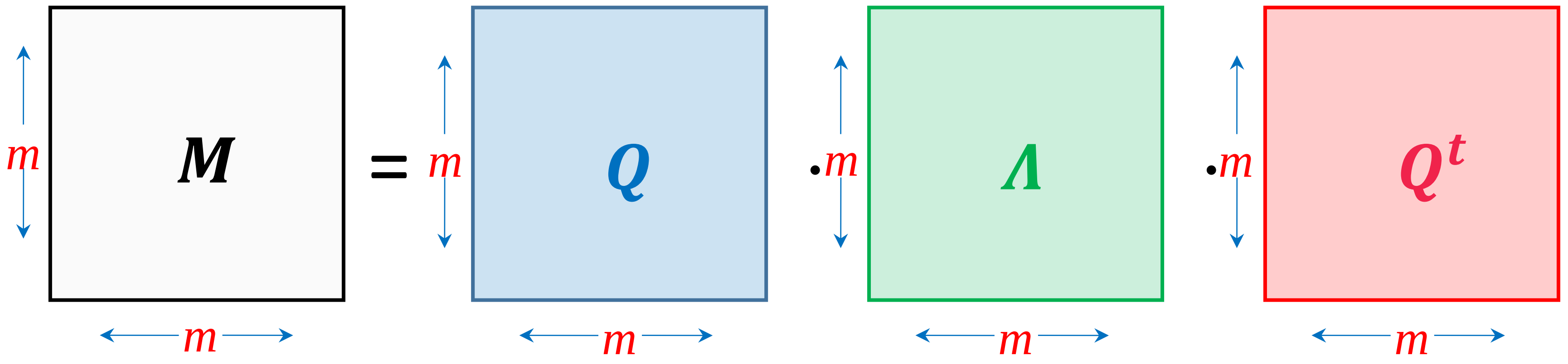
- Calculating the principal components
  - The PCs can also be obtained by eigenvalue decomposition (ED) of the covariance matrix.
  - The PCs can be calculated by singular value decomposition (SVD) of the data matrix.
  - If we standardize the variables, we would have the correlation instead of the covariance.

    - A covariance matrix and a correlation matrix

$$\begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 \end{bmatrix} \qquad \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} \\ \rho_{21} & 1 & \rho_{23} & \rho_{24} \\ \rho_{31} & \rho_{32} & 1 & \rho_{34} \\ \rho_{41} & \rho_{42} & \rho_{43} & 1 \end{bmatrix}$$

Principal Component Analysis (PCA)

- Matrix Decompositions
  - Eigenvalue decomposition (ED)
    - A square matrix $\boldsymbol{M}$ is decomposed as $\boldsymbol{M}=\boldsymbol{Q}\ \boldsymbol{\Lambda}\ \boldsymbol{Q}^t$.
    - All the matrices have the same size: $Size(\boldsymbol{M})=Size(\boldsymbol{Q})=Size(\boldsymbol{\Lambda})=m\times m$.



Principal Component Analysis (PCA)

- Eigenvalue decomposition (ED)

  - A square matrix $\boldsymbol{M}$ is decomposed as $\boldsymbol{M}=\boldsymbol{Q}\ \boldsymbol{\Lambda}\ \boldsymbol{Q}^t$.
  - Here, $\boldsymbol{\Lambda}$ is a diagonal matrix that contains the "eigenvalues."

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_m \end{bmatrix}$$

  - A square matrix $\boldsymbol{M}$ is decomposed as $\boldsymbol{M}=\boldsymbol{Q}\ \boldsymbol{\Lambda}\ \boldsymbol{Q}^t$.
  - The columns of $\boldsymbol{Q}$ are the so-called "eigenvectors."

$$\boldsymbol{Q} = \begin{bmatrix} \uparrow & \cdots & \uparrow \\ \boldsymbol{q_1} & \cdots & \boldsymbol{q_m} \\ \downarrow & \cdots & \downarrow \end{bmatrix}$$

  - Between an eigenvector and its eigenvalue, we have the following relation:

$$\boldsymbol{M}\boldsymbol{q_i} = \lambda_i\ \boldsymbol{q_i}$$

  - Between any two eigenvectors, we have the following orthogonality condition:

$$\boldsymbol{q_i} \cdot \boldsymbol{q_j} = \delta_{ij} \quad \Leftrightarrow \quad \boldsymbol{Q}\boldsymbol{Q}^t = \boldsymbol{Q}^t\boldsymbol{Q} = \boldsymbol{I}$$

Applications of PCA

- Dimensional Reduction
  - The number of principal components (PCs) is equal to the number of variables, say k.
  - The original variables X_i can be expressed in terms of the PCs:

$$X_i = \beta_{1,i}PC_1 + \beta_{2,i}PC_2 + \cdots + \beta_{k,i}PC_k$$

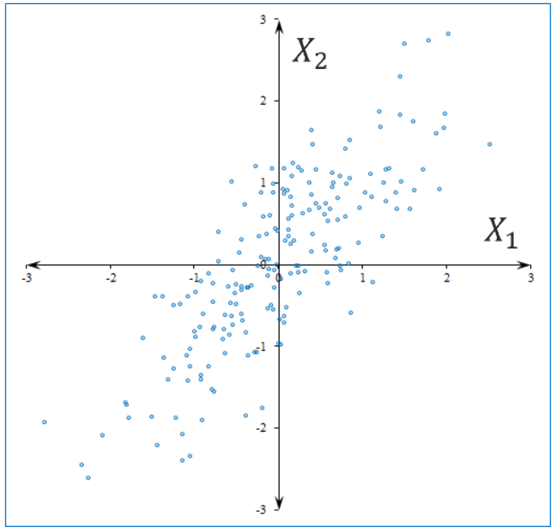
  - The PCs are ordered by variance: $\sigma_1^2 > \sigma_2^2 > \sigma_3^2 > \cdots > \sigma_k^2$
  - Thus, we can reduce dimension starting from the last PC, ($q < k$):

$$X_i \approx \beta_{1,i}PC_1 + \beta_{2,i}PC_2 + \cdots + \beta_{q,i}PC_q$$ "Reduced dimension input"

- ✓ Pros
  - ✓ We can simplify the data and reduce overfitting error.
  - ✓ We get only the most salient features.
- ✓ Cons
  - ✓ Loss of details
  - ✓ It is difficult to interpret intuitively.

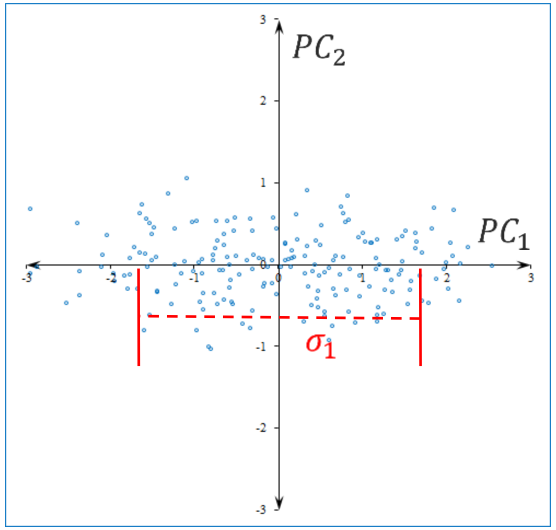
Applications of PCA

- Dimensional Reduction
  - Suppose a dataset with two variables that can be conveniently visualized on a plane:



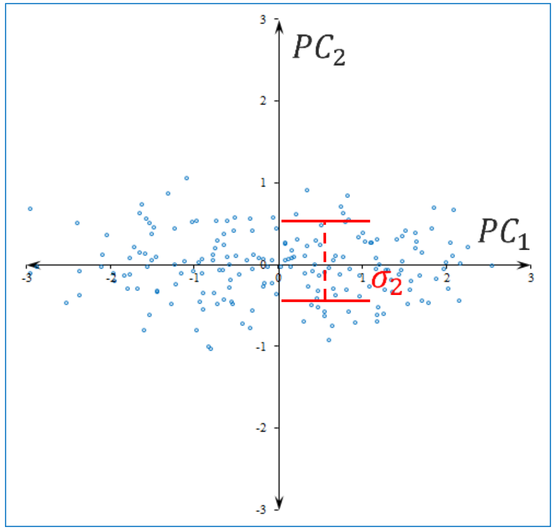
Applications of PCA

- Dimensional Reduction
  - The observations can be represented using the $PC1$ and $PC2$ as new coordinate axes:



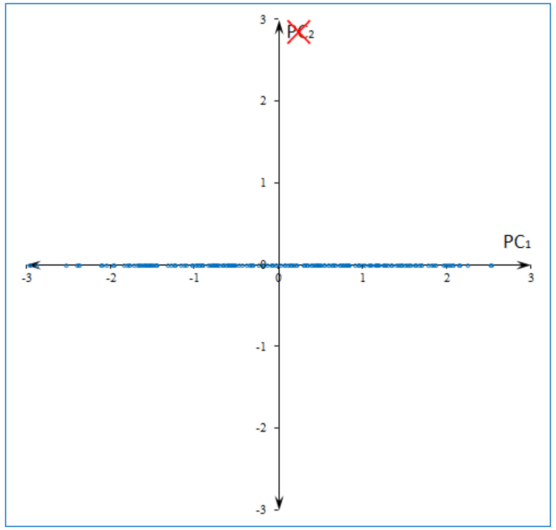
Applications of PCA

- Dimensional Reduction
  - The observations can be represented using the $PC1$ and $PC2$ as new coordinate axes:



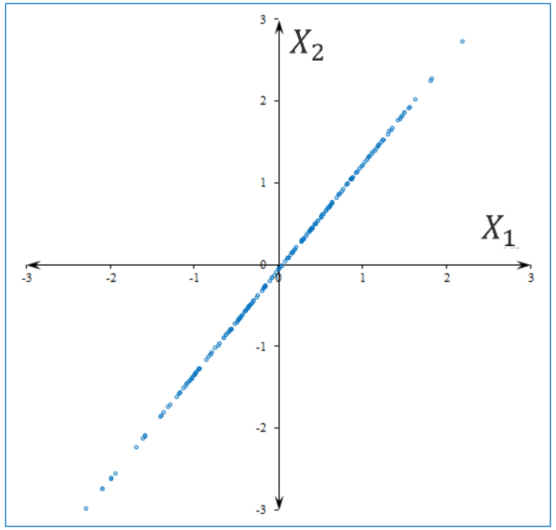
Applications of PCA

- Dimensional Reduction
  - We can eliminate the direction represented by PC_2 that corresponds to the smaller variance:



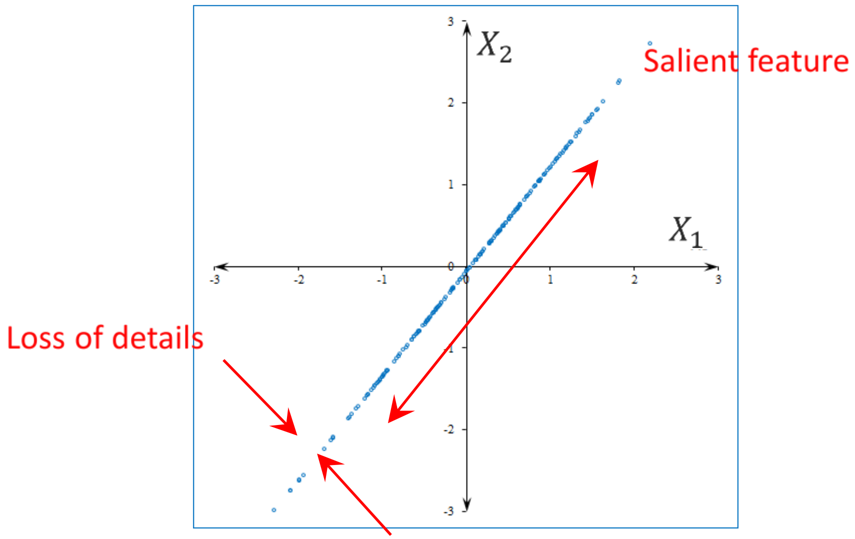
Applications of PCA

- Dimensional Reduction
  - Now, we can go back to the original coordinate system and show the “reduced dimensional input”:



Applications of PCA

- Dimensional Reduction
  - Now, we can go back to the original coordinate system and show the “reduced dimensional input”:



  - We can notice that details have been lost leaving only the most salient feature.

Applications of PCA

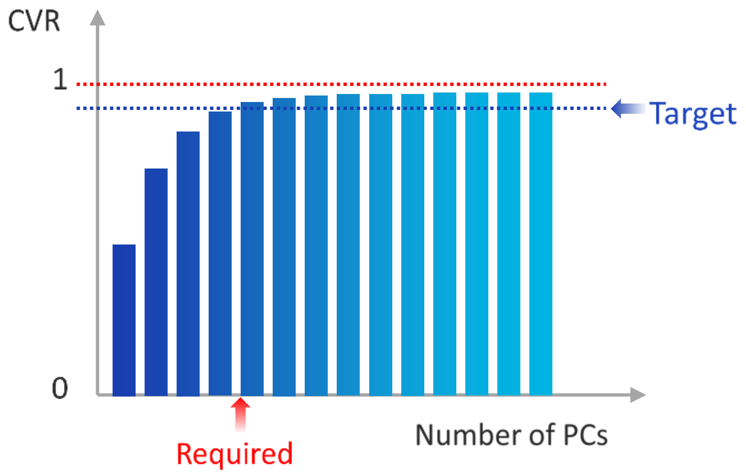
- Dimensional Reduction
  - The total variance is: $\sigma_{total}^2 = \sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \cdots + \sigma_k^2$
  - Then, we can calculate the cumulative variance ratios:

$$CVR_1 = \frac{\sigma_1^2}{\sigma_{total}^2}$$

$$CVR_2 = \frac{\sigma_1^2 + \sigma_2^2}{\sigma_{total}^2}$$

$$CVR_3 = \frac{\sigma_1^2 + \sigma_2^2 + \sigma_3^2}{\sigma_{total}^2}$$

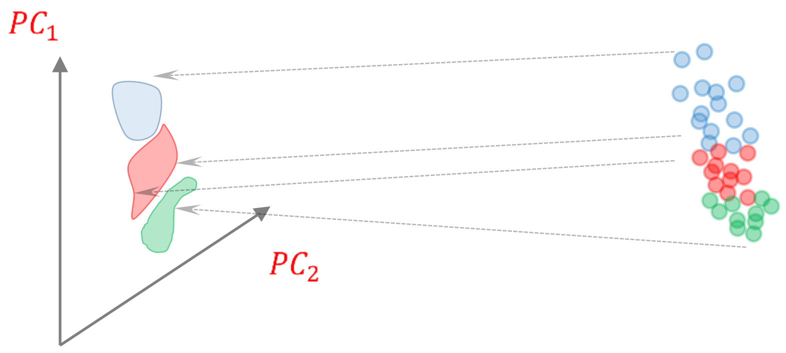
$$\vdots$$



  - We can set a target CVR and determine the required number of PCs.

Applications of PCA

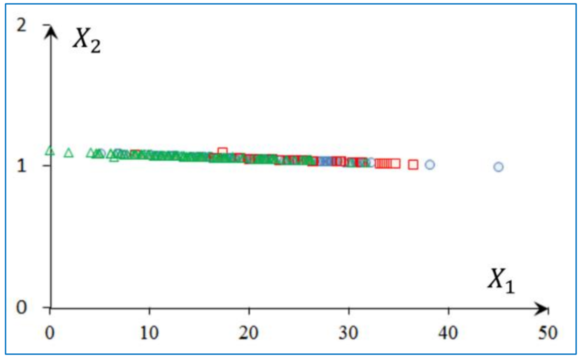
- High Dimension Visualization
  - $PC1$ and $PC2$ are the directions of the largest and the second largest variance.
  - $PC1$ and $PC2$ define the most spread-out plane on which to project the high dimensional coordinates

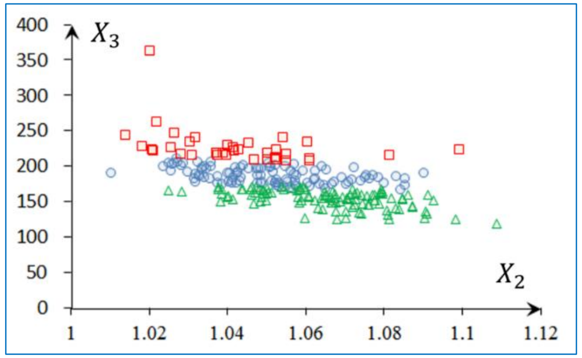


  - It is easy to implement as it uses only the first two components of the transformed scores.

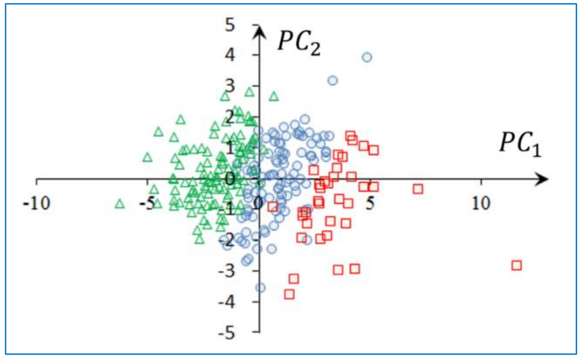
Applications of PCA

- High dimension visualization





Projected onto the original variable set



Projected onto the plane defined by PC_1 and PC_2

Table of contents

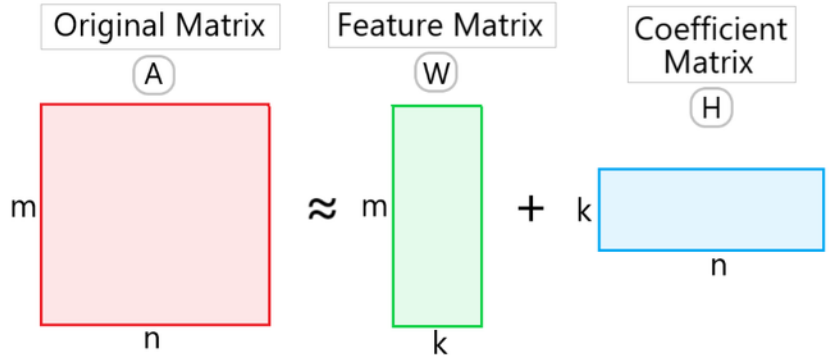
- Principal Component Analysis (PCA)
- Non-Negative Matrix Factorization (NMF)
- Linear Discriminant Analysis (LDA)
- Practical Notes

Non-Negative Matrix Factorization (NMF)

- Dimensional Reduction
  - Down a large dataset into smaller meaningful parts while ensuring that all values remain non-negative
    - For a matrix A of dimensions m × n where each element is ≥ 0 NMF factorizes it into two matrices W and H with dimensions m × k and k × n respectively where both matrices contain only non-negative elements:

$$A_{m\times n} \approx W_{m\times k} H_{k\times n}$$

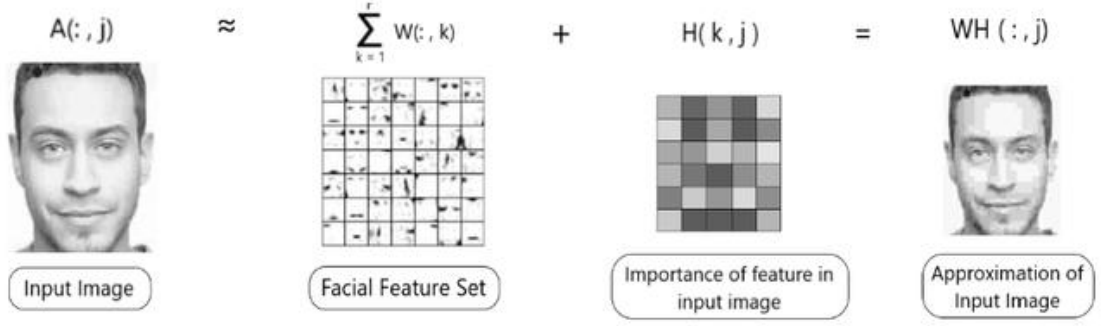
- **A** → Original input matrix (a linear combination of W and H)
- **W** → Feature matrix (basis components)
- **H** → Coefficient matrix (weights associated with W)
- **k** → Rank (dimensionality of the reduced representation where k ≤ min(m, n))



- NMF helps identify hidden patterns in data by assuming that each data point can be represented as a combination of fundamental features found in W

Non-Negative Matrix Factorization (NMF)

- Dimensional Reduction
  - NMF decomposes a data matrix A into two smaller matrices W and H using an iterative optimization process that minimizes reconstruction error:
    1. Initialization: Start with random non-negative values for W and H.
    2. Iterative Update: Modify W and H to minimize the difference between A and W × H.
    3. Stopping Criteria: The process stops when:
       - The reconstruction error stabilizes.
       - A set number of iterations is reached.



Common optimization techniques for NMF include:

- Multiplicative Update Rules: Ensures non-negativity by iteratively adjusting W and H.
- Alternating Least Squares (ALS): Solves for W while keeping H fixed, and vice versa, in an alternating manner.

Table of contents

- Principal Component Analysis (PCA)
- Non-Negative Matrix Factorization (NMF)
- Linear Discriminant Analysis (LDA)
- Practical Notes
    - *A. Dimensional reduction with PCA.*
        - *i. Read in data and explore.*
        - *ii. Visualize the data.*
        - *iii. Visualize the reduced dimensional input by PCA.*
        - *iv. Analysis of the cumulative variance ratio (CVR).*
    - *B. Dimensional reduction with NMF.*
        - *i. Visualize the reduced dimensional input by NMF.*
    - *C. Optimized high dimensional visualization with PCA.*
        - *i. Simulate data.*
        - *ii. Visualize on the plane defined by PC1 and PC2.*

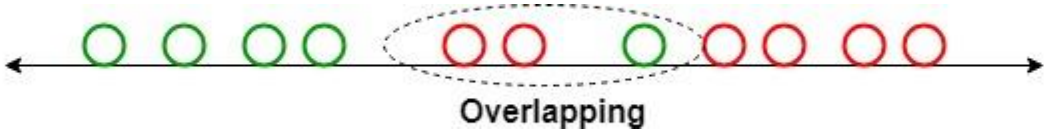
Table of contents

- Principal Component Analysis (PCA)
- Non-Negative Matrix Factorization (NMF)
- Linear Discriminant Analysis (LDA)
- Practical Notes

Linear Discriminant Analysis (LDA)

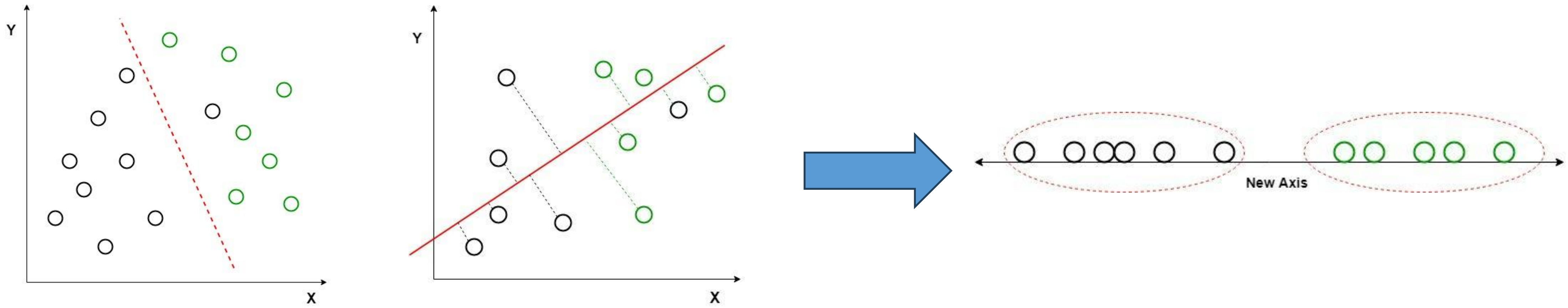
- Dimensional Reduction
  - Reducing the dimensionality of data while retaining the most significant features for classification tasks
  - It works by finding the **linear combinations of features that best separate the classes in the dataset**
  - Separate two or more classes by converting higher-dimensional data space into a lower-dimensional space

  - Core Assumptions of LDA
    - Gaussian Distribution: Data within each class should follow a Gaussian distribution.
    - Equal Covariance Matrices: Covariance matrices of the different classes should be equal.
    - Linear Separability: A linear decision boundary should be sufficient to separate the classes.



Linear Discriminant Analysis (LDA)

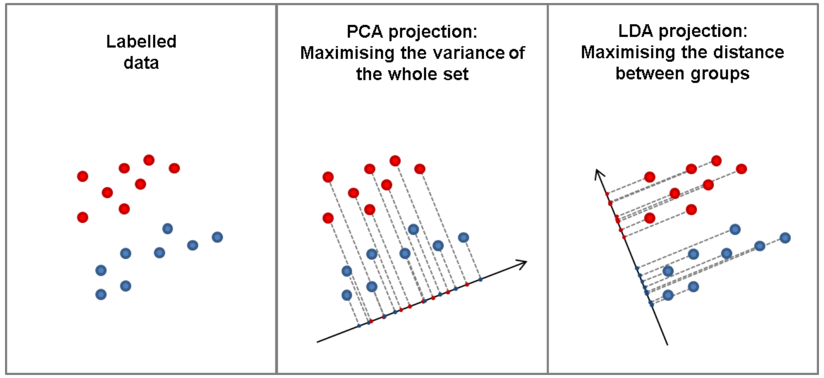
- Dimensional Reduction
  - How does LDA work?
    - LDA works by finding **directions in the feature space that best separate the classes**.
    - It does this by maximizing the difference between the class means while minimizing the spread within each class.



    - It uses both axes (X and Y) to generate a new axis in such a way that it maximizes the distance between the means of the two classes while minimizing the variation within each class.
    - This transforms the dataset into a space where the classes are better separated.
    - It shows how LDA creates a new axis to project the data and separate the two classes effectively along a linear path. But it fails when the mean of the distributions are shared as it becomes impossible for LDA to find a new axis that makes both classes linearly separable. -> In such cases, we use **non-linear discriminant analysis**.

Linear Discriminant Analysis (LDA)

- Dimensional Reduction
  - LDA versus PCA projections



  - Extensions to LDA
    - Quadratic Discriminant Analysis (QDA): Each class uses its own estimate of variance (or covariance) allowing it to handle more complex relationships.
    - Flexible Discriminant Analysis (FDA): Uses non-linear combinations of inputs such as splines to handle non-linear separability.
    - Regularized Discriminant Analysis (RDA): Introduces regularization into the covariance estimate to prevent overfitting.

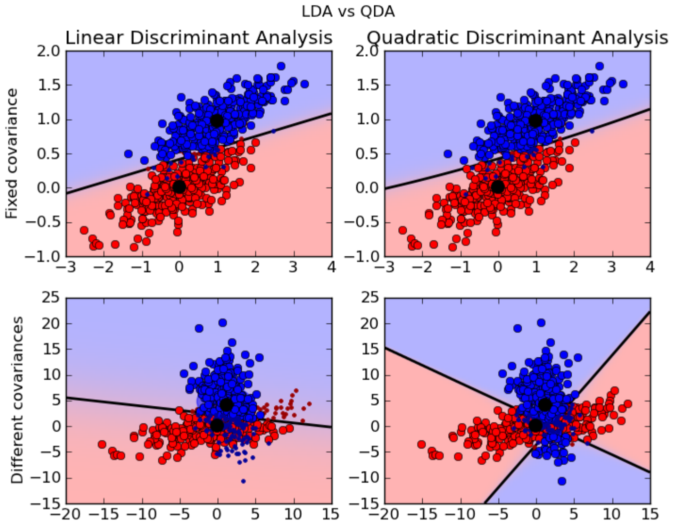


Table of contents

- Principal Component Analysis (PCA)
- Non-Negative Matrix Factorization (NMF)
- Linear Discriminant Analysis (LDA)
- Practical Notes