Principal Component Analysis

Preliminaries: Data matrix X as an nxp matrix

- Data matrix written with **features as columns** and **records as rows**

	Feature 1	**Feature 2**	**---**	**Feature p**
record 1				
record 2				
---				
record n				

- Thus the data matrix X is an nxp matrix
  - Note: Some books/articles follow the convention of features as rows and records as columns

Case Study: MNIST Handwritten Digits Database

MNIST Example

Yann Lecun

- MNIST: Popular dataset of handwritten images
- Lecun's famous 1998 work
    - LeNet-5
    - http://yann.lecun.com/exdb/mnist/ (http://yann.lecun.com/exdb/mnist/)

(MNIST Example By Josef Steppan - Own work, CC BY-SA 4.0 https://commons.wikimedia.org/w/index.php?curid=64810040 (https://commons.wikimedia.org/w/index.php?curid=64810040)")

MNIST Dataset Details

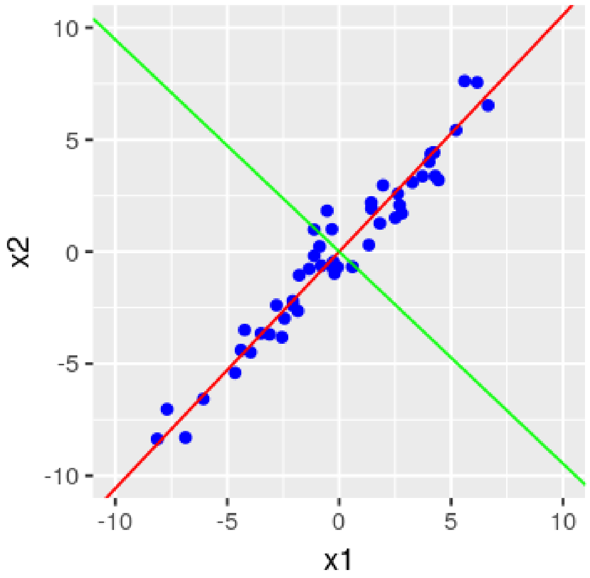
- Dataset of 60000 handwritten images
- Each image 28x28 pixels,
- Each pixel: 8 bit 'greyscale' value, ie 0 to 255
- Thus X is a 60000x784 data matrix:

	pixel 1	pixel 2	---	pixel 784
Image #1	51	27	--	126
---	--	--	--	
Image #60000	65	32	--	121

PCA Applications

- PCA is used to a) find patterns in data b) for dimensionality reduction in various areas such as
    - Finance
    - Bioinformatics
    - Psychology
- Some specific applications:
    - Image compression
    - Facial recognition
    - Computer vision

Motivation: Principal Components



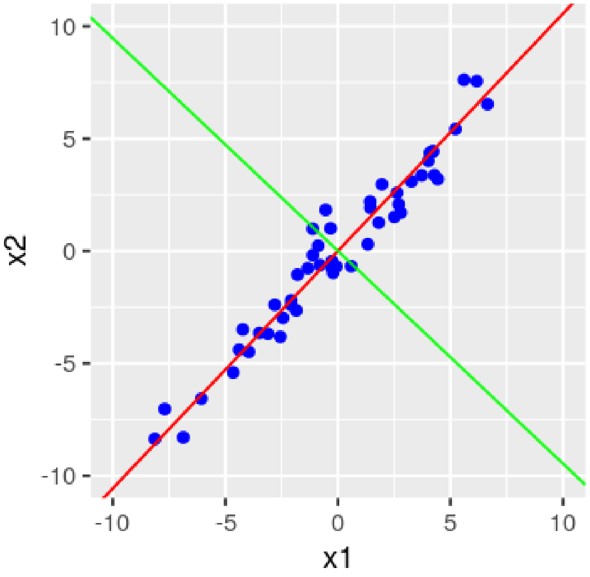
- The plot shows a sample of 50 data points (centered)
  - Each point has two values: X1 and X2 (the features)

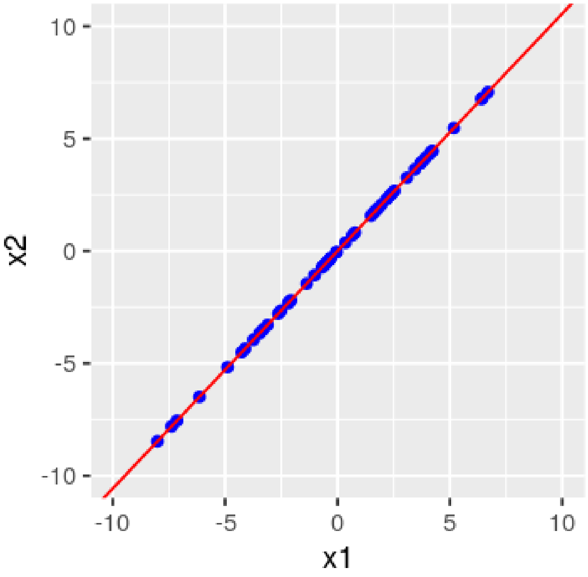
- **First Principal Component:**
  - Data variation is maximum along the red line
  - This direction is the **First Principal Axis**
  - This axis represents a new feature which is a linear combination of original features:
    eg Here the new feature is $X1' = 0.69X1 + 0.72X2$
  - $X1'$ is the **component** of each data point along the principal axis and is called the **First Principal Component**

- **Second Principal Component:**
  - Data variation is next-best along some direction perpendicular to the first principal component
  - This direction is the **Second Principal Axis** (green line)
  - In this example, data variation in this direction is small
  - The feature $X2' = -0.72X1 + 0.69X2$ is the **Second Principal Component**

- For a dataset with n features, there are **n Principal Components**

Motivation: Dimensionality Reduction



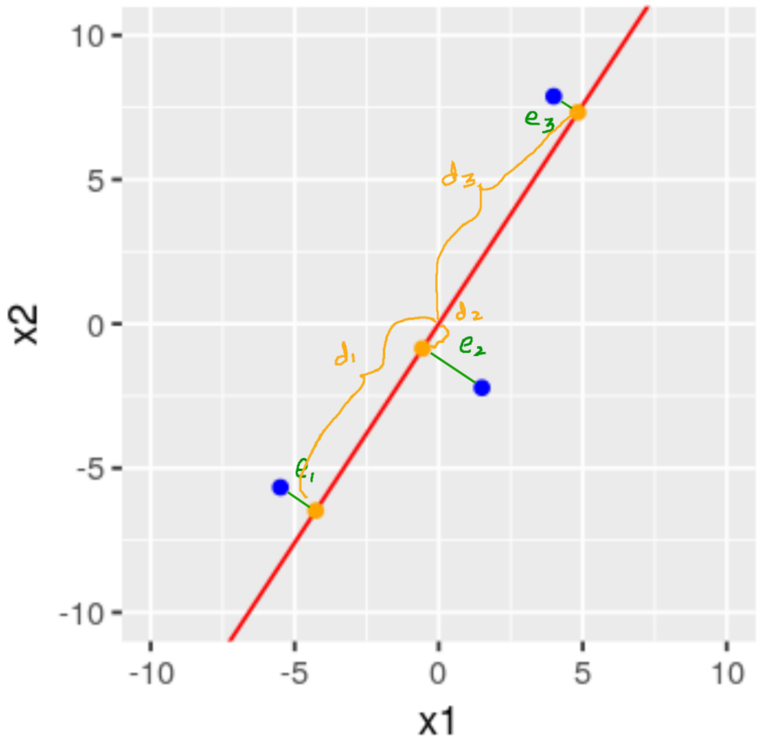


- **Left Plot: Original Data**:
  - Data has maximum variation along the red line
  - Variation of the data along the green line is small
- **Right Plot: 1D approximation**

- **Right Plot: 1D approximation**
    - So the data can be represented by only the First Principal Component
    - The original data can be approximated by projection of the data points along the first principal direction
    - Thus the data is now reduced to one dimension from two
    - Instead of two features $X1, X2$, a single new feature $X1'$ represents the data:
      $X1' = 0.69X1 + 0.72X2$
    - Second principal component can now be neglected
        - That is, the values of the feature $X2'$ are almost zero
          $X2' = -0.72X1 + 0.69X2 \approx 0$

Principal Components from Covariance Matrix

PC: Maximum variance / minimum SSE



- The First Principal Component (along red line)
  - Has maximum variance $s^2 = (d_1^2 + d_2^2 + d_3^2)/2$
  - or Equivalently, minimum Sum of Squared Error $\mathrm{SSE} = e_1^2 + e_2^2 + e_3^2$
- For any other choice of direction, SSE is higher and s^2 is lower than the above values

Pincipal Components From Covariance Matrix
Steps:

- **Step 1:** Center the data matrix X_R to X
- **Step 2:** Find covariance matrix, C
- **Step 3:** Do eigendecomposition of $C, C = VSV^T$
- **Step 4:** Find the Principal Components: These are columns of the matrix XV
- **Step 5:** Do dimensionality reduction, if possible

Centering the Data Matrix:

- **Centering:** Mean of row vectors subtracted from the data matrix

- **Example:**

  - Let the raw data matrix, X_R be
    $$X_R = \begin{bmatrix} 2 & 3 \\ 1 & -2 \\ 3 & 8 \end{bmatrix}$$

  - Mean of all row vectors is :
    $$X = \begin{bmatrix} 2 & 3 \end{bmatrix}$$

  - Then the centered matrix, X is
    $$X = \begin{bmatrix} 0 & 0 \\ -1 & -5 \\ 1 & 5 \end{bmatrix}$$

Covariance Matrix

- Consider a centered dataset X with two features: $X = [X_1, X_2]$, where X_1, X_2 are column vectors
  - Each data sample is in a row
  - Columns are features

- The covariance matrix C is given by:

$$C = \frac{X^T X}{(n-1)}$$

where

$$X^T X = \begin{bmatrix} X_1^T \\ X_2^T \end{bmatrix} * \begin{bmatrix} X_1 & X_2 \end{bmatrix}$$

$$= \begin{bmatrix} X_1^T X_1 & X_1^T X_2 \\ X_2^T X_1 & X_2^T X_2 \end{bmatrix}$$

Example: Covariance Matrix from $\boldsymbol{X}$

- Let the dataset X be
  $$\begin{bmatrix} 1 & 1 \\ 2 & -1 \\ -3 & 0 \end{bmatrix}$$
- Then the covariance matrix is given by
  $$\frac{1}{2}\begin{bmatrix} 14 & -1 \\ -1 & 2 \end{bmatrix}$$

Properties of Covariance Matrix

- C is symmetric
- C is positive semidefinite

Properties of a Symmetric Matrix, A

- A has real eigenvalues
- Eigenvectors of A corresponding to distinct eigenvalues are orthogonal
- An 'orthonormal set' of eigenvectors can always be constructed for A

Spectral Theorem: Eigendecomposition of a Symmetric Matrix A

- A can be decomposed as
  $A = VSV^T$
  - here V: Orthogonal matrix, columns are unit eigenvectors of A
  - S: Diagonal matrix. Diagonals are eigenvalues of A

Properties of an Orthogonal Matrix V

- V preserves length of a vector
- V 'rotates' or 'flips' coordinate axes
- Columns of V form an 'orthonormal set': ie these are unit vectors which are mutually orthogonal

Principal Components from Covariance Matrix:

- Let C be the covariance matrix of centered dataset X
- Then C is positive semidefinite and can be diagonalised as follows:

  $$C = VSV^T$$

  where

  - V: Orthogonal matrix of eigenvectors
    - The eigenvectors are the Principal Directions or Principal Axes of the data
  - S: Diagonal matrix with non-negative eigenvalues λ_i in decreasing order
    - λ_i are the variances of the respective principal components

Principal Components from Covariance Matrix:

- **Consider the Matrix $\boldsymbol{P} = \boldsymbol{XV}$**
    - **jth column of $\boldsymbol{XV}$:** This is the jth Principal Component
    - **ith row of $\boldsymbol{XV}$:** gives the coordinates of the ith data point in the new PC space

- **Variance Explained**
    - For an ith principal component,

      Variance Explained (%) = $\frac{\lambda_i * 100}{\sum_j \lambda_j}$

- **Cumulative Variance Explained**
    - For an ith principal component,

      Cumulative Variance Explained (%) = $\frac{\sum_{k=1}^{i} \lambda_k}{\sum_j \lambda_j} * 100$

- **Loadings or Factor Loadings:**
  - The columns of V give principal directions. But these are unit vectors and so do not indicate data variation along the directions.
  - To address this, a **Loading** matrix is defined as follows:
    $L = V\sqrt{S}$
  - ith column of L gives principal directions **scaled** by standard deviation in that direction.
  - Thus, as against the column vectors of V which are unit vectors, column vectors of L have standard deviations in the corresponding directions as their lengths
  - Thus columns of L give a direct indication of the dominance of a particular principal direction

Example:

- **Given Centered Data Matrix, X:**

$$X = \begin{bmatrix} 1 & 1 \\ 2 & 3 \\ -3 & -4 \end{bmatrix}$$

- **Covariance Matrix**

$$C = \frac{X^T X}{(n-1)}$$

$$= \begin{bmatrix} 7 & 9.5 \\ 9.5 & 13 \end{bmatrix}$$

- **Eigendecomposition of Covariance Matrix**
$$C = VSV^T = \begin{bmatrix} 0.591 & -0.807 \\ 0.807 & 0.591 \end{bmatrix} \begin{bmatrix} 19.96 & 0 \\ 0 & 0.038 \end{bmatrix} \begin{bmatrix} 0.591 & 0.807 \\ -0.807 & 0.591 \end{bmatrix}$$

  - Thus, variances along the 2 principal directions are: $19.96 \text{ and } 0.038$ resp

  - **Variance explained by the First Principal Component** = $\dfrac{19.96 * 100}{19.96 + 0.038} = 99.8\%$ (same as above)

- **The Principal Components are given by:**

$$P = XV = \begin{bmatrix} 1 & 1 \\ 2 & 3 \\ -3 & -4 \end{bmatrix} \begin{bmatrix} 0.591 & -0.807 \\ 0.807 & 0.591 \end{bmatrix} = \begin{bmatrix} 1.4 & -0.2 \\ 3.6 & 0.2 \\ -5 & 0.1 \end{bmatrix}$$

- **Inferences:**

    1. The first principal component (first column of P) is dominant
    2. Second component (second column of P) is very small and can be ignored thus reducing the dataset dimension to one
    3. First singular value $\sigma_1 = 6.319$ dominates
    4. The single dominant feature (first principal component) is then given by first column of V : $X1' = 0.591X1 + 0.807X2$
    5. **Variance Explained** by the First Principal Component: = $\dfrac{6.319^2 * 100}{6.319^2 + 0.274^2} = 99.8\%$, which justifies ignoring second principal component

- **Loadings:**
$$L = V\sqrt{S} = \begin{bmatrix} 0.591 & -0.807 \\ 0.807 & 0.591 \end{bmatrix} \begin{bmatrix} 4.47 & 0 \\ 0 & 0.19 \end{bmatrix} = \begin{bmatrix} 2.64 & -0.16 \\ 3.60 & 0.11 \end{bmatrix}$$
    - Thus, length of first column of L is 4.47, the sd along first principal direction