# Dimension reduction

SCI 2000–Guest Lecture

Max Turgeon

24 November 2020

University of Manitoba

# Motivation

# Dimension reduction

- Most data is *multivariate*.
    - We measure more than one variable on each patient.
    - We measure more than one characteristic on each player of a cricket team.
- By construction, **images** are multivariate data.
    - Otherwise it would be a single pixel...
- **Dimension reduction methods** transform multivariate data to help us find "hidden" structures.
- Linear dimension reduction uses linear transformations
    - Rotation, reflection, rescaling and projection onto lower dimensional space.
    - After transformation, relationships may be easier to see.

# Principal Component Analysis

# Main definition i

- **PCA**: Principal Component Analysis
  - Dimension reduction method
- Let $\mathbb{Y}$ be a $n \times p$ matrix.
  - $n$ observations, each contain $p$ measurements.
- We are looking for a linear transformation $W$ ($p \times k$) such that
  - The columns of $\mathbb{Y}W$ are orthogonal.
  - The sample variance of column $j$ is larger than that of column $j + 1$.

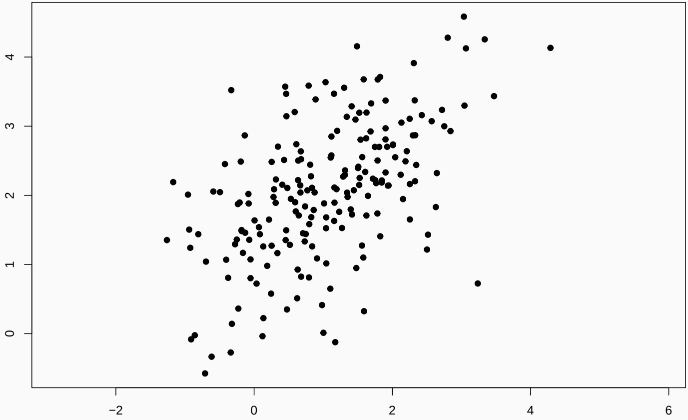
**PCA Theorem**

Let $\lambda_1 \geq \cdots \geq \lambda_p$ be the eigenvalues of the sample covariance matrix $S$, with corresponding unit-norm eigenvectors $w_1, \ldots, w_p$. The PCA solution is given by the matrix

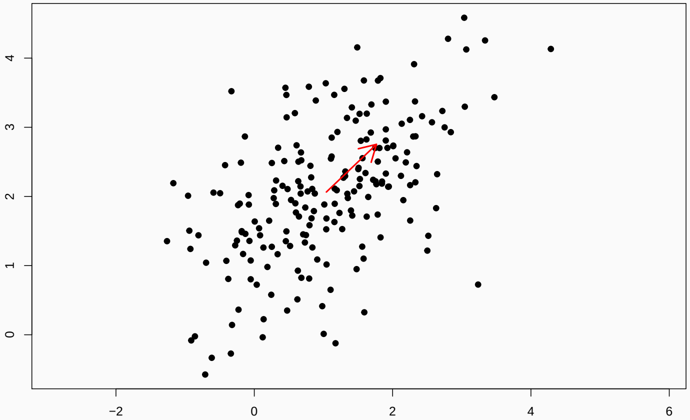
$$W = \begin{pmatrix} w_1 & \cdots & w_k \end{pmatrix},$$

whose $j$-th column is $w_j$.

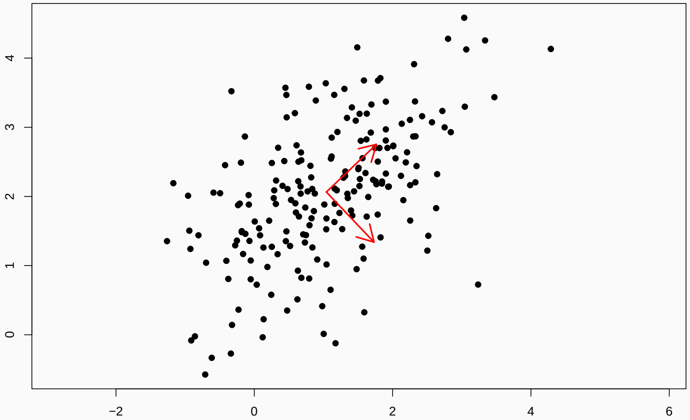
# Example i



# Example ii



# Example iii



# Properties of PCA i

- Some vocabulary:
  - $\mathrm{Z}_j = \mathbb{Y}w_j$ is called the $j$-th **principal component** of $\mathbb{Y}$.
  - $w_j$ is the $j$-th vector of **loadings**.
- Note that we can take $k = p$, in which case we do not reduce the dimension of $\mathbb{Y}$, but we *transform* it into a matrix of the same dimension, but orthogonal columns.
- Each linear transformation $\mathbb{Y}w_j$ contributes $\lambda_j / \sum_k \lambda_k$ as percentage of the overall variance.

# Properties of PCA ii

- **Selecting** $k$: One common strategy is to select a threshold (e.g. $c = 0.9$) such that

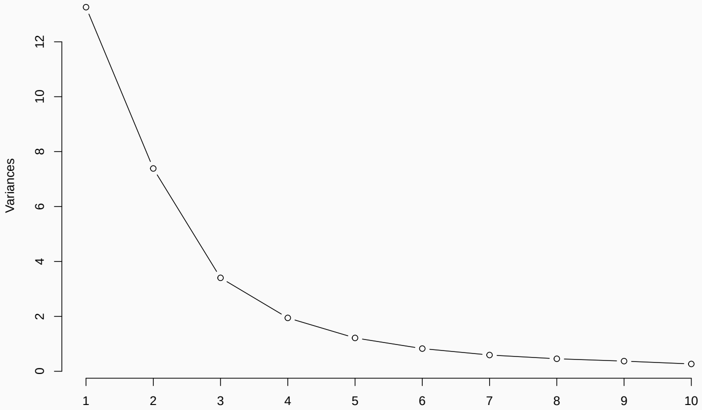
$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \geq c.$$

# Demo

See demo in Jupyter notebook.

# Scree plot

- A **scree plot** is a plot with the sequence $1, \ldots, p$ on the x-axis, and the sequence $\lambda_1, \ldots, \lambda_p$ on the y-axis.
- Another common strategy for selecting $k$ is to choose the point where the curve starts to flatten out.
    - **Note**: This inflection point does not necessarily exist, and it may be hard to identify.


Variances
0
2
4
6
8
10
12
1
2
3
4
5
6
7
8
9
10

# Geometric interpretation of PCA i

- The definition of PCA as a linear combination that maximises variance is due to Hotelling (1933).
- But PCA was actually introduced earlier by Pearson (1901)
  - *On Lines and Planes of Closest Fit to Systems of Points in Space*
- He defined PCA as the **best approximation of the data by a linear manifold**
- Let's suppose we have a lower dimension representation of $\mathbb{Y}$, denoted by a $n \times k$ matrix $\mathbb{Z}$.

# Geometric interpretation of PCA ii

- We want to *reconstruct* $\mathbb{Y}$ using an affine transformation

$$f(z) = \mu + W_k z,$$

  where $W_k$ is a $p \times k$ matrix.
- We want to find $\mu, W_k, \mathsf{Z}_i$ that minimises the **reconstruction error**:

$$\min_{\mu, W_k, \mathsf{Z}_i} \sum_{i=1}^{n} \|\mathsf{Y}_i - \mu - W_k \mathsf{Z}_i\|^2.$$

# Geometric interpretation of PCA iii

**Eckart–Young theorem**
The reconstruction error is minimised by taking $W_k$ to be the matrix whose columns are the first $k$ eigenvectors of the sample covariance matrix $S$.

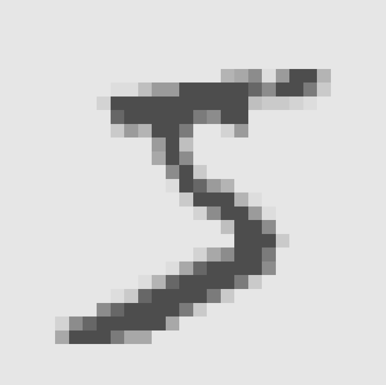
Equivalently, we can take the matrix whose columns are the first $k$ *right singular vectors* of the centered data matrix $\tilde{\mathbb{Y}}$.

# MNIST dataset

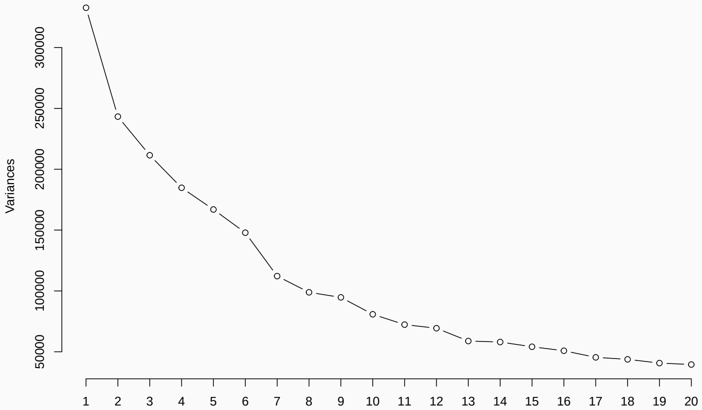
# Description of data

- The MNIST dataset contains images of hand-written digits (0 to 9)
    - It can be downloaded from `http://yann.lecun.com/exdb/mnist/`
- Each image is 784 gray-scale pixels ($28 \times 28$).
- The data is separated into a *training* and *testing* dataset.
    - 60,000 training samples
    - 10,000 testing samples

# Data Visualization i

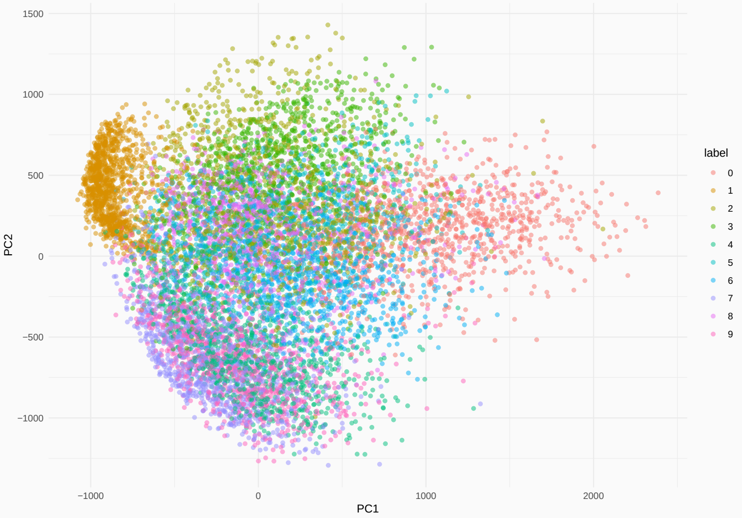


# Data Visualization ii

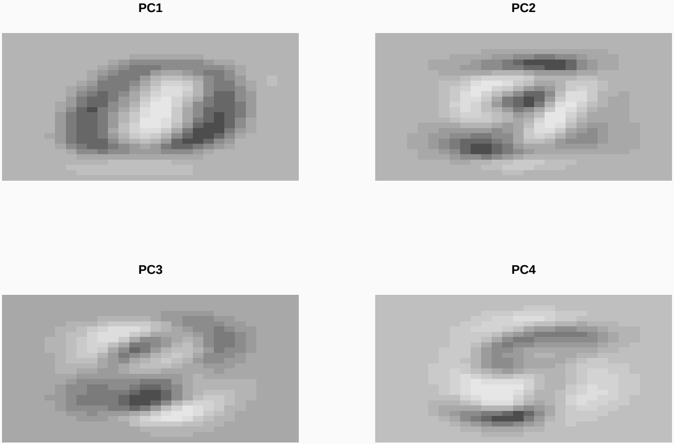


# Data Visualization iii

# Data Visualization iv

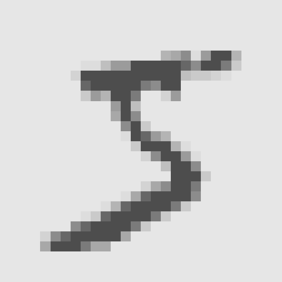


# Data Visualization v



# Data Visualization vi

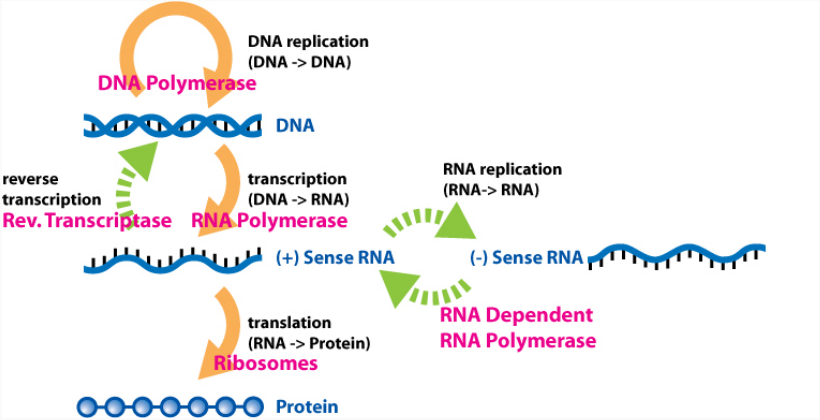
**Original**



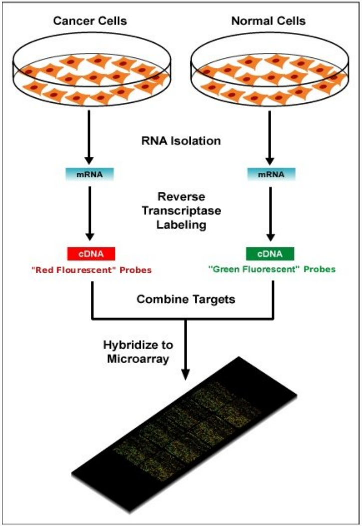
**Approx**

# Application: Gene expression analysis

# Central Dogma of Molecular Biology



# Microarray Analysis

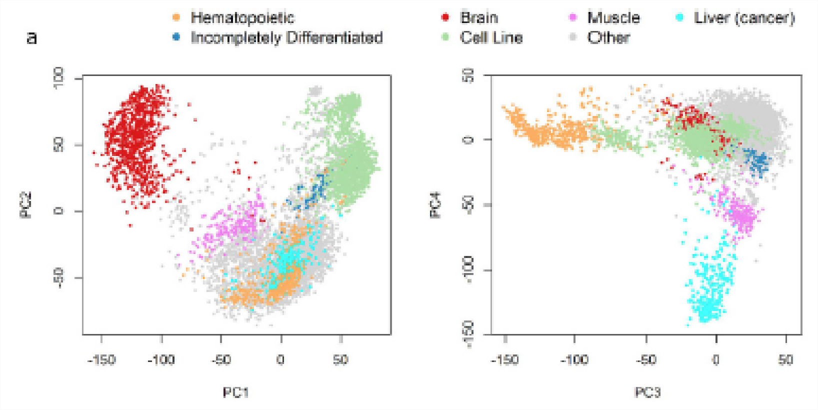


# PCA and Gene expression analysis i

- The following figures come from Lenz *et al.* (2016) in Nature–Scientific Reports.
    - *Principal components analysis and the reported low intrinsic dimensionality of gene expression microarray data*
- They collected data on transcript abundance for over 45,000 transcripts. The data was collected on 5372 samples, corresponding to 369 different tissues, cell lines, or disease states.
- **Main idea**: proteins work together, so transcript should be highly correlated.

# PCA and Gene expression analysis ii

- The authors argue that only the first 3-4 principal components of this gigantic dataset are biologically relevant!



# Summary

- Dimension reduction transforms the data to make hidden structures more visible.
  - Different methods highlight different features.
- PCA can be used for visualization of image data.
  - Transform through PCA and look at the first few dimensions.
- PCA can be used with all types of highly structured data.
  - Gene expression can be measured using high-resolution images.
  - These images are turned into abundance values.
  - PCA is then used to transform these abundance values.