# Part 1: overview

- Principal Component Analysis (PCA)
  - Maximal variance principle
  - Single dataset
- Partial Least Squares (PLS)
- Two-way Orthogonal PLS (O2PLS)
- Post-hoc analyses using external databases

# Introduction

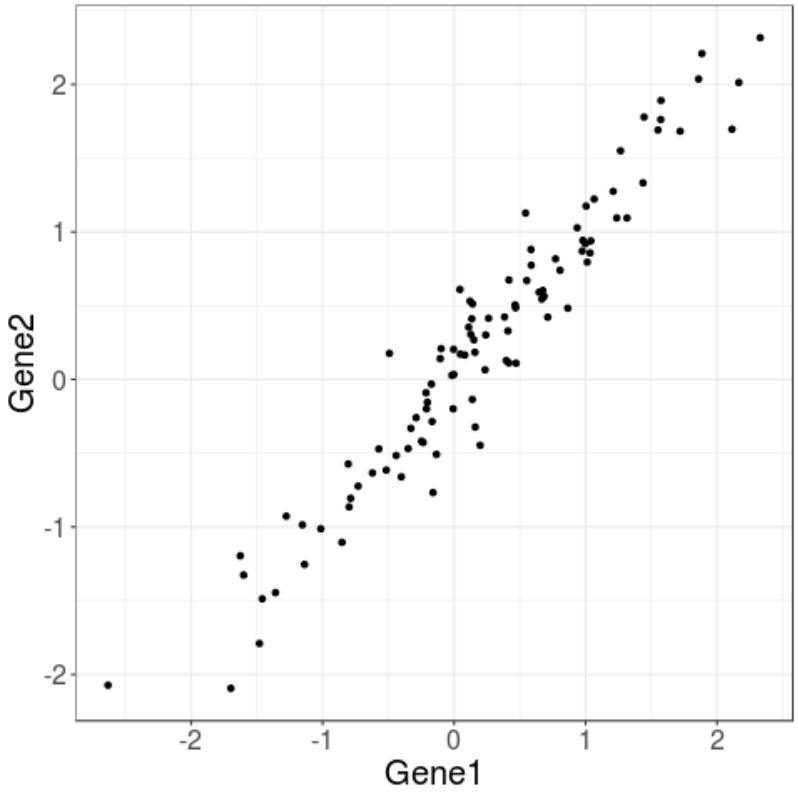
- Suppose we have a dataset $X$, with $N$ rows and $p$ columns
- E.g. methylation values, glycan abundances, etc

**How to inspect these variables and their correlations?**

| | Gene 1 | Gene 2 |
|---|---|---|
| **Sample 1** | $x_{1,1}$ | $x_{1,2}$ |
| **Sample 2** | $x_{2,1}$ | $x_{2,2}$ |
| **Sample 3** | $x_{3,1}$ | $x_{3,2}$ |

# Example: bivariate data

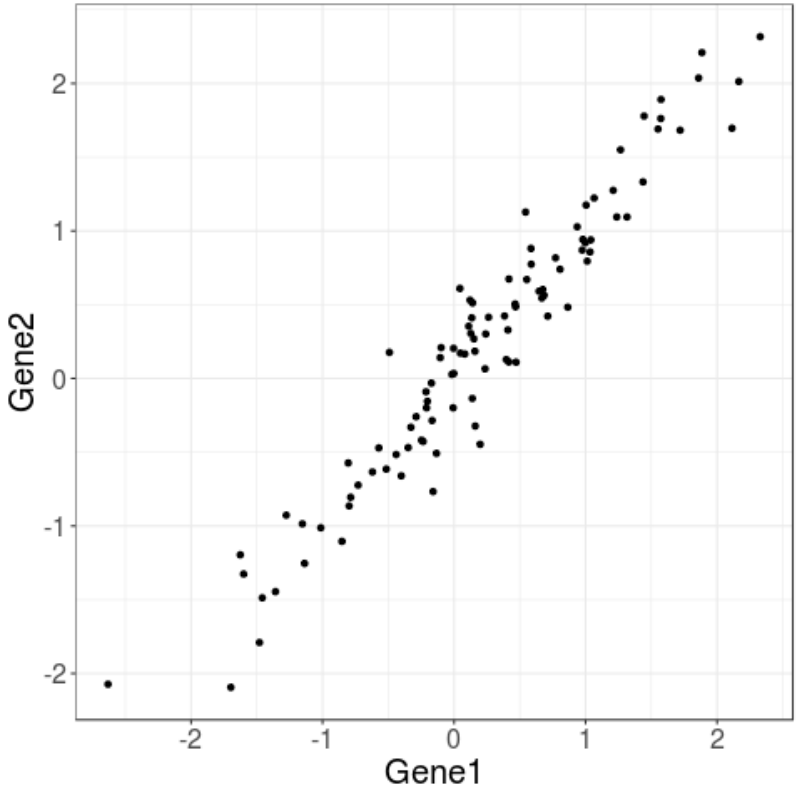
- Suppose we have two genes
- They are highly correlated




- How would you represent these data? Do you need both dimensions?
- In general, aggregate data across direction of maximal variance

# PCA principle: maximal variance

- $X$ is our dataset, where each column has zero mean
- Consider a *linear combination* of $X$




**Which one has the highest variance?**

- 1*Gene 1 + 1*Gene 2?
- 1*Gene 1 – 1*Gene 2?

# PCA principle: maximal variance

- $X$ is our dataset, where each column has zero mean
- A *linear combination* of $X$ can be written as the product of $X$ and a *weight vector* $w$
- The vector $w$ has p numbers
- Resulting product: $Xw$ (dimension of this product?)
- The *sample variance* of $Xw$ is the "matrix-squared"
- $\frac{1}{N}(Xw)^\top(Xw) = \frac{1}{N}w^\top X^\top Xw$

**Message:** for each $w$, we can calculate the variance

# PCA algorithm: estimation

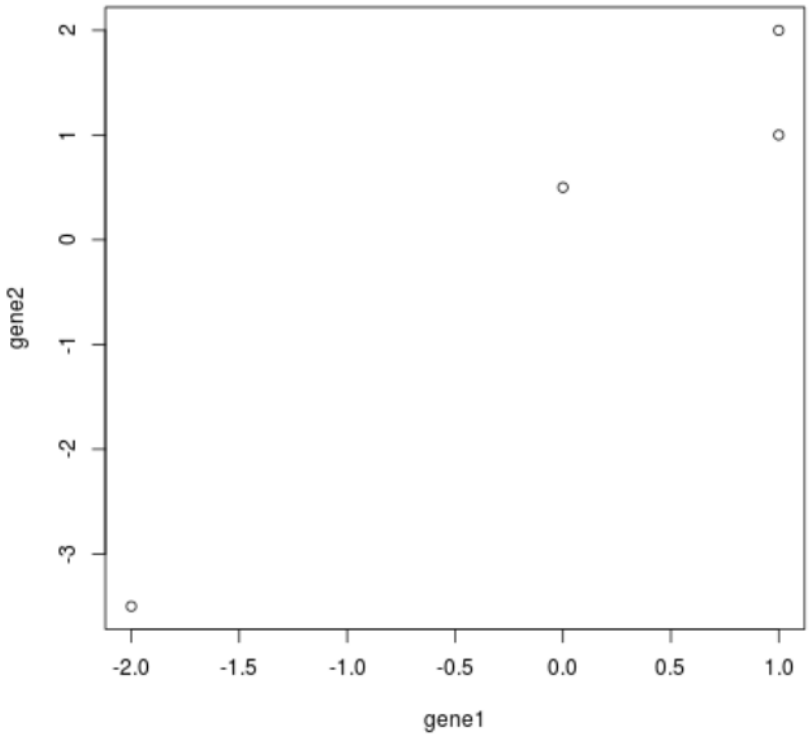
- Objective: maximize variance of linear combination
- The combination is relative
  - Multiply numbers by 0.1, or 1, or 100 times is equivalent
- Can be calculated using
  - A singular value decomposition (*svd* in R)

**Message**: first principal component is first (right) singular vector

# An example in R

```
gene1 <- c(-2, 0, 1, 1)
gene2 <- c(-3.5, 0.5, 2, 1)
plot(gene1, gene2)
svd(cbind(gene1, gene2))$v
```

```
          [,1]       [,2]
[1,] 0.5007639  0.8655839
[2,] 0.8655839 -0.5007639
```




- The best weight is (0.5,0.87)
- So: 0.5 times gene 1 and 0.87 times gene 2 gives the highest variance

# PCA: interpretation of the results

- Weights $w$ are numbers for each feature (gene) indicating the importance for that principal component
- A weight is also called a loading
- These loadings are relative, the squares sum up to 1
- The result of projecting the data onto these loadings are called scores: $t = Xw$
- These scores $t$ indicate the importance of each sample for that component

# Principal component analysis: summary

- For a given dataset $X$, we want to inspect variables and their correlations
- Based on a bivariate scatterplot, we find the direction of maximal variance
- This direction is represented by weights for each feature, calculated by a singular value decomposition
- The projections of the data onto these weights are called the scores.
- One can interpret or plot the weights and scores to understand which features/samples are most important.

# Down syndrome data

- Methylation and glycomics data on Down syndrome patients and controls
- 29 trios: a DS patient and its sibling and mother
- A subset of 450k methylation intensity values
  - We only consider chromosome 21 and targeting a gene
  - Yields 3322 CpG sites
- Ten (plasma N-)glycan peaks
  - Divide by the total abundance, apply log

# Exercises

1. Load the Down Syndrome data and describe the datasets.
   - Use boxplots, histograms and frequency tables
2. Perform PCA on glycomics, using *svd*, and inspect the loadings of the first component
3. Calculate and plot the scores of the first component

# Part 1: overview

- Principal Component Analysis (PCA)
  - Maximal variance principle
  - Single dataset
- Partial Least Squares (PLS)
- Two-way Orthogonal PLS (O2PLS)
- Post-hoc analyses using external databases