Morning session: overview

Unsupervised learning: dimension reduction and data integration

- Principal Component Analysis (PCA)
  - Maximal variance principle
- Partial Least Squares (PLS)
  - Maximal covariance principle
- Two-way Orthogonal PLS (O2PLS)
  - Multi-omics data integration

Introduction

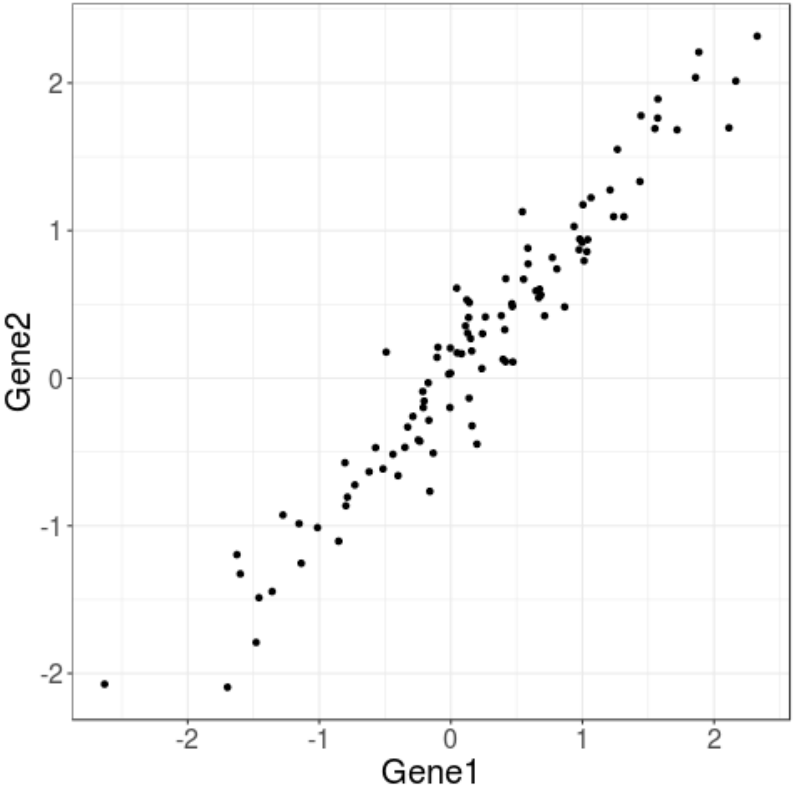
- Suppose we have a dataset X, with N rows and p columns
- Make it concrete: gene expression values, metabolite abundances, questionnaires, etc

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
- 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

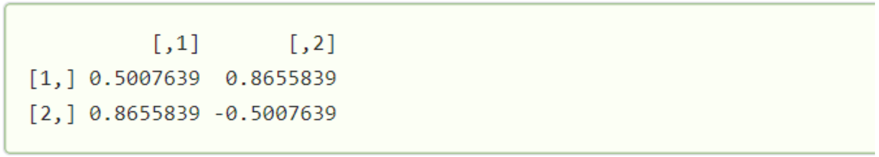
- Objective: maximize $w^\top X^\top X w$ over w
- Such that the squared elements in w, written as $w^\top w = 1$ (why?)
- Apply the Lagrange method:
  - Maximize $w^\top X^\top X w \ - \lambda w^\top w$
  - Differentiate w.r.t. w and set to zero: $X^\top X w = \lambda w$
- Resulting w is the first eigenvalue of $X^\top X$
- Equivalently: first right singular vector of X

Thus: first principal component is first eigenvector

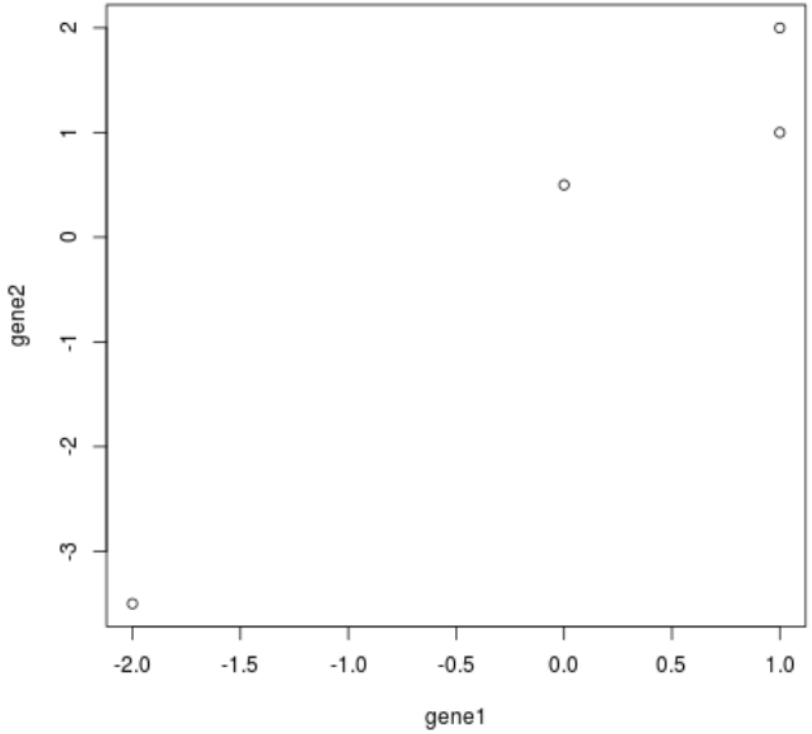
An example in R

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

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



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



PCA: interpretation of the results

- Weights w are numbers for each feature (gene) indicating the importance for that principal component
- These weights are relative, the squares sum up to 1
- The result of projecting the data onto these weights 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

Morning session: overview

Unsupervised learning: dimension reduction and data integration

- Principal Component Analysis (PCA)
  - Maximal variance principle
- Partial Least Squares (PLS)
  - Maximal covariance principle
- Two-way Orthogonal PLS (O2PLS)
  - Multi-omics data integration