# Connecting Nutrition, Health and Environment

Perla Troncoso Rey, Wiktor Jurkowski, Earlham Institute

26 - 27 January 2017

## Contents

## List of Figures

## List of Tables

## Part I

# Part I

## 1 Introduction to the data

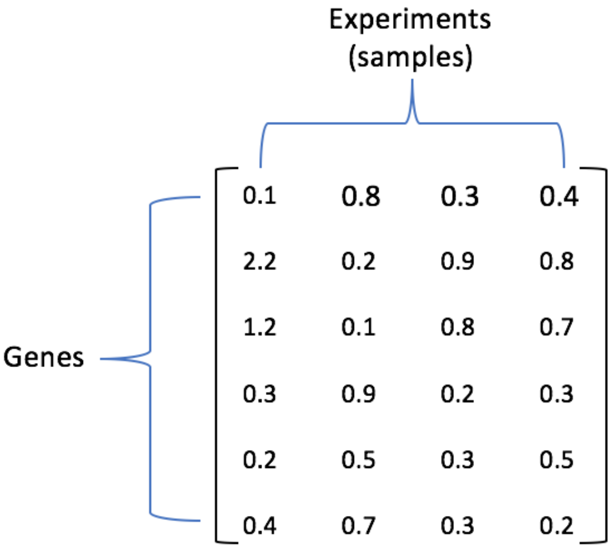
Some details

### 1.1 Exploring the expression data

#### 1.1.1 Principal Component Analysis

Principal Component Analysis, PCA, is a mathematical technique that is used to explore data, specially high-dimensional data. PCA is a method of extracting some of the most important trends in high-dimensional data.

Example of data is shown in Figure 1



| | | | |
|---|---|---|---|
| 0.1 | 0.8 | 0.3 | 0.4 |
| 2.2 | 0.2 | 0.9 | 0.8 |
| 1.2 | 0.1 | 0.8 | 0.7 |
| 0.3 | 0.9 | 0.2 | 0.3 |
| 0.2 | 0.5 | 0.3 | 0.5 |
| 0.4 | 0.7 | 0.3 | 0.2 |

Figure 1: Example of expression data with samples across the columns and individual genes down the rows.

high dimensionality: as many dimensions as number of genes The result of each experiment can be thought as a kind of space, where each each feature is a coordinate in the space. there are typically thousands of genes the structure of pattern in the data extends to all the dimensions

Tools to perform PCA:

- MATLAB

- R

How it works:
The mean represents the average of the values in the data:

$$\bar{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

The variance provides the the spread of the data:

$$\text{Var}(\mathbf{X}) = \sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{\mathbf{X}})^2 \tag{2}$$

For example, 3 show two distribution with the same mean but different variance.

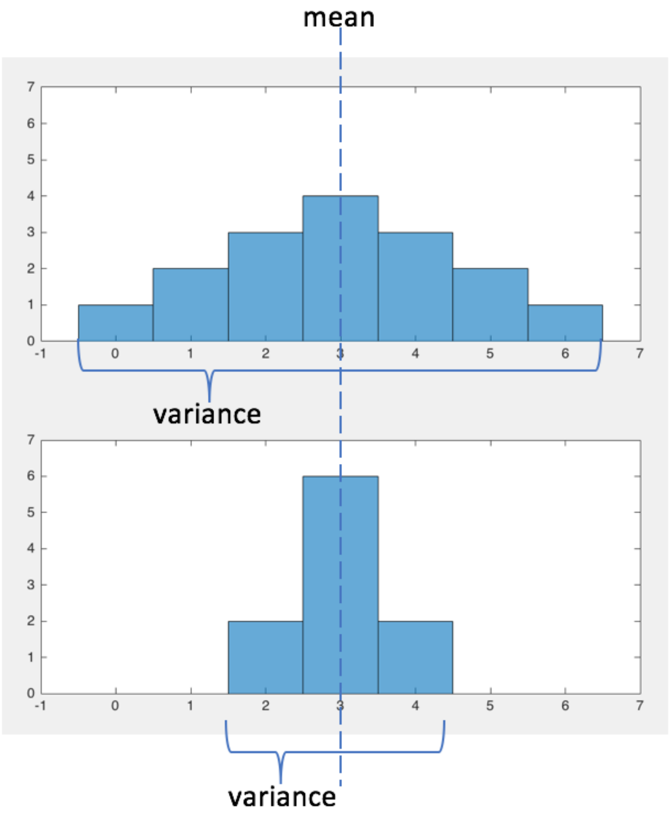


Figure 2: Two distributions with the same mean but different variance.

(i.e. the data points are at the same location but with a different strength. So the third statistic we'll need is the covariance)

The covariance represents the degree of co-dependence of two variables, i.e., it measures the co-dependency of two variables, given by:

$$\mathrm{Cov}(\mathbf{X}, \mathbf{Y}) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{\mathbf{X}})(y_i - \bar{\mathbf{Y}}) \tag{3}$$

Increases with increasing co-dependency and variance. Just as the variance measures the degree to which a set of data varies, the co-variance is a measure of the way two sets of data vary together.

$$\mathrm{Cov}(\mathbf{X}, \mathbf{X}) = \mathrm{Var}(\mathbf{X}) \tag{4}$$

The covariance also increases in magnitude as the variance of each of the two datasets increases.

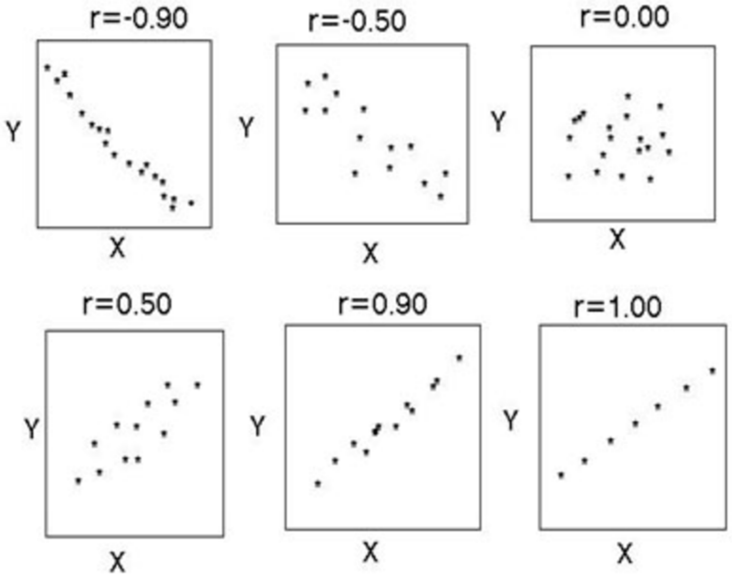


Figure 3: Examples of correlation

Coordinate transformations

- Two dimensional space described by coordinates (i.e. a point in space described by $x_1$ and $y_1$ such that $\mathbf{v} = [x_1, y_1]$).

xxxxxxxxxxxxxxx

PCA benefits (merits)

- A powerful tool to visualise high dimensional data
- Shows quantified difference among observations
- Used to assess data quality and discover relationships between data points

Example 1:

PCA Plot of gene expression data

Each dot is a gene expression from a sample in each category (class) from a patient, and is coloured by its sub-type

The three axis are the three principal components and the numbers represent the percentage of variance that is captured by each component

The first component captures the most variance, whereas the second and third capture only a small percentage

Dots of the same subtype tend to cluster together which means that samples of the same subtype have similar transcritption profiles.

The distance on the dots on each axis should not be treated equally (as each component captures a different percentage of variance). difference on the first component should be taken into more consideration.

Example 2: Random data

Simulated gene expression data by random numbers

This is how a random dataset would look like in a PCA plot:

- dots of different classes mix all together

- the first three components capture almost equal and small variance

- from the plot once would conclude that the different subtypes are not distinct from each other or that subtype has no influence on tumor cell transcriptome

DATA PREPARATION

- Use microarray gene expression data as an example

- Gene expression data is usually stored in a tab delimited text file. The extension of such files could be .csv, .soft, .xls(x), etc. Use Excell of Sublimetext to open and preview the file.

- Gene expression values must be normalised before PCA plotting.

PCA

- We need to transpose the matrix because the function requires the rows of the input matrix to be observations and the columns variable, which means rows to be the gene expression profiles and columns to be the genes.

- There are three outputs to the function:

1. The first output is the coefficient matrix (not used here!)

2. The second output is the scores, which are the transformed coordinates by PCA.

3. The third output, pca variance, stores how much variance each component captures.

Looking into detail on the outputs:

- The first several components capture most variance of the data.

- The score matrix has the same arrangement as expression matrix, which are rows as gene expression profiles and columns are genes. We pick the first three columns, namely the first three components. The first component will be the x-axis, the second to be the y-axis and the third component to be the z-axis.

Running PCA

PCA Plot using MultiPEN

Parameters:

gene expression data

groups

#### 1.1.2 Hierarchical Clustering

### 1.2 Interaction Network

#### 1.2.1 Compile the network with Cytoscape

# Part II

# Part II

## 2 Ranking genes (Feature Selection)

Feature selection with MultiPEN from expression leves and network

## 3 Pathway Analysis

Objective

Steps

Input data A list of ranked genes. For this session we will use the rankings from feature selection in ExampleOutputsMultiPENRankings_lambda0.0001.txt.

To run the R script type in the terminal:

Rscript enrichmentGO.R ../ExampleOutputs/MultiPENRankings_lambda0.0001.txt output/