Processing of Biological Data

Prof. Dr. Volkhard Helms
Winter Semester 2018/2019

Saarland University
Chair of Computational Biology

Exercise Sheet 1

Due: Nov 15, 2018 14:15

Submit your solutions on paper, hand-written or printed at the *beginning* of the lecture or in building E21, Room 3.03. Alternatively you may send an email with a single PDF attachment. If possible, please include source code listings. Additionally hand in all source code via mail to duy.nguyen@bioinformatik.uni-saarland.de.

Principal Component Analysis (PCA) and Data Imputation

Exercise 1.1: Principal Component Analysis (40 points)

(a) Describe step-by-step the PCA technique on matrix $\mathbf{X}$ (with $\mathbf{I}$ observations and $\mathbf{J}$ variables), from the pre-process until the transformed matrix is obtained (15 points).

(b) In this section, we will apply the PCA technique to a toy dataset ("pca_toy.txt" in the supplement, with 50 observations and 4 variables: a, b, c and d) by using the `prcomp` function in `stats` R package. The documentation of `prcomp` can be found in:
https://www.rdocumentation.org/packages/stats/versions/3.5.1/topics/prcomp

- Plot the transformed data w.r.t. the first two principal components (PC1 and PC2) (10 points)
- Which variables are the most important in PC1 and PC2? Why? (10 points)

(c) Why do we need to standardize the data before doing PCA? (5 points)

Exercise 1.2: Data Imputation

Imputation based on a given data distribution (40 points)

The main idea behind this method is to impute missing proteomics data which have expression below the detection limit (http://www.nature.com/nmeth/journal/v13/n9/fig_tab/nmeth.3901_SF3.html). Basically, the imputation can be broken down into the following steps:

1. Calculate the mean and standard deviation of the current data.
2. Derive the new mean and standard deviation for the missing data based on the current distribution. The new mean should be in the lower quantile of the distribution since we want to simulate the low expression data. The new standard deviation could be derived by taking a fraction of the current standard deviation.
3. Generate the new data based on the new mean and standard deviation from the previous step.

(a) The "ms_toy.txt" file contains an example of proteomics data (6960 proteins × 6 samples). Use any of the 6 samples and write a script to impute the missing data (which are indicated by "NA") for the sample of your choice by following the steps mentioned above. (20 points)

Hint: in R, use function `qnorm` *to derive the new mean and function* `rnorm` *to generate the data.*

(b) Plot the distribution of the sample with the imputed data in a similar fashion as Fig. 1 with different combinations of new means and standard deviations. What is the effect of different means and standard deviations? Which configuration (mean & standard deviation) is the most logical/desirable? (15 points)

Hint: in R, use function `hist` *and function* `plot`

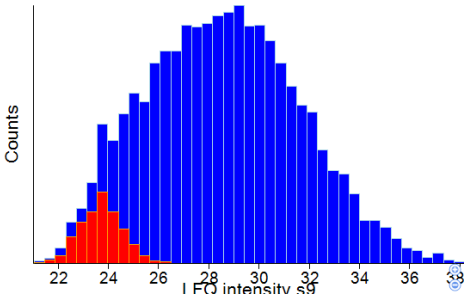

Figure 1: Sample Plot (blue is the overall distribution, red is the imputed data)

(c) What is the shortcoming of this method? (5 points)

Local Least squares imputation (20 points)

Write an R script to perform Local Least squares (LLS) imputation on control samples (`ctrl.1`, `ctrl.2` and `ctrl.3`) of the proteomics data mentioned above ("ms_toy.txt") using `llsImpute` function in `pcaMethods` R package. More information about `llsImpute` can be found in the following link: https://www.rdocumentation.org/packages/pcaMethods/versions/1.64.0/topics/llsImpute. Write the new control data into a text file.

Hint: LLS imputation is not possible if all samples have missing data. Therefore, those cases should be removed before imputation.

Have fun!