SPRING SEMESTER 2017
BIS 692B, CBB 645B, STAT 645B
Statistical Methods in Genetics and Bioinformatics

Homework #2
Due 2/2/2017 (submit homework in class, printed or handwritten, with source code attached)

(1) Testing dependency by correlation of distance

We introduced distance correlation in the class. Let $(\mathbf{X},\mathbf{Y}) = \{(X_k, Y_k): k=1,\ldots,n\}$ denote a set of observations, where X is a p-dimensional vector and Y is a q-dimensional vector. We defined distance correlation through $V_n^2(\boldsymbol{X},\boldsymbol{Y})$, $V_n^2(\boldsymbol{X})$, and $V_n^2(\boldsymbol{Y})$. Let

$$a_{kl} = |X_k - X_l|_p, \bar{a}_{k.} = \frac{1}{n}\sum_{l=1}^{n} a_{kl}, \bar{a}_{.l} = \frac{1}{n}\sum_{k=1}^{n} a_{kl}, \bar{a}_{..} = \frac{1}{n^2}\sum_{k,l=1}^{n} a_{kl},$$
$$A_{kl} = a_{kl} - \bar{a}_{k.} - \bar{a}_{.l} + \bar{a}_{..}.$$

Similarly, let

$$b_{kl} = |Y_k - Y_l|_q, \bar{b}_{k.} = \frac{1}{n}\sum_{l=1}^{n} b_{kl}, \bar{b}_{.l} = \frac{1}{n}\sum_{k=1}^{n} b_{kl}, \bar{b}_{..} = \frac{1}{n^2}\sum_{k,l=1}^{n} b_{kl},$$
$$B_{kl} = b_{kl} - \bar{b}_{k.} - \bar{b}_{.l} + \bar{b}_{..}.$$

In these notations, $|X_k - X_l|_p$ and $|Y_k - Y_l|_q$ represent distance measure in the p-dimensional and q-dimensional space, respectively. Define

$$V_n^2(\boldsymbol{X},\boldsymbol{Y}) = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}B_{kl}.$$

Prove that

$$V_n^2(\boldsymbol{X},\boldsymbol{Y}) = S_1 + S_2 - 2S_3$$

where

$$S_1 = \frac{1}{n^2}\sum_{k,l=1}^{n} |X_k - X_l|_p |Y_k - Y_l|_q,$$
$$S_2 = \frac{1}{n^2}\sum_{k,l=1}^{n} |X_k - X_l|_p \frac{1}{n^2}\sum_{k,l=1}^{n} |Y_k - Y_l|_q,$$
$$S_3 = \frac{1}{n^3}\sum_{k=1}^{n}\sum_{l,m=1}^{n} |X_k - X_l|_p |Y_k - Y_m|_q.$$

(2) Correlation measures and differential expression

Please download the data from the class website.

This is a breast cancer data set from The Cancer Genome Atlas (TCGA). Each column represents the gene expression values for one sample. The data contain matched tumor

and normal samples: sample 1 and sample 61 are the tumor and normal samples from patient 1, sample 2 and sample 62 are the tumor and normal samples from patient 2… In total there are 120 matched tumor and normal samples.

[2.A.] **Correlation analysis**.
In the class, we talked about several correlation measures. Consider Pearson Correlation, Spearman Correlation, (**and others if you have time, bonus**). Do you see any difference when using different correlation measures? If so, find out what causes the difference.

[2.B.] **Differential expression**.
Identify differentially expressed genes between tumor and normal. You may use any criterion you like, for example, certain cut-off for *p*-values, some procedures for multiple hypothesis testing or the locfdr procedure. Explain and justify your choice.
Note: differentially expressed genes are supposed to have different mean expression values between normal and tumor samples.