Homework 2

4 月 4 日前提交至 elearning

请提交R 或 python 的完整代码及产生的表和图.

1. 实现 “Adaptive Thresholding for Sparse Covariance Matrix Estimation” 中的 Table 1（只要复现 hard thresholding 部分，即表格的右半部分）。注意：复现的表中的数字和论文中不会完全相同（seed 不同）。

2. 实现 “Adaptive Thresholding for Sparse Covariance Matrix Estimation” 中的 Figure 3（只需要复现图 (a) 和 (c)）。注意：复现结果会与文中不同，每位同学得到的也会不同，请在 heatmap 下方标注你得到的 zeros 的比例。

   data 获取方式：

   library(plsgenomics)

   data(SRBCT)

   提示：根据 dim(SRBCT$X)，共有 83 个 samples，前 63 个为 paper 中所指的 training sample。其他更多内容请查看 R help 文档及 paper 的 Section 5.2。

   另注：论文中 Section 5.2 有个小笔误，请大家看下图红色圈出的符号，它代表的是样本标准差，它的平方才是样本方差。请大家 coding 的时候注意一下。

$$F = \frac{1}{k-1}\sum_{m=1}^{k} n_m(\bar{x}_m - \bar{x})^2 \Big/ \left(\frac{1}{n-k}\sum_{m=1}^{k}(n_m-1)\hat{\sigma}_m^2\right),$$

where $n = 63$ is the sample size, $k = 4$ is the number of classes, n_m, $1 \le m \le 4$ are the sample sizes of the four types of tumors, $\bar{x}_m$ and $\hat{\sigma}_m$ are the sample mean and sample variance of the class m, and $\bar{x}$ is the overall sample mean. Based on the F values, we chose the top 40 and bottom 160 genes. We also ordered

注意事项：

1. 提交的代码应为.R 或.Rmd 或.py 或.ipynb 格式文件（请对代码进行适当的注释）。提交其他格式文件的酌情扣分。

2. 提交的代码应可以直接运行得到结果。如果代码内容有所缺失，则缺失的部分一律按照完全错误处理。

3. 为公平起见，若无特殊原因，截止时间后提交的作业满分为 20 分 (满分为 30 分)。

4. 若发现抄袭，抄袭和被抄袭的作业均按零分处理。