18.465 PROBLEM SET 7 DUE FRIDAY, APRIL 10, 2015

This PS will involve testing multiple hypotheses and the "pvclust" library in R and the program "pvclust" in it for hierarchical clustering (of columns of a data matrix, unlike hclust methods which cluster rows), which gives evaluation of bootstrap support for each node. The library also contains a data matrix "lung" which will be studied. So begin an R session with library(pvclust).

1. (10 pts.) First, we'll consider a toy example. The code "toymatf" is on the course website and described in a handout. Take a randomly generated data set x = runif(15). Create a data matrix for it by mat = toymatf(x). Run pvclust on it as in the pvclust documentation:

```
mat.pv = pvclust(mat,nboot = 100)
plot(mat.pv, cex = 0.8, cex.pv = 0.7)
```

and print the plot with dev.print() . The observed bootstrap support for each node is given as a percentage in green on the upper right of the node. (In this problem, ignore the "au" numbers given in red on the upper left of the nodes.)

2. (30 points) The "lung" data set, a 916× 73 table, is in the pvclust library. One can load it by the command data(lung). Represent it as a matrix, say lungmat = as.matrix(lung). Consider some 300 rows of that data set, different for each student, as follows:

lungmat[1:300,]	D. Hunter
lungmat[301:600,]	name Y. Mao
lungmat[601:900,]	name A. Yu

The data set "lung" is to some extent described in the pvclust documentation. It comes from research described in the paper by Garber et al., 2001, which is being distributed. The data set gives gene expression levels originally for 918 cDNA clones, said in the paper to represent 835 unique genes. In the pvclust documentation, it says that some observations of "duplicate genes" have been removed, but actually it seems only in two cases, leaving 916 cDNA clones, each of which gives a row of the data matrix. There are 73 columns.

Date: 18.465, April 3, 2015 .

(a) Do clustering with pvclust of your data set. Make sure that your output plot shows the degree of bootstrap support for each node in the dendrogram (the main point of the procedure). The pvclust software can produce an output in color where there is some information conveyed by the color. Please produce such output if it's feasible for you.
(b) Let's see how well the bootstrap works in this case. Let's consider the dendrogram that Garber et al. found as shown in their Fig. 1, based on the full data of 916 genes (with a couple of duplications making 918) as showing the "true" clustering. Consider each cluster such that the node (edge) where it's assembled has bp of 90 or more in your dendrogram. Are such clusters true ones? It would not be surprising if the cluster is of two tissue samples from the same tumor (_p and _c) or of a tumor and a metastatasis of it, as to a _node. It will be more interesting if it's samples from different patients.
(c) Do the same as in (b) but for the red "au" numbers rather than the green "bp" ones. Do these bootstrap support estimates work as well as "bp" does?

(The last part, (d), appears at the end of the problem.)

Background information: The paper says tissues were taken from "67 human lung tumors representing 56 patients." In the column headings one can see 5 with "_normal" which were sampled from non-cancerous lung tissue in the patients having tumors elsewhere in their lungs (as the paper says, "5 normal lung specimens were studied"). The first column is from non-cancerous "fetal lung tissue for comparison."

The paper says there were 41 adenocarcinomas (the particular focus of the paper), whereas I at first saw 39 in the column headings; it says there were five each large cell and small cell lung cancers (LCLC and SCLC, respectively), but I at first saw 4 of each; and 16 squamous cell cancers (SCCs), but I at first saw 13. Most of these discrepancies can be reconciled as follows. The paper says "Eleven of the tumors were sampled twice, either as"
[i] "a primary tumor/metastatic lymph node pair,"
[ii] "a primary tumor/intrapulmonary metastasis pair,"
[iii] "a pair of metastases from the same patient...,"
[iv] "or a central/peripheral biopsy pair from the same tumor."

In category [i] were 6 patients, who beside a patient code with their primary tumor type had in addition with the same patient code "_node" instead of the cancer type.

In category [iv] were three patients, each of whom also appeared twice, with at the end of their codes "_p" and also "_c".

In category [ii] were two patients. One had two codes 313-99MT_Adeno and 313-99PT_Adeno. The other presented a more complicated situation and was in addition the only patient in category [iii]. Namely, patient 319_00 appeared three times with following codes MT1_Adeno and MT2_Adeno (which clustered together) and PT_Adeno (which clustered far away).

Thus, eleven patients contributed a total of 23 tissue samples (two each for 10, and three for one) of the 67. There are 44 other tissue samples, each coming from a unique individual. It seems to me therefore that there were in fact 55, not 56 total individuals in the study sample, 54 patients and one fetus.

Tissues from the same patient are given in consecutive columns of the data set, for example in columns 2 and 3. Assigning each "node" column to the same type as the primary tumor in the preceding column, I found the numbers of types of cancers then agreed except for LCLC. Looking at the Fig. 1 dendrogram, I see exactly 4 LCLC's, and verified that none had an additional sample from the same patient (none was involved in any of the pairs or triples [i], [ii], [iii], [iv]). The "combined" case reportedly had both LCLC and SCSC so this may give the fifth LCLC.

A small fraction of the data numbers are missing, "NA," as I see twice in the first two rows. The pvclust documentation suggests that it computes correlations only using coordinates that are available in both columns.

The listing of tissue samples across Fig. 1 of the paper is unfortunately not in the same order as the columns of the data set. I'll write here what I found for column numbers of different types. Column numbers in parentheses mean another sample from the same patient as in the preceding column. Such duplicates generally do cluster together as shown by the pairing bars in Fig. 1 of the paper; the left arrow shows a third sample from one patient clustering relatively far from the other two (right arrow).

Fetal: 1
Squamous cell: 2, (3), 14, 19, 22, 23, 25, (26), 27, 32, (33), 53, 59, 62, (63), 71; a total of
16 columns as the paper says.
Adenocarcinoma: 4,5,6,7,8,(9),(10),11,12,13,15,(16),17,18,21,24,28,29,(30),31,34,35, 36,(37), 39,40,41,45,46,50, 54,55,(56),58,61,64,66,(67),69,70,72;
a total of 41 columns as the paper says;
Combined: 20
LCLC: 38,44,51,57

SCLC: 42,47,(48),52,73
normal: 43,49,60,65,68

The clustering shown in Fig. 1 has some interesting features. The normal tissues do cluster together with each other and the fetal tissue, but as if they were a subcluster of the adeno larger cluster. It might then be easier to distinguish the normals from other cancer types than from adenocarcinomas.

The Garber et al. paper, source of the data for "lung", shows in a dendrogram on p. 13785 a clustering of the tissues where "Adeno" tissues are clustered into three groups, and some outliers. The three groups appear to differ substantially in mortality as shown in Fig. 4 on p. 13788.

The paper emphasizes adenocarcinoma in its title and otherwise. For some parts of the list of genes forming the rows, the main interest is differentiating the different groups of adenocarcinomas.

If I read correctly, the group members were as follows, where "_Adeno" is omitted since it appears with all of the following.

....

Group 1: codes 181-96, 132-95, 198-96, 156-96, 187-96, 180-96, 199-97_p, 199-97_c, 12-00, 137-96, 68-96, 257-97, 204-97, 11-00, 320-00_c, 320-00_p, 319-00PT, 313-99MT, 313-99PT.

There are 19 total tissue samples listed, from apparently 16 different patients, where 199-97, 320-00, and 313-99 appear to have two samples from each of these patients. The column numbers in "lung" belonging to Adeno group 1 are:

4, 5, 9, 11, 13, 15, (16), 24, 28, 29, (30), 34, 35, 40, 45, 46, 50, 55, (56).

Here again, column numbers in parentheses indicate a further tissue sample from the patient in the preceding column.

....

Group 2: codes 185-96, 178-96, 306-99, 306-99_node, 226-97, 222-97, 165-96.

There are 7 total tissue samples listed, from apparently 6 different patients, where 306-99 had two tissue samples. The column numbers in "lung" for Adeno group 2 are:

21, 31, 39, 64, 66, (67), 72.

- - - - -

Group 3: codes 218-97, 223-97, 80-96, 265-98, 184-96, 184-96_node, 234-97, 319-00MT1, 319-00MT2

There are 9 total tissue samples listed, from apparently 7 different patients, (the paper p. 13785 says "nine patients," but that seems not right), where 184-96 and 319-00 each had two tissue samples. The column numbers in "lung" for Adeno group 3 are:

7, 8, (10), 18, 36, (37), 58, 61, 69.

Actually columns 8, 9, and 10 are three tissues from one patient coded 319-00. Column 9, the "presumed primary" tumor, was clustered in Group 1, but columns 8 and 10 from "intrapulmonary metastases" in this group 3. The case is mentioned in the paper on p. 13785, near end of first column, and on p. 13788, end of first column and beginning of second.

- - - - - - - -

Adeno outliers:

code 69-96, column 6, clustered among squamous cell cancers (SCC).

codes 299-99 and 161-96, columns 17 and 54, clustered with each other and next with small cell cancers (SCLC).

codes 191-96, 147-96, and 237-97, columns 12, 41, and 70, clustered with large cell cancers (LCLC).

(d) The clustering of adenocarcinomas into Groups 1, 2, and 3 was apparently based on the dendrogram in the paper itself. Did you find that any of the above six outliers would have been assigned to one of the three groups, based on your rows?

3. (10 points) When one uses tests such as the Mann–Whitney–Wilcoxon or Kolmogorov–Smirnov two-sample tests, the test statistic has discrete values, only finitely many for given m and n. Thus when one forms order statistics of p-values in applying the Benjamini–Hochberg procedure, extended to possibly discrete p-values via Benjamini and Yekutieli's results, there may be ties among them and so among their order statistics, as in $p_{(i-1)} < p_{(i)} = p_{(i+1)} = \cdots = p_{(j)} < p_{(j+1)}$. Among the tied p-values $p_{(r)}$, $i \leq r \leq j$, can it happen under the procedure that some of the corresponding hypotheses are rejected and not others? Why or why not?

4. (20 points) Having in mind the "lung" data set, let's look into whether the Benjamini–Yekutieli hypotheses hold for data given by rows of the "lung" matrix (or similar data matrices). We know that the 916 rows of "lung" come from 835 distinct genes. Suppose we're applying m (e.g. 300 or 916) tests, whose p-values are identical in some subsets (e.g. two or more clones of one gene) or otherwise jointly independent (which may or may not be a fair assumption for the given set of genes). Then show that the "PRDS" assumption of Benjamini and Yekutieli does hold. *Hint*: let D be an increasing set. For each $i \in I_0$ (index of a gene for which the hypothesis of no difference in effects is

true), let F be the set of all indices j with $1 \le j \le m$ with p-values $p_j = p_i$ and G the set of other indices. So a point x can be represented as (y, z) where $y = \{p_k\}_{k \in F}$ and $z = \{p_k\}_{k \in G}$. We want to find the conditional distribution of x given $p_i = u$, what is it? For each given value of z, how does the event $(y, z) \in D$ depend on u, i.e. for what kind of set of values of u will it hold?

5. (30 points) For the 300 rows of "lung" you used in problem 2, to compare two kinds of tissues, suppose you have sub-matrices of m columns for one type and n for another, say typea and typeb. The column numbers for different types were given in Problem 2. In this case *do not* include "node" members (column numbers in parentheses, lymph node for patient whose lung cancer sample is in previous column) since we need an independence assumption for the columns. For patients with _c (central) and _p (peripheral) samples from the same tumor, include only one of the two. Also, for patient 319-00, who contributed three tissue samples, 319-00PT, 319-00MT1, and 319-00MT2 (one "primary tumor" and two "metatstatic tumors") include *only* 319-00PT (column 9, primary tumor, Adeno group 1) and not the metastatic tumors (columns 8 and 10, Adeno group 3). Garber et al.'s data are displayed at the website

http://genome-www.stanford.edu/lung_cancer/adeno

and then clicking on "data". It can be seen that the rows as well as columns are in different orders than in "lung". The data set has 918 rows. The "IMAGE:24661" row (identical except for beginning with "GENE917X" vs. "GENE915X") appears twice on the Stanford website. One of these was deleted from "lung," which has only one "IMAGE:24661," in the last, 916th row. Some other duplicating row apparently was deleted to leave 916.

You can see what the three Adeno subclusters were for the full data set in Fig. 1, Adeno groups 1, 2, 3. Test, for your rows, the one of the following indicated:
(a) LCLC vs. all other cancer tissues (i.e., all tissues other than normal, fetal_lung, and LCLC). Omit "combined" (column 20) from this (patient had both LCLC and SCLC). For rows 601-900.
(b) Likewise, SCLC vs. all other cancer tissues. Again omit "combined." For rows 301-600.
(c) Squamous cell vs. all other cancer tissues. For rows 1-300
(d) Adeno Group 1 vs. Adeno Group 3. For your rows, whatever they were.

Then you can test for differences in gene expression in your rows as follows: initialize a 100 by 2 matrix called pv, say, with pv[j,1] = j for each j, and however you initialize pv[j,2]; then run

```
for(j in 1:300)
{
pv[j,2] <- ks.test(typea[j,1:m],typeb[j,1:n])$p.value
}
```

if you are testing one type against a large, heterogeneous grouping. Replace "ks.test" by "wilcox. test" if testing one type against another (specifically, Adeno 1 vs. Adeno 3).

Sort the p-values (vector from the second column of pv) from smallest to largest; decide by the Benjamini–Hochberg method with $q = 0.1$ for which of the original rows the hypothesis of equality should be rejected. And so, see if the Garber et al. conclusions hold for your data rows.

Note: the Benjamini–Hochberg method originally applied to test statistics with continuous distributions, at least for true null hypotheses, which doesn't hold for the Mann–Whitney–Wilcoxon test (nor the Kolmogorov–Smirnov test), but this restriction is removed with the Benjamini–Yekutieli improvement. Recall that we found in an earlier pset that for location differences, which seem of interest here, the Mann–Whitney–Wilcoxon test is more powerful than the Kolmogorov–Smirnov test. When comparing two types, we might therefore prefer the Wilcoxon test. But to see if one type is different from many others combined, the difference might not be in location, so one might better use ks.test instead of wilcox.test.

Also, for two data sets of size 4 for example, the smallest possible p-value is $2/\binom{8}{4} \doteq 0.0286$, which is not very small, and after a correction for multiple tests is likely to show no significant differences. Thus, if one of two types being compared is relatively small, the other one needs to be large, such as all or many of the rest of the columns.

In each case, compare the number of rejected hypotheses you find via Benjamini–Hochberg with the number found via Bonferroni, namely, rejecting those hypotheses H_j for which the p-value is $\leq q/m$.

6. Extra credit problem (up to 50 points).

It would be interesting to discover which of the 916 rows of "lung" represent clones of the same gene. We are told that there are 835 distinct genes. So we'd like to do a clustering of the rows, rather than of the columns as previously. For hclust, clustering of rows is actually the default. A technical problem is the non-numerical "NA" (not available) entries in the data frame which appear fairly often. The pvclust code documentation p. 7 under “Usage” includes codes “ use.cor = ”pairwise.complete.obs” ” so maybe entries where a coordinate is “NA” are left out in evaluating correlations.

Cluster the rows into $m = 835$ clusters by an hclust method, say “average,” which we hope would correspond to the unique genes, and thereby estimate which rows represent different clones of the same gene. See how well the clusters correspond to those indicated by hand-written brackets in the margin of the printout containing descriptions of genes from the Stanford data base. (Caution: trying the “single” hclust method produced clusters of sizes 9, 8, 5, 4, 4,... whereas from the marking of the Stanford rows, one would expect clusters of sizes 4, 3, 3,...,3, 2,2,.......,2,1,........1.