STA 437/1005, Fall 2008 — Assignment #3

Due on November 24, at start of lecture. Worth 10% of the course grade.

This assignment is to be done by each student individually. You may discuss it in general terms with other students, but the work you hand in should be your own. In particular, you should not leave any discussion of this assignment with any written notes or other recordings, nor receive any written or other material from anyone other than your instructor by any other means, such as email.

For this assignment, you will analyse a data set on gene expression in yeast, at different points in its cell cycle. The data was collected for the analysis reported in the following paper:

> P. T. Spellman, *et al.* (1998) "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization", *Molecular Biology of the Cell*, vol. 9, pp. 3273-3297.

The paper above used other methods of analysis, but for this assignment, you should analyse the data using Principal Component Analysis (PCA). I have written some R functions for doing PCA that you should use. (R has a built-in function for PCA, but it doesn't handle data sets where the number of variables is greater than the number of cases.) These functions, the data, some R hints, and links to the text of the paper and to the original source of the data are available from the course web page:

`http://www.utstat.toronto.edu/~radford/sta437`

Look in the section for Assignment 3. (Note that it isn't necessary for you to read the original paper or consult the original data source to do this assignment.)

Yeast is a single-celled organism whose celluar organization is similar to that of humans (and other "Eukaryotes"). It has approximately 6000 genes, many of which have analogues among the 25000 human genes. Except when they decide to have sex, yeast reproduce by cellular division (producing "buds"). The yeast cells go through a cycle, in which they start out small, grow over time, divide, and then start growing again. During the course of this cycle, the activities of some groups of genes increase and decrease, as their functions are required for only part of the cycle. Some other groups of genes are active during the whole cycle, or may not be active at all during an experiment, if their functions (eg, for mating, or coping with adverse conditions) are not needed in the environment of the experiment.

The purpose of the study above was to determine which genes have activities that vary during the yeast cell cycle. Data on the activity (also called the "expression") of nearly all yeast genes was obtained using a DNA microarray. Measurements of gene activity cannot be obtained in this way for single cells; only the average activity for a large number of cells can be measured. To see what happens during the cell cycle, it is therefore necessary to obtain large numbers of cells that are all at (approximately) the same point in the cell cycle. In the data you will see, this was done in two different ways. For the "alpha" data, a cell culture was exposed to a pheromone that had the effect of synchonizing all the cells at a particular point in their cycle. Samples of cells were then taken at seven minute intervals after that, as they all grew and divided in approximate synchrony. For the "cdc15" data, the cells were arrested at a certain point in their cycle by low temperature, and then allowed to continue in approximate synchrony when the temperature was raised, with measurements taken every 10 or 20 minutes thereafter. These two methods of synchronization may have unintended side effects, and the different experimental environments result in different growth rates for the cells, so the data is not completely comparable. The experimenters expected that the "alpha" data will show about two cell cycles, and that the "cdc15" data will show about three cell cycles.

The data for the two experiments ("alpha" and "cdc15") that you will see is actually the log base 2 of the ratio of average gene activity in a large number of cells thought to be at a certain point in the cell cycle to the average gene activity of a large number of cells that are in all stages of the cell cycle. Positive numbers therefore indicate that a gene is more active than it typically is; negative numbers that it is less active than it typically is. The technology isn't perfect, so some data was missing. I replaced these missing measurements by zero, since that is a "neutral" value in this context. I also eliminated genes for which all data was missing for one or both experiments.

Previous research had identified 104 genes as probably varying in activity over the cell cycle. I have provided data for you identifying these genes. (Actually, only 92 are identified, since two systems of yeast gene naming are used, and I didn't succeed in converting all the names.) The original analysis used these genes as a guide to identifying other genes that vary in activity over the cell cycle, but for this assignment, you should use this data only at the end of the analysis, as described below.

The data is stored in a file with a header line giving the names of 43 variables, with one line after that for each of 5894 genes. The line for each gene starts with the gene name, after which the values of the 43 variables are given. The 43 variables are as follows:

1	prev	1 if gene was previously identified as varying over the cycle, 0 otherwise
2	alpha_0	Activity ratio of this gene just after administering the alpha pheromone
3	alpha_7	Activity ratio 7 minutes after administering the alpha pheromone
4	alpha_14	Activity ratio 14 minutes after administering the alpha pheromone
	...	
19	alpha_119	Activity ratio 119 minutes after administering the alpha pheromone
20	cdc15_10	Activity ratio 10 minutes after resuming growth at higher temperature
21	cdc15_30	Activity ratio 30 minutes after resuming growth at higher temperature
	...	
43	cdc15_290	Activity ratio 290 minutes after resuming growth at higher temperature

This data can be read with the `read.table` function, which will set the column names of the data frame it reads to the variable names, and the row names of the data frame to the names of the genes. You should convert this data frame to a matrix with the `as.matrix` function, since for this analysis you will need to transpose it (using the `t` function), so that the genes become the variables (ie, $p = 5894$), and the different time points become the cases (eg, $n = 42$ if both experiments are analysed together). Remember that the `prev` indicator is to be used only at the end of the analysis.

A major problem with analysing this data is that the 5894 genes are far too many to look at manually, and also greatly exceed the number of observations. It therefore makes sense to try to use PCA to reduce the number of variables from 5894 to something more managable. For this assignment, you should look at the first $k = 4$ principal components, and see whether these contain enough information to see the cell cycle, and identify which genes vary in activity during it.

There are several options when applying PCA to this data. First, we could apply PCA using all 42 observations from both experiments, or using just the 18 observatins from the "alpha" experiment, or using just the 24 observations from the "cdc15" experiment. Second, we could use the variables without scaling them (ie, use the covariance matrix), or we could first divide each variable by its standard deviation (ie, use the correlation matrix). This gives six combinations of options which you should try. You should always center the variables (ie, subtract their mean). (Note that centering and scaling will be done automatically by the `pca.vectors` function that I provide, if the right option is set.)

Having found the first 4 principal component vectors, you can find the projections of the observations on these vectors, which reduces the data set to one with 42 observations and 4 variables. Note that you can find the projections for all 42 observations even if you used only a subset of them to find the principal component vectors.

Once you have such a reduced data set with 4 principal component values, you should plot these values for the 18 observations from the "alpha" experiment and for the 24 observations from the "cdc15" experiment, and see whether any of these values seem to follow cycles. For each of the six ways of finding principal components, you should try to identify two variables that vary periodically (not in exactly the same way), and then make a 2D plot of the observations for each experiment with respect to these two principal components. We would hope to see something resembling circular movement as the cycle proceeds. Comment on what you have found, including whether using the covariance or the correlation matrix worked better.

Next, using whatever principal components seem like the best indicators of the cell cycle, you should devise some way of categorizing a gene as varying in activity over the cell cycle, or not. The paper above identified approximately 800 genes as varying in activity over the cell cycle. You should also select approximately 800 genes that seem the most cyclic. Finally, you should see what fraction of the previously identified cyclic genes are in the set that you identified. It is only at this last stage that you should look at the `prev` indicator in the data file, which identifies these genes previously thought to be cyclic. You should discuss why you used the method you chose for categorizing genes, and how well it worked.

You should hand in your discussions, the R output and R plots that support your conclusions, and a listing of the R commands you used.