Data analysis assignment: Principal Components for Microarray data.

Due in class Tue April 6th.

For this assignment, you will use microarray time-course data from the yeast cell cycle contained in the file **yeast_cycle.xls**. This is a preprocessed subset of the data first presented and analyzed in Spellman et al. 1998.

- T=15 equi-spaced time points cover 2+ cycles (cells were synchronized with the “cdc” method to allow observation of cycling behaviors).

- The data is from spotted arrays, and represents normalized log-ratios (red to green; the sample on the green channel came from un-sychronized cells) . ~800 genes were originally selected as showing periodic expression behavior, and hence being cell-cycle related. Of these, you have N=679 for which no entries were missing. Thus, you do not have to perform normalization, imputation of missing values, or preliminary filtering (identification of relevant genes) for this assignment.

- The file also contains row-standardized data for the first 12 columns/time points.

- THESE ARE THE SAME DATA YOU SAW ANALYZED WITH PCA IN CLASS.

Part 1: Repeat the analysis seen in class (Notes C; also Holter et al., 2000). You can use Minitab, or any other software of your choice). Your report should contain plots and tables of the PCA outputs, and articulated explanations of what these represent and how they can be interpreted.

As you work, make sure you address the following points. We use dimension reduction (PCA in this case) to:

1. Evaluate the “complexity” (dimensionality) of the data.
2. Identify basic expression patterns underlying the data, and interpret them.
3. Visualize the data in low-dimension through projections.
4. Eliminate noise and/or artifacts through low-dimensional reconstructions of the data.
5. Identify genes “close” to basic expression patterns whose interpretation is of particular interest.

Part 2: Code and apply re-sampling or random permutation procedures, as computational means to provide a statistical assessment of the results in Part 1. (Note: the CSE people are likely to take charge of this, but it is paramount that everyone in the group follows what is being done and be able to reuse/reproduce the algorithms) .

Concentrate on one or two pieces of output (eigenvalues and the scree plot; first two or three eigenvector and their plot; ranked proximities of genes to an eigen-direction or plane and their plot) and chose one among the following four options:

1. Evaluate sampling variability. This requires you to bootstrap the data (i.e. resample rows with replacement), and recompute the output of interest on each bootstrap data set (take N as the size for each, and produce M=small of them).
2. Evaluate stability. This requires you to implement perturbations by deletion (i.e. delete rows, or equivalently resample rows without replacement), and recompute the output of interest on each perturbed data set (take 0.8N as the size for each, and produce M=small of them).
3. Construct a "chance background" (null scenario). This requires you to implement random permutations (to scramble away certain features of the data while preserving others), and recompute the output of interest on each permuted data set (the size of each is N, and produce M=small of them). Pay special attention to what you are scrambling and what you are preserving. Hints: randomly permute cells within each column; randomly permute cells across the whole matrix.
4. Make up your own computational assessment scheme!

Results from Part 2 can be presented through tables containing relevant intervals, and/or superimposing bands on the plots representing output from the original data. For instance:

Eigenvalue	Actual	Simulated center (mean, or median)	Simulated L (mean-aSD, or qth quantile)	Simulated H (mean+aSD, or (100-q)th quantile)
Lambda_1				
. . .				
Lambda_T