

Stat597D, Spring 2001.

Homework assignment #F1

The excel file hmwF1_data.xls contains 16 columns and 801 rows. The first row contains column labels, the first column contains gene identification codes (orf), and the remaining entries contain data for N=800 genes over T=15 conditions.

This is a subset of the publicly available cdc yeast cell cycle data from Stanford. You have already encountered this data in some of your readings, and you will encounter it again in readings concerning clustering analyses.

The 15 conditions represent equally spaced (20 minutes apart, starting from minute 10) observation times on cycling yeast cells. The duration of a cell cycle is approximately 90—110 minutes, so the data ought to recover two full cycles, and some more.

The 800 genes represent a subset of 2467 annotated genes from yeast. These genes were selected as being the one whose expression profiles were most consistent with a periodic behavior, and thus the most likely to be involved in cell cycle related functions and processes.

The entries represent log ratios (base 2) of expression at the given time over expression in a reference condition. The numerators values are obtained from a synchronized culture, while the denominator ones (reference condition) are obtained from an unsynchronized culture, which ought to produce expressions “averaged” over the natural cycling. Beware of the last few time points, as it is claimed that synchronization might be failing, making the observations unreliable.

The entries are not centered, nor standardized, neither by row – gene, nor by column – time point. Missing entries are indicated by asterisks.

Our objective is to gather and interpret information on the structure of this data. In particular, we want to attempt a dimension reduction of the data set, interpret the relevant directions it produces, and investigate genes that seem to play a special role in this structure (e.g. position in the data cloud). In this respect, the excel file hmwF1_add.xls contains orf identifications and standardized short descriptions for the whole set of 2467 genes (this will provide you, through the orf, a short description for any gene(s) you might have isolated – the biologist in the group might then be able to form some interpretation and/or retrieve more information on the gene(s) via data base queries).

The crucial issues concern

1. Whether or not to consider all time points (some researchers have limited attention to the first 12 time points, discarding the last 3).
2. How to deal with missing entries (simply discard rows, or try to fill them – overall row average; linear interpolation over neighboring time points; column averaging on rows/genes that are similar to the row in question in its non-missing parts¹).
3. Whether to center and standardize the data by row.

¹ This requires a definition of profile similarity, possibly of an appropriate distance to tell how far apart two rows/profiles are – something we will deal with when we will study clustering methods.

4. Whether to standardize the data by column (i.e. refer to the correlation, as opposed to the variance/covariance matrix, in our analysis; recall that centering by column is “implicit” in both PCA and FA).
5. Whether to use a simple PCA, or frame our analysis through the decomposition model behind an FA (recall that if we perform FA with the principal components method, we are using the same directions as in PCA).
6. How many directions to retain (considering percentages of retained variability, eigenvalues above average, scree-plots shapes, possibly tests – but beware of the gaussian requirement implicit in their use).
7. Whether to rotate (i.e. change coordinate system) in the subspace spanned by the directions we retain, to improve interpretation (although the directions produced by PCA are appealing as natural variability axes of the data, we might still want to rotate to a new coordinate system of, say, the first principal plane, if this substantially helps interpretation).
8. Whether to repeat the analysis after deletion of a certain collection of genes (for example to show that without them, the number of directions to retain decreases).

You can perform your analysis using any software you want.

Minitab will allow you to

- Perform PCA with the var/cov or with the correlation matrix (without having to standardize columns on your own), with various outcome storage options. You are on your own for rotations.
- Perform FA, with the var/cov or with the correlation matrix, the principal component and maximum likelihood method, various rotation options, and various outcome storage options.
- Use graphics.

Your write-up should detail and motivate your choices relative to points 1—8 above. You can report results for various options (e.g. PCA with var/cov AND with correlation, PCA with AND without certain point/genes) in a comparative fashion, if you find the comparison interesting. The write-up should contain relevant software output and plots, but with parsimony and appropriate editing (stay within a max of 15 pages, including figures and tables).