

## Stat597D, Spring 2001. Homework assignment # 4

### Microarray Analysis Section: (Francesca)

This assignment concerns a subset of the data in Golub et al. 1999, which investigated the classification of leukemia (thanks to Jessica Martinelli, who is working on these data for her thesis, for helping with the preparation of the assignment). The original data set, as well as additional information of various types, can be found at [http://waldo.wi.mit.edu/MPR/data\\_set\\_ALL\\_AML.html](http://waldo.wi.mit.edu/MPR/data_set_ALL_AML.html).

Samples of cells were obtained from bone marrow or peripheral blood of 72 individuals with Leukemia, and expression levels for each sample were recorded with Affimetrix (one-color) arrays. The arrays had 7,129 locations. Circa 6,800 of them are relative to human genes. The others are controls and replicates. Together with expression, a number of other variables were recorded for the samples (see the website). Among those is a classification of Leukemia-types as ALL (lymphoblastic Leukemia) and AML (acute myeloid leukemia), and further as T-cell or B-cell if ALL.

Expression levels were *normalized* to ensure comparability across samples/arrays. The first sample was taken as “baseline” (normalization factor = 1), and normalization factors for each of the other 71 samples were computed through a regression based-procedure involving genes with “P” (present) Affimetrix calls in all samples. Information on the normalization procedure and a list of the normalization factors can be found at the website). After normalization, levels were “*thresholded*” by setting to 100 all values  $\leq 100$ , and to 16,000 all values  $\geq 16,000$ . Let  $X_{ij}$ ,  $i=1 \dots 7,129$ ,  $j=1 \dots 72$  indicate the normalized and thresholded levels.

A filtering was then applied, to eliminate from the analysis genes presenting insufficient variation across samples. After computing minimum and maximum of  $X_{ij}$  over  $j=1 \dots 72$  for each  $i$ :  $min_i$ ,  $max_i$ , all  $i$ 's such that  $max_i/min_i \leq 5$  or  $max_i - min_i \leq 500$  were eliminated. This identified a subset of 3,571 genes, which is the one used in following studies of the same data (Dudoit et al., 2000, West et al., 2000). In order to construct a feasibly small data set for the assignment, we applied a much more stringent filtering based on the same criterion: we eliminated all  $i$ 's such that  $max_i/min_i \leq 30$  or  $max_i - min_i \leq 3000$  were eliminated, which left us with a subset of 886 genes. You are strongly encouraged to read the analyses performed in Golub, Dudoit and West, but notice that since you are working on a much smaller set of genes, your results may very well not resemble theirs.

Golub et al. also divided the samples at the outset into a training set (38 samples), and a validation set (34 samples).

The data in the excel sheet hmw\_F2\_data.xls consists of

- Two initial columns containing short description and identification of genes (gene accession numbers)
- 72 pairs of columns, one for each sample. The first column in the pair (samp\_j) contains normalized and thresholded expression levels, the second (call\_j) the affimetrix call (present or absent, used in the normalization procedure).

- 4 additional columns containing  $max_i$ ,  $min_i$  and  $max_i/min_i$ ,  $max_i - min_i$  for each gene/row (used in the filtering step).
- 886 rows corresponding to genes
- A row classifying samples into TRA=training and VAL=validation; this expresses Golub's original partition of the sample. You are not bound to use the same for your validation procedure (e.g. you could use the whole data set, and perform cross-validation leaving out a certain percentage at a time). This row is provided only so that you can reconstruct Golub's sets, if you want to compare your analysis to theirs.
- A row classifying samples into ALL and AML. This is a binary response,  $Y_{bi}$ .
- A row classifying samples into B-cell, T-cell (sub-classification of ALL) and AML. This is a categorical response in three categories,  $Y_{tri}$ .

Note that expression levels are normalized and thresholded, but they are not centered or standardized, neither by gene nor by sample. Moreover, they are on the original scale (no log transformations). The various authors that worked on these data have used log transformations, centering and standardization in various ways and at various stages of their analysis.

- Decide whether to apply logs, and row/column centering or rescaling (briefly motivate your choices).
- Apply the 2-phase dimension reduction strategy described in class to identify:
  - the single direction (derivative predictor  $F(Y_{bi})_1$ ) that best separates ALL and AML
  - the plane or single direction (derivative predictors  $F(Y_{tri})_1$ ,  $F(Y_{tri})_2$ , or simply  $F(Y_{tri})_1$ ) that best separates B-cell, T-cell and AML

The  $X$ -based phase does not involve the response (so it is the same in both cases), and is performed through a singular value decomposition. As for the  $Y$ -based phase, use discriminant analysis or SIR (of course here the outcome will differ depending on the response you are using).

- Since the trinary response is obtained by a further sub-classification of the binary one, an interesting question concern the relationship between the identified subspaces. Is the direction for  $Y_{bi}$  very close to (contained in) the plane/direction for  $Y_{tri}$ ? As usual, you can evaluate this using determination coefficients for the regression of  $F(Y_{bi})_1$  on  $F(Y_{tri})_1$ ,  $F(Y_{tri})_2$ , or simply of  $F(Y_{bi})_1$  on  $F(Y_{tri})_1$ .

- Restricting attention to the binary  $Y_{bi}$ :
  - Fit a logistic model for  $Y_{bi}$  on  $F(Y_{bi})_1$ , and classify the 72 samples accordingly. Provide a misclassification rate.
  - Rank genes in terms of how close their profiles over the samples are to the single best direction (i.e. in terms of determination coefficient – squared correlation coefficient – for the regression of  $X_i$  on  $F(Y_{bi})_1$ , as discussed in class).

Recall that discriminant analysis and logistic regression routines are available in Minitab, while a SIR routine is available in ARC. But you can use other statistical software, and concerning the 2-phase dimension reduction, you can also implement the needed matrix calculations and decompositions using S+ or Matlab.

Now pick one among the following four questions (you need to work out only ONE):

1. Perform a cross-validation study for the model in d.(i), dividing the 72 samples into training and validation groups.

2. Using the short descriptions, and other information you may get from data base searches, give a biological interpretation of the top-ranking genes from d.(ii).
3. (This is an open-ended question) Design and implement a permutation or re-sampling study to decide how many of the top-ranking genes from d.(ii) to designate as relevant for  $Y_{bi}$ .
4. (This is a theoretical question). Consider a generic regression problem with  $p$  predictors, say  $W_1 \dots W_p$ , a binary response  $Y_{bi}$ , and a trinary response  $Y_{tri}$  obtained by further partitioning one of the classes of the binary response. You have, say  $T$  observations, and you can consider the overall, between and within covariance matrices of the predictors, as in the class notes. Choose one between discriminant analysis and SIR. Using algebra on the corresponding “standardized” versions of the between matrices, and their spectral decompositions, try to work out a proof that the single direction that best separates the binary response is always contained into (coincident with) the plane (direction) that best separates the trinary response.

**Computational Methods Section:** (David)

Suppose that  $X$  takes on the possible values 0, 1, 2, whereas  $Y$  takes on values 0 and 1. The conditional distribution of  $X|Y$  is binomial with parameters  $[2, (Y+1)/4]$ . The conditional distribution of  $Y|X$  is binomial with parameters  $[1, (2X^2+1)/(2X^2+2)]$ .

Implement a Gibbs sampler to estimate the joint distribution  $(X, Y)$ . Present the results in a 2 by 3 table. Discuss how you choose a burning period.. Are you able to use this table (or another method) to find the true joint distribution?