SD-08

Mixed Models Analysis of Microarray Experiments Using Pooled Error Estimates

Yuan Liu and James Blum, UNC-Wilmington, Wilmington, NC

Abstract:
Our current efforts involve modification of "gene by gene" models for analyzing microarray experiments to include the possibility of similar error distributions among groups of genes, thereby reducing the complexity of the overall model and enhancing the power of effect tests. This paper presents implementations of these ideas using the SAS System, including multiple grouping methods, diagnostics and summarization. Comparisons of the proposed and current methods will be discussed in relation to applications to experiments from recent literature.

1. Introduction

Background

In microarray experiments, typically several thousand gene expressions are measured simultaneously with the goal of determining differential expression related to some characteristic of the subjects in question. This characteristic may be different levels of a treatment application, response categories, quantitative response or a combination of several of these. Among many other methods, ANOVA models have been proposed for use in determining the nature of differential expression.

ANOVA and Mixed model

Every measurement in a microarray experiment is associated with a particular combination of an array in the experiment, a dye (red or green), a variety or treatment (or some combination of these, and a gene. Let y_{ijkgr} be the log fluorescent intensity from the r^{th} spot for gene g on array i for dye j and variety k. A typical ANOVA model for a micorarray experiment (Kerr et al., 2001) can be the form of

$$y_{ijkgr} = \mu + A_i + D_j + (AD)_{ij} + G_g + (AG)_{igr} + (DG)_{jg} + (VG)_{kg} + \varepsilon_{ijkgr}. \qquad (1.1)$$

Here, μ is the overall mean expression level; the array effects A_i account for differences between arrays averaged over all genes, dyes, and varieties; the dye effects D_i account for differences between the average signal form each dye; $(AD)_{ij}$ is the term accounting for effects of the interaction between the array and the dye. These three kinds of effects account for overall variation in array and dyes and also are considered as "global" effects. They are not of interest, but accounting for them amounts to data normalization. In addition to these "global" normalization terms, there are source of variation to consider at the level of individual genes. They are the terms of G_g ,$(AG)_{igr}$, $(DG)_{jg}$ and $(VG)_{kg}$ in ANOVA models (1.1) and are considered as "gene-specific" effects. The gene effects G_g account for the expression level of

genes averaged over the other factors. The $(AG)_{igr}$ terms account for the average effect of the spot on array i for gene g. The $(DG)_{jg}$ terms account for the effect of dye j on gene g. The variety-by-gene term $(VG)_{kg}$ represent levels of signal intensity for genes that can specifically be attributed to the RNA production differences varieties under study, thus being the effect that is of primary interest in our analysis.

Similarly model 1.1 can be specified in two stages (Wolfinger *et al.* 2001): normalization model and gene-specific model:

Normalization model: $y_{ijkgr} = \mu + A_i + D_j + (AD)_{ij} + \gamma_{ijkgr}$ (1.2)

Gene-specific model: $\gamma_{ijkgr} = G_g + (AG)_{igr} + (DG)_{jg} + (VG)_{kg} + \varepsilon_{ijkgr}$ (1.3)

Normalization plays an important role in the earlier stage of microarray data analysis, removing experiment-wide systematic effects that could bias inferences. Most common is red-green bias due to differences between the labeling efficiencies and scanning properties of the two fluoresces. Locally weighted scatter-plot smoothing (LOWESS) is also widely used to attempt to normalize these differences. The normalization methods of using statistical models (model 1.1 and 1.2) assume various effects are additive, and hence they are similar to global normalization which subtracts a global constant to get normalized values. (Gordon *et al.* 2002). Normalization model (1.2) mainly contains the "global" terms in model 1.1. For the remainder of this discussion, we will assume some appropriate normalization has been applied prior to conducting the analysis using the gene-specific model.

A practical advantage of two-stage model is computational feasibility. The two-stage model can be viewed as a computationally tractable re-formulation of model (1.1). General statistical software (SAS included) has difficulty handling model (1.1) because of the large number of parameters, especially when considering some effects as random in the application of a mixed model. In this case complex computational methods are involved and it will be difficult to fit a full mixed model in a single attempt. These two separate models provide a conceptually and computationally efficient means to analyze microarray data.

Local Pooling of Errors (LPE)

An early proposal of then concept of local pooling of errors (LPE) was put forth by Jain *et al.* (2003). Their approach is designed to improve estimates of variability, and the power of statistical tests, by pooling the errors across genes that have similar expression intensity values. This method is based on the assumption that the variance of individual gene expression measurements is a (likely non-linear) function of intensity. Without going into technical details, the method locally combines information across genes at similar intensity level for estimating error variances. The LPE statistic for the median (log-intensity) difference for each gene under the two compared the conditions is then calculated as:

$$Z = \frac{Med_1 - Med_2}{\sigma_{pooled}}$$

Where $Med_i, i = 1,2,$ is the median intensity of the i th sample;

$$\sigma^2_{pooled} = k \cdot [\sigma_1^2 / n_1 + \sigma_2^2 / n_2]$$

Where n_1 and n_2 are number of replicates in the two array samples being compared; $\sigma_i^2, i = 1,2,$ is the estimate of variance for the i th group.

While we will use a different approach to local pooling, one that is suitable to the two-stage modeling discussed above, the motivation is similar. One would not expect an assumption of a common residual variance across several thousand genes to be reasonable, so a single ANOVA model assuming homogeneity would not be appropriate. However, fitting a gene-specific model one gene at a time (Wolfinger, *et al.* 2001) models each gene with a different variance parameter. While this is certainly workable, it would also seem likely that the error variances are not entirely heterogeneous. Therefore, it should be possible to combine information on genes that show similar residual error thus increasing the effective sample size and power of the resulting tests.

2. Method

The method proposed here will be centered on the two-stage model, but will actually be comprised of three stages of modeling. The normalization model is run in the initial step (We will assume a base 2 logarithmic transform prior to fitting the normalization model.), and the residuals from this model are used in the response in the gene-specific model, which is run one gene at a time. Residuals for each gene specific model are determined as well.

Normalization model: $y_{ijkgr} = \mu + A_i + D_j + (AD)_{ij} + \gamma_{ijkgr}$ (2.1)

Gene-specific model: $\gamma_{ijkgr} = G_g + (AG)_{igr} + (DG)_{jg} + (VG)_{kg} + \varepsilon_{ijkgr}$ (2.2)

These models are fit using PROC MIXED from the SAS/STAT software, taking the array effect and all interactions involving it as random. Residuals are computed from the OUTP= option in the MODEL statement. It should be noted that care must be taken with respect to variable names here. The default name for the residual is "RESID" from the OUTP= option; however, at the gene-specific level we are already using a residual from a previous mixed model and attempting to generate residuals for another model. If we have not renamed the residual from the normalization model, SAS will attempt to use the same name for both sets of residuals and only the first will be preserved.

Once the residuals from the gene specific models are found, some measurement of variation, σ_g, is computed for each (we will choose standard deviation). Next, the genes will be grouped based on these measures of variation. Several methods are possible, one of the most basic is to break the distribution into k quantiles, $q_1,\ldots,q_k$. Then if, for gene g, $q_i < \sigma_g \leq q_{i+1}$ $\{i = 1, \ldots, k\text{-}1\}$, that gene is placed in group i. (We may assume that $q_1 = 0$.) In SAS, the UNIVARIATE procedure can be used to construct and store the quantiles, which are then changed to macro variables so that group assignments can be done in a DATA step.

Once the groupings are determined, model (2.2) is run again, this time by groups of genes instead of by individual genes. Within every group the effect of interest is for the variety-by-gene term (VG), for which we can extract tests for variety specific differences for each gene via the SLICE= option (*e.g.* LSMEANS variety*gene /SLICE=gene). Finally, the p-values from the slices are run through the MULTTEST procedure to make adjustments for multiple testing, with the Bonferroni method being a popular choice for the adjustment.

3. Further Considerations

3.1 Group "Checking"

There is no guarantee that grouping genes based on quantiles of residual variation will produce groups that have homogeneous variance, particularly if the number of quantiles/groups selected is small relative to the number of genes. Even if the number of groups is large, there is no assurance that each group will be homogenous. Since the residuals are available however, homogeneity within groups is an assumption that can be checked or tested. Standard plots for assessing homogeneity can be constructed from the absolute values of the residuals or the squared residuals, and Levene's test for homogeneity can be constructed from these as well (in PROC GLM, for example).

In tests of this procedure on data sets with arrays consisting of 5,000 to 7,000 genes, groups of 50 to 70 genes have been constructed. Some groups (less than 10 in tests thus far) have failed to demonstrate homogeneity in all trials. As a remedy for this eventuality, each group that fails is bisected and re-tested until the resulting groups pass the homogeneity test. If plots are constructed of the squared or absolute residuals, a more refined, but manual, decision can be made as to how to divide these groups.

3.2 Other Grouping Methods

As opposed to starting with equal size groupings based on equally spaced quantiles, grouping of genes could be based directly on the magnitude of the variation itself. For example, if the minimum and maximum values of σ_g (σ_{min} and σ_{max}) are known, and a specific distance d is chosen, groups can be constructed with group i defined as all genes such that $\sigma_{min} + (i\text{-}1)d < \sigma_g \leq \min\{\sigma_{min} + i \cdot d, \sigma_{max}\}$, $i = 1,2,\ldots$ This allows for a much more specific definition of a neighborhood of variances, which goes more directly to the production to homogeneous groups. The groups are not likely to be equal in size, which poses little difficulty except for the common occurrence of having several hundred genes in a group at the lower end of the variance spectrum.

In this case, computational difficulties can become an issue, particularly on the average desktop computer.

Another approach is to use a clustering algorithm to determine the groups, such as a k-means procedure using the FASTCLUS procedure. Here the user has indirect control over the size of the neighborhood of variances through the number of clusters, so if any groupings fail the homogeneity test, increasing the number of clusters should cause the most heterogeneous groups to separate. It may also be feasible to search a hierarchical joining of the variability estimates for an appropriate cut-off resulting in homogeneous groups.

3.3 Other tests

Given that the F-test for the effect slices will produce several significant results, there is likely further inference that one would wish to conduct in these cases. In instances where a particular gene shows a significant difference in expression across several varieties or treatment conditions, the next likely step is to conduct pair-wise comparisons among the varieties for each of those genes. This problem is a bit cumbersome.

Consider a situation where there are four varieties in question and the genes are set in groups of 50, which would be a reasonable case in practice. If one wishes to construct contrasts to handle the pair-wise comparisons, the coefficient set is of size 200. Some creative programming using the SAS Macro Language can simplify this, but it is difficult to make fully automatic.

The user could also ask for all pair-wise comparisons in the LSMEANS statement by using the DIFF option; however, this must be requested for the (*VG*) effect, resulting in many unwanted comparisons. Rectifying this can be done in a few relatively simple steps. First, the set of all pair-wise comparisons can be reduced to the set of comparisons among varieties for each gene. If, for example, the comparisons are conducted on an effect labeled as gene*variety in SAS, then the output data set will contain the variables gene and _gene. So reduction of the set of comparisons amounts to selecting the set of observations where these variables match. Presuming that the set of by gene slices have been reduced to those that are significant following a suitable multiple testing adjustment, the two resulting data sets can be merged, keeping only the initially significant genes. Finally, the raw p-values for the pair-wise comparisons can be adjusted in some suitable fashion.

4. Conclusions

At the time of this writing, study of the methods proposed here has been limited to examining some existing microarray data. For the *Saccharomyces cerevisiae swi/snf* data set of Sudarsanam *et al.* (2000) studied by Wolfinger *et al.* (2001), the proposed method find four to seven times as many significant genes versus the one gene at a time method, depending on how conservatively the gene groups are constructed. Extensive simulation of these methods is warranted (and is being undertaken) to study the effects of various choices of grouping on power of tests and false discovery rates. Hopefully, this will lead to the ability to construct mixed-models for microarray

data in a plausible, yet parsimonious, manner; adequately describing the nature of the data but maximizing computational and statistical efficiency.

5. References

Jain, N., Thatte, J., Braciale, T., Ley, K., O'Connel, M., Lee, J. (2003) *Bioinformatics* 19:1945-51.

Kerr, M.K., Churchill, G. (2001) *Genetical Research* 77:123-28.

Smyth, G.K., Yang, Y.H., Speed, T. (2002) *Methods Mol Biol* 224:111-36.

Sudarsanam, P., Vishwanath, R.I., Brown, P.O., Winston, F. (2000) *Proc Natl Acad Sci* 97:3364-9.

Wolfinger, R.D., Gibson, G., Wolfinger, E.D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., Paules, R. (2001) *J Comp Biol* 8:625-37.

6. Contact Information

Your comments and questions are valued and encouraged.
Contact the author at:

James Blum
University of North Carolina-Wilmington
601 S. College Road
Wilmington, NC 28403-5970
Work: (910) 962-4299
Fax: (910) 962-7107
Email: blumj@uncw.edu
Web: http://people.uncw.edu/blumj