A random varaince model for differential gene detection in small sample microarray experiments

by

George W. Wright and Richard M. Simon

National Cancer Institute Biometrics Research Branch 6130 Executive Blvd MSC

7434 Bethesda MD 20892-7434

Abstract

Motivation: Microarray techniques provide a valuable way of characterizing the molecular nature of disease. Unofortunately expense and limited specimin availability often lead to studies with small sample sizes. This makes accurate estimation of within gene variance difficult, since variance estimates made on a gene by gene basis will have few degrees of freedom, and the assumption that all genes share equal varaince is incorrect.

Results: We propose a model by which the within gene varainces are drawn from a inverse gaussian distribution, whose parameters are estimated across all genes. This results in a test statistic that is a minor variation of those used in standard linear models. We demonstrate that the model assumptions are valid on experimental data, and has more power than standard tests to pick up large changes in expression, while not increasing the rate of false positives.

1 Introduction

Microarray technology allows a scientist to view the expression of thousands of genes from an experimental sample simultaneously. By observing changes in expression levels across multiple samples it is possible to generate and test a multitude of hypotheses relating gene expression to other characteristics of the samples. One of the primary goals of microarray analysis is to identify genes whose expression level varies between different classes of samples. Testing whether a single observed quantity varies across different classes of observations is a problem that is well understood, but the volume of information available on each sample creates new analysis challenges.

In modeling gene expression we use a linear model framework. This model is very versatile, and can be applied to a large number of different experimental designs. Further, other familiar tests such as the t-test, paired t-test and F-test, and Analysis of Variance (ANOVA) are actually special cases of the more general linear model. Kerr and Churchill (2001) suggested the application of ANOVA to the red and green log intensity values as an alternative to the use of log ratios. But even at the level of log ratio data, use of an ANOVA type model can provide a very powerful analysis framework, allowing the researcher to account for multiple competing factors that might influence gene expression.

A question that arrises when fitting a linear model to microarray data is what

estimate to use for the variance of the residual error. One approach is to form a single linear model that pools the residual sum of squares across all genes into a single variance estimate. This method makes the assumption that once all of the factors included in the model have been taken into account, all genes are equally variable. In practice we find that this is not a valid assumption. It is impossible to take into account all reasons for genetic variation in a single linear model. Those factors that have not been taken into account will vary from gene to gene, and be incorporated into the residual, leading to large differences in the residual variance across genes. A second approach is to form a separate linear model for each gene, and estimate the gene-specific residual variance using only the data from that gene. This method has the disadvantage that there is no information shared between genes. If the sample size is small, there may be very few degrees of freedom available to estimate the residual variance, leading to statistical tests with low power.

We propose a hybrid approach, in which it is assumed that the variance of the residuals change from gene to gene, but represent random selections from a single distribution. By observing the residual sum of squares within each gene,we estimate the form of this distribution. Then for an individual gene we adjust the observed residual sum of squares in light of the distribution. By sharing the variance estimate across multiple genes, we can form a better estimate for the true residual variance of a given gene, and effectively boost the residual degrees of freedom. The test statistics

produced by this shared variance model are very similar in form to those for standard linear models, meaning that this model can be easily implemented using standard statistical packages.

2 Model formulation

We will denote by y_{ij} the normalized expression values for sample i and gene j. How the expression values are actually formulated will depend on the application. If we wish to follow the methodology of Kerr and Churchill., then the y_{ij} will represent normalized log intensity values of a single channel. Alternatively, the y_{ij} can represent the normalized log-ratios of a 2 color array, or normalized log signal for Affymetrix GeneChip$^{\mathrm{TM}}$ Arrays

The type of model we wish to consider is the following

$$y_{ij} = x_i' \beta_j + \varepsilon_{ij} \tag{1}$$

The x_i's are vectors of design variables specific to the sample, the β_j's are vectors of unknown coefficients that are specific to a particular gene, and the ε_{ij}'s are unobserved residuals with mean 0 and unknown variance. The x_i's represent the characteristics of the samples that we wish correlate with gene expression. For notational simplicity we have suppressed the intercept term that is often included in linear models. But if required this can be represented by setting the first component of each x_i vector to be identically equal to 1, in which case the first component of the β vector will be

the intercept.

In practice the investigator is often interested in determining the gene expression between two or more classes, in which case the x_i' will vectors of class indicators. For example when testing between treated and untreated samples, the first component would be 1 for the treated cases and 0 for the untreated cases, while the second component would be 0 for the treated cases and 1 for the untreated cases. Models involving more than 2 classes can be represented by increasing the dimension of x_i. This model is not restricted just to class comparison problems. For example, it may be reasonable to consider fitting a regression of gene expression against time, in which case x_i could be the continuous variable representing elapsed time.

The β_j's are variables that connect the sample characteristics to the expression of a single gene. For example in the case of treated and untreated samples outlined above, the first component of each β_j would represent the average expression of gene j for the treated cases, while the second component would represent the average expression for the untreated cases.

Let n denote the number of samples, m denote the number of genes, and k denote the dimension of each x_i and β_j. We use X to denote the $n \times k$ dimensional design matrix whose columns are x_i, and make the additional simplifying assumption that X is of full rank. For X not of full rank it is always possible to reparameterize the model into a lower dimensional model with X of full rank.

The crux of our model is in the handling of the ε_{ij}'s. We assume that for each gene j,

$$\varepsilon_{ij} \sim N\left(0, \sigma_j^2\right) \tag{2}$$

with the σ_j^2's being random variables themselves with an inverse Gamma distribution. That is

$$P(\sigma^{-2} = x) \sim G(x; a, b) \equiv \frac{x^{a-1} \exp(-x/b)}{\Gamma(a) b^a} \tag{3}$$

for some unknown parameters a and b. This choice of the Inverse-Gamma distribution as a prior distribution of variance is a standard choice in Bayesian analysis due to its computational convenience, but as we will show later, it models true variance structure of microarray data surprisingly well. We will refer the linear model described above as the Randomized Variance Model, or RVM.

This use of a distribution for σ is similar to a model suggested by Baldi and Long (2001) with some exceptions. First, our analysis is frequentist perspective rather Bayesian. Second, we consider the more general case of linear models, while Baldi and Long concentrated on a 2 class distinction. Third, we make the assumption that the a and b parameters in the prior of σ_j^2, are the same for all genes, and show how they can be estimated from the data. Finally, we make no assumptions about a prior distribution for β_j. This was a conscious decision on our part since the distribution of β depends on how the genes on the array were chosen. A significant proportion of the genes on the array are likely to have constant expression across samples. Any

prior distribution on the mean structure would have to include a singular component at zero, and any hypothesis tests on β would be heavily dependent on the choice of prior.

3 Linear Hypothesis testing for β

Hypothesis tests for β_j will be performed on a gene by gene basis, therefore in this section and the next, we will suppress the j subscript, and concern ourselves with tests of a single gene. In these sections we will also make the temporary assumption that the parameters a and b are known constants. In a later section we will show how to use the full set of data to estimate values for a and b, and argue that the substitution of these estimates for the true values will not invalidate the tests. For the sake of brevity, mathematical proofs of all claims have been omitted, but are available in the supplementary materials.

Following the standard framework of hypotheses on linear models, we wish to test the hypothesis $H_0 : \beta \in \omega$, where ω is a linear subspace of R^k. We will use $r \equiv k - \dim(\omega)$ to denote the number of linear constraints imposed by ω. In class comparison problems, the subspace is likely to be that all of the components of β are equal; representing that all classes have the same average expression. In that case r will be equal to the number of classes minus 1. In regression problems, the constraint will be that some of the components of β are equal to zero, meaning that change in

the respective component of x_i has no effect on gene expression.

In standard linear models in which there is no distribution for σ, the maximum likelihood estimate for β over R^k is given by

$$\widehat{\beta} \equiv (X'X)^{-1}Xy \tag{4}$$

and the maximums likelihood estimate under ω, is

$$\widehat{\widehat{\beta}} \equiv (X'_\omega X_\omega)^{-1}X_\omega y \tag{5}$$

where X_ω represents the design matrix X projected into the subspace ω.

To test of the hypothesis, $H_0 : \beta \in \omega$ against the alternative $H_1 : \beta \in R^k$ we would consider the respective sum of square residuals,

$$\widehat{SS} = \left\| y - X'\widehat{\beta} \right\|^2, \text{ and } \widehat{\widehat{SS}} = \left\| y - X'\widehat{\widehat{\beta}} \right\|^2 \tag{6}$$

and then the likelihood ratio test statistic would be

$$\mathcal{F} = \frac{n-k}{r} \frac{\widehat{\widehat{SS}} - \widehat{SS}}{\widehat{SS}} \tag{7}$$

which under H_0 has an F distribution with r and $n-k$ degrees of freedom.

If we repeat this type of analysis under the RVM assumption of a distribution for σ^2, expressed in equation (3) we obtain the same maximum likelihood estimates, $\widehat{\beta}$ and $\widehat{\widehat{\beta}}$, for β within R^k and ω respectively, but there is a change in the likelihood ratio test statistic. The residual sum of squares in the denominator of the $\mathcal{F}$ is replaced

with

$$\widetilde{SS} = \widehat{SS} + 2b^{-1} \tag{8}$$

and the number of degrees of freedom in the denominator change from $n - k$ to $n - k + 2a$. Thus the adjusted statistic becomes

$$\tilde{\mathcal{F}} = \frac{n - k + 2a}{r} \left(\frac{\widehat{\widehat{SS}} - \widehat{SS}}{\widehat{SS}} \right), \tag{9}$$

which under H_0 will have F distribution with r and $n - k + 2a$ degrees of freedom (see supplementary material).

To interpret this change, it is useful to consider

$$\hat{\sigma}^2 = \frac{\widehat{SS}}{n - k} \text{ and } \tilde{\sigma}^2 = \frac{\widetilde{SS}}{n - k + 2a} \tag{10}$$

which are the maximum likelihood estimates for σ^2 under the standard linear model and under RVM. After some algebra we find that

$$\tilde{\sigma}^2 = \frac{(n - k)\hat{\sigma}^2 + 2a\,(ab)^{-1}}{(n - k) + 2a}. \tag{11}$$

ab is the mean value for σ^{-2} under its distribution, so $(ab)^{-1}$ could be thought of as an estimate of σ^2 based on the shared variance distribution, and $\tilde{\sigma}^2$ as a weighted average of this and the sample variance for the specific gene. The degree to which each of these is weighted, depends on the number of samples, and the parameter a. The larger the sample size the more confident we will be in $\hat{\sigma}^2$ our sample estimate of σ^2. On the other hand, large values for a will indicate a highly peaked Gamma

distribution, making it more likely that the true σ^2 is close to $(ab)^{-1}$. The increase in the degrees of freedom from $n-k$ to $n-k+2a$, indicates the additional information about σ^2 provided by the distribution. Again a larger value for a represents a more informative distribution and so more certainty about $\tilde{\sigma}^2$ as an estimate of σ^2, that in turn translates as more degrees of freedom.

4 Application to t and F tests and ANOVA

A very common problem in micro array analysis is that of determining genes that are differentially expressed between 2 or more tissue varieties. To represent this in our framework, we set k equal to the number of groups, and set the pth component X_j to an indicator of membership of sample j in variety p. Just as our hypothesis tests for the linear model under the RVM assumption was similar to the standard tests for linear models, so also are our tests for differences across varieties of similar form to the t and F tests that are often used.

For testing between 2 varieties, with sample means $\widehat{\mu}_1$, $\widehat{\mu}_2$, sample variances $\widehat{\sigma}_1^2$, $\widehat{\sigma}_2^2$, and sample sizes n_1, n_2 the standard test statistic is

$$t = \frac{\widehat{\mu}_1 - \widehat{\mu}_2}{\widehat{\sigma}_{pooled}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}, \quad \text{where} \quad \widehat{\sigma}_{pooled} = \frac{(n_1-1)\widehat{\sigma}_1 + (n_2-1)\widehat{\sigma}_1}{n_1+n_2-2} \tag{12}$$

and this has a t distribution with $n-2$ degrees of freedom.

In the case of the RVM model,

$$\widetilde{t} = \frac{\widehat{\mu}_1 - \widehat{\mu}_2}{\widetilde{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}\ , \quad \text{where} \quad \widetilde{\sigma} = \frac{(n-2)\widehat{\sigma}^2_{pooled} + 2b^{-1}}{(n-2) + 2a} \tag{13}$$

and the degrees of freedom increase to $n - 2 + 2a$.

For testing $k > 2$ varieties, it is common to consider the following F-test statistic, involving mean sums of squares (MSS)

$$\mathcal{F} = \frac{\text{MSS(between varieties)}}{\text{MSS(within varieties)}} \tag{14}$$

which under the null hypothesis has an F distribution with $k - 1$ and $n - k$ degrees of freedom. The RVM model is this case will be identical with the exception that

$$\widetilde{\text{MSS}}\text{(within varieties)} \equiv \frac{(n-k)\text{MSS(within varieties)} + 2b^{-1}}{(n-k) + 2a} \tag{15}$$

replaces MSS(within varieties), and the degrees of freedom for the denominator of the F statistic becomes $n - k + 2a$, instead of $n - k$, (See supplementary material)

For more complicated ANOVA models, such as Latin square designs, interaction effects, or models involving nested effects, the tests in the case of RVM are identical to those in the classical ANOVA case, with the exception that the residual sum of squares is increased by the amount $2b^{-1}$, and the residual degrees of freedom is increased by $2a$.

5 Estimation of a, b

In the previous sections we assumed complete knowledge of a and b. In practice this will not be the case, and so a, and b must be estimated. Use of estimated a and b in the tests of the previous section will violate the strict frequentist formulation that we have used so far. However, unlike σ_i and β_i, the values a and b are constant across genes, and so it is possible to use the data from all of genes to form an estimate. This means that our estimates of these parameters can be based on thousands of data points and can in principal be extremely accurate. Also the RVM tests are not very sensitive to the exact values of a and b. For this reason we can use substitute estimated values for a, and b in the above tests without invalidating them.

To estimate a and b we consider $\widehat{\sigma}_j^2 = \widehat{SS}_j/(n-k)$ the empirical estimates of σ_j^2. It can be shown (see supplementary material) that under the assumptions of RVM

$$ab\left(\widehat{\sigma}_j^2\right) \sim F_{(n-k),2a} \tag{16}$$

Therefore we can estimate the parameters a and b by fitting an F distribution to the observed $\widehat{\sigma}_j^2$.

Although computationally simpler, we recommend against using a method of moments estimator. The higher moments of the F distribution are infinite, and so such an estimator will be unstable. Instead we recommend numerically maximizing the likelihood under the F distribution with respect to a, and b. This provided much

more accurate estimates for a and b, and while computationally intensive, was not overly so, since this estimation needs to be performed only once for the entire data set.

6 Simulation results

To evaluate the tests presented, we generated 2000 simulated data sets, of the expressions of 6000 genes from 2 groups of 5 For each gene, a random σ_j^2 value was chosen from an inverse gamma distribution with $a = 3$, $b = 1$. The σ_j^2 values were chosen separately for each data set. Independent normally distributed random numbers with mean 0 and variances were then assigned as gene expression values. We chose the values $a = 3$, $b = 1$ because estimates of a and b in actual microarray data were found to be in the vicinity of these values. For 3000 of these genes, an amount between 0.1 and 2.0 was added to the expression of samples from the first group (150 genes at every 0.1 units between 0.1 and 2.0). These 3000 genes would represent genes that were truly differentially expressed between the two groups. The remaining 3000 genes would represent non-differentially expressed genes. This procedure was then repeated for an additional 2000 data sets, with 10 samples in each set.

For each gene in each data set, we tested for mean differences between sample groups according to 3 different one-sided t-tests. For the first t-test, a pooled variance estimate across the 6000 genes included in the data set was used, and the t statistic

was assumed to have infinite degrees of freedom. For the second t-test, the variance estimate was made separately within each gene, and 8 degrees of freedom were used. For the final t-test we estimated a and b from the observed residual variances for the data set, and then used the modified t-test presented in equation 13.

Table 1 shows the mean and variance over the 4000 simulations of the values estimated for a and b,first by method of moments and then by Maximum likelihood. We observe that the Maximum likelihood estimates were unbiased, and had low variance, as opposed to the Method of moments estimates which were highly variable, even when based on a large number of genes. We observed that there was a strong inverse correlation between our estimates of a and of b, .such that ab^{-1} was estimated with extreme accurately In light of equation (11), the estimation of b is important only in so far as ab^{-1} is estimated accurately. Therefor any error in our statistic through inaccurate estimation of a and b will really be through the mis-estimation of a.

Table 2, shows the proportion of false positives, (i.e. non-differentially expressed genes that the test declared to be differentially expressed) for each of the tests at different levels. We observe that both the RVM test, and the t-test with the variances estimated within individual genes, appear to have the proper false positive rates. However, using a single value for all variances results in an exesive false-postitive rate. This is due to the fact that highly variable genes, may exhibit large fold differences even when they are not actually differentially expressed. When these fold

differences are then divided by the relatively small pooled residual variance estimate, they will appear significant.

Figure 1 shows the power for detecting true differences between groups according to RVM t-test, and the standard t-test in which the variances were computed individually for each gene. The X axis represents the true difference between the means of the two groups, and the Y axis shows the proportion of such genes that were found to be significant at P=0.01 and P=0.001. Power estimates for the test in which the variance was pooled across genes were not included since this test failed to control type one error and so was not comparable to the other tests. We observe that the RVM t-test generally performed better than the standard t-test, particularly when there was a large difference between the two groups, and a very small p-value was required for significance. The difference was less prominent when there were 10 samples in each group than when there were 5 samples in each group. This is due to the fact that as sample size increases, the RVM test statistic approaches the standard t-statistic.

7 Experimental results

Although the RVM test performed well in simulations in which the model assumptions held exactly, we needed determine how well the model assumptions hold in actual

data. To check this, we looked at two different sets of data. The first set (referred to as the DLBCL data set) came from Rosenwald et. al. (2002) and included 7399 genes measured on 274 Lymphoma samples which were divided into three groups (ABC, GCB, Type 3) according to their gene expression values. The second set (referred to as the BRCA data set) from Hedenfalk et. al. (2001) included 3226 genes measured on 22 breast cancer samples, which were also divided into 3 classes (Sporadic, BRCA-1 and BRCA-2) according to their mutational status.

The primary assumption made by the RVM model is that the within group variance is distributed according to an inverse gamma distribution. This choice of distribution was made purely for computational convenience, and there is no intrinsic reason why the gene variances should follow this distribution. Since we are unable to observe the true within group variances directly to determine whether they followed a inverse gamma, we instead observed the sample variances to determine the extent to which they could be fitted to the F distribution described in equation (16) as should be the case if the true variances followed the inverse gamma distribution. We found that in both data sets, the distribution of the observed residuals was virtually indistinguishable from the corresponding F distributions with fitted parameters, (Fig 2) Implied by our model assumptions. Other data sets have also been investigated, (data not shown) and in each case the F distribution appears to be a very good fit to the observed residuals.

In order to check the relative type 1 and type 2 errors of the various tests we restricted ourselves to the DLBCL data, and in particular the 83 ABC patients and 134 GCB patients. With such a large sample size we can determine with great accuracy which genes were truly differentially expressed between these two groups, and the size of the difference. By selecting small subsets of these groups, we can determine the extent to which the various methods can detect the differentially expressed genes in small samples, while controlling the number of non-differentially expressed genes found to be significant. We repeatedly took sub-samples of size 10 (5 GCB, 5 ABC) and 20 (10 GCB, 10 ABC) an calculated for each gene the p-value for a the difference between the samples observed for the 2 sets. This was repeated 2000 times, and the proportion of times the gene was found to be significant at p=$\{0.001, 0.005, 0.01\}$ was recorded.

To avoid having subsets in which no data for a gene was available, we excluded all those that were missing values in more than 20% of the total data. This resulted in 6,027 genes. A t-test was performed on the entire data set to determine those genes that were truly differentially expressed between the two groups. This test resulted in 1,621 genes were found to be significant at the 0.001 level were declared differentially expressed and used to compute power, while 2,916 genes had a p-value greater than 0.05, were declared non-differentially expressed and used to compute the false positive rates. The remaining 1,490 gene were on the border line between

differentially expressed and non-differentially expressed, were not used in any of the power or type 1 error calculations. In order to consider the effect of true fold difference in our ability to detect genes, we calculated for each gene the mean log ratio within the ABC and GCB subtypes on the total data, then used the difference between these two means rounded to the nearest 0.1.

The results of this analysis are similar to what we found in the simulation. The type 1 error of the t-test and random variance tests are very close to the desired values in both the 10 sample and 20 sample subsets, while the constant variance model had a much larger type 1 error rate than was desired, indicating that this test was invalid (Table 3). Again we also observe that the RVM model preforms much better than the t-test in identifying those genes with large true fold differences between classes.

8 Discussion

Analysis of microarray data clearly indicates that all genes are not equally variant within a sample, therefore some method of estimating an individual gene's variability must be taken into account when determining statistical significance. While directly estimating the variance within each gene works well for large samples, when the sample size is small such an estimate can be imprecise resulting in a test of low power. We have proposed a model through which information from the entire set of

genes is used to influence the variance estimate of a single gene, while still allowing the differences in gene variances. This model can be applied to any linear regression framework, and so is very versatile

This model has the advantage that it is very simple to implement, requiring only slight modifications to tests that are already available in most statistical packages. We have demonstrated that the underlying assumption of our model, that the within gene variances are distributed according to an inverse gamma distribution , appear to be correct in actual experimental data. As a result, our model correctly controls type 1 error, in these experimental situations. Our method has greater power than does the standard t-test in detecting genes with large fold differences. Since these are the genes that are most likely to be of biological interest, we feel that this new method provides an easily implemented improvement to standard techniques where genes are analyzed individually.

References:

Baldi P., and Long AD (2001), "A Baysian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes". *Bioinfomatics* 17, 509-519.

Hadenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, Meltzer P, Gusterson B, Esteller M, Kallioniemi O, Borg A and Trent J. (2001) Gene expression

profiles of hereditary breast cancer. *N Engl J Med* 344, 539-548.

Kerr K and Churchill G (2001), "Statistical Design and the design of gene expression microarrays". *Biostatistics* 2: 183-201

Rosenwald A., Wright G., Chan W. C., Connors J. M., Campo E., Fisher R. I., Gascoyne R. D., Muller-Hermelink H. K., Smeland E. B., Giltnane, J. M., et al. (2002). "The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma", *N Engl J Med* 346, 1937-47.

Scheffe H. (1959) *The Analysis of Variance,* John Wiley, New York.