國立交通大學

統計學研究所

碩士論文

利用混合模型對微陣列資料
做分群變異數分析

Cluster ANOVA with Mixtures
(CANOVAM)
for Microarray Data

研究生：柳超毅

指導教授：盧鴻興　博士

中華民國九十三年六月

利用混合模型對微陣列資料
做分群變異數分析

Cluster ANOVA with Mixtures
(CANOVAM)
for Microarray Data

研究生：柳超毅　　Student：Chao-I Liu

指導教授：盧鴻興 博士　　Advisor：Herry Horng-Shing Lu

國立交通大學理學院

統計學研究所

碩士論文

A Thesis
Submitted to Institute of Statistics
College of Science
National Chiao Tung University
in Partial Fulfillment of the Requirements
for the Degree of Master
in
Statistics
June 2004
Hsinchu, Taiwan, Republic of China

中華民國九十三年六月

誌謝

這本論文的完成，首先要感謝我的指導老師盧鴻興老師在這一年多來給我的教導。使我學習了研究的方法和態度，另外也要感謝我的父母給我精神上和經濟上的支持，使我能夠順利的完成學業。在日常生活中志浩、政輝…等 411 研究室的同學們，我渡過快樂的每一天。牛哥、泰賓學長和一起打球的朋友，不但讓我能夠在兩年中愉快的運動，也給了我許多課業和生活上的寶貴意見。最後我要感謝幫我詳細閱讀了我的論文的口試委員陳素雲老師、黃冠華老師和許文御老師，並且給了我許多寶貴的意見。

柳超毅　　　謹誌于
國立交通大學統計研究所
中華民國九十三年六月

中華民國九三年六月利用混合模型對微陣列資料
做分群變異數分析

研究生：柳超毅　　指導教授：盧鴻興　博士

國立交通大學統計學研究所

中文摘要

在微陣列資料分析中使用變異數分析時殘差通常是一個稀疏的分配. 因此, 我們嘗試對微陣列資料使用混合的變異數分析來做模型的建立, 希望使模型中的實驗因子更單純並且讓殘差更有彈性. 在混合模型中的參數使用 EM 演算法來估計具有較低的複雜度和單調收斂的性質. 在混合模型中, 分群的組數是用貝氏資訊法則並且利用主因子分析來選擇組數的初始值. 然後基因在被分組之後, 基因在每一組中的表現可以對多維常態分配的殘差使用簡單的變異數分析來建立模型. 因此, 對分群後的基因統計的估計和推論可以使用傳統的變異數分析, 包含最小平方估計法和 F 檢定. 在提出利用混合模型對微陣列資料做分群變異數分析這個新的建議之後, 基因可以透過簡單的變異數分析更有彈性的被分群. 在實證研究中也驗證了在各種不同的微陣列資料中 CANOVAM 是可行的.

Abstract

Fitted residuals of ANOVA models for microarray data typically follow a sparse distribution. Hence, we are motivated to model microarray data by ANOVA with mixtures to have model simplicity for experimental factors and flexibility for residual sparsely. The parameters in mixtures are estimated by the generalized EM algorithms with low complexity and monotonic convergence. The number of clusters in mixtures is determined by the Bayesian information criterion and the initial estimate is generated by the projection to principal components. Then, genes are clustered so that the expressions of genes in every cluster can be modeled by a simple ANOVA model with a multivariate Gaussian distribution of residuals. Hence, statistical estimation and inference for every cluster of genes will be performed as the classical ANOVA, including least square estimation and F tests. By this new approach of clustered ANOVA with mixtures (CANOVAM), genes are clustered by simple ANOVA models with flexibility. Empirical studies are also investigated, which confirm the practical feasibility of CANOVAM for microarray data in various experiments.

Tables of Contents

Chapter 1. Introduction

Microarray is a high-throughput and powerful technique for revealing the patterns of coordinately regulated genes (Brown and Botstein, 1999). However, microarray data are also notorious for their noises like experimental errors, biological variations, and instrumental offsets. In addition, the number of RNA samples assayed is typically small in comparison to the large number of genes in an array. Therefore, statistical methods are necessary to model uncertainty in microarray data.

Analysis of variance (ANOVA) is a procedure for constructing statistical tests by partitioning the total variance into different sources. ANOVA has been applied for microarray data in a series of studies (Kerr *et al*., 2000, Kerr *et al.*, 2002, Kerr *et al.*, 2002 Chi and Churchill, 2003, Dudoit *et al.*, 2003). A recent review of ANOVA methods for testing differential expression of genes in microarray experiments is reported in Cui and Churchilll (2003). After background correction, they proposed the ANOVA model in two stages. The first stage uses the following model that does not involve the effects related to gene-specific effects:

$$y_{ijgr} = \mu + A_i + D_j + (AD)_{ij} + r_{ijgr}, \tag{1.1}$$

where y_{ijgr} is the logarithm of signal intensity. The indices represent array (i), dye(j), gene (g) and measurement (r). The notation μ is the overall mean expression level; (A) is the effect of the array on the measured intensity; (D) is the effect of the dye on the measured intensity; (AD) is a term accounting for effects of the interaction between the array and the dye; and (γ_{ijgr}) is the residual. The constraints are that the sum of every effect is zero to avoid the problem of identification similar to those in Appendix 2. In the second stage, gene-specific effects are modeled in terms of the residuals (γ_{ijgr}) of the first stage model of (1.1). The gene-specific model is:

$$r_{ijr} . = G + (VG)_{ij} + (DG)_{j} + (AG)_{i} + \varepsilon_{ijr} . \tag{1.2}$$

In this stage, *(G)* is the average intensity associated with a particular gene; *(AG)*$_i$ is the effect of the array on that gene; and *(DG)*$_j$ is the effect of the dye on that gene. The error term *(*ε_{ijr}*)* is assumed to be independently and identically distributed with mean 0 and a common variance. The variety-by-gene effect *(VG)* is the term that is of primary interest in microarray analysis. This two-stage specification of the model was proposed by Wolfinger *et al.* (2001). For Affymetrix data, the model will be different. In particular, there are no dye effects and there are probe sets with perfect matched and mismatched pairs for Affymetrix data.

For fixed-effects ANOVA, hypothesis testing involves the comparison of two models under the null and alternative hypotheses. In this setting we consider a null hypothesis of non-differential expression (H_0: all *(VG)* values are equal to zero) and an alternative hypothesis with differential expression among treatment conditions (H_1: at least one *(VG)* value is not equal to zero). We can compute F statistics via gene-by-gene basis from the residual sum of squares (RSS):

$$F = \frac{(rss_0 - rss_1)/(df_0 - df_1)}{rss_1 / df_1} . \tag{1.3}$$

Where rss_0 and df_0 are the residual sum of squares and degrees of freedom for the null model (or hypothesis) respectively. Similarly, rss_1 and df_1, are the residual sum of squares and degree of freedom for the alternative model (or hypothesis) respectively. But this F statistic may not follow a standard F distribution because the distributional assumptions of normality may fail in practice. Hence, it is necessary to establish the inference of F statistic by nonparametric approaches, like permutation and bootstrap tests. However, permutation tests will have computational difficulty for a large number of disturbances and bootstrap tests will involve more computational costs (Efron and Tibshirani, 1993, Good, 2000).

In this study, we observe that residuals of the first-stage ANOVA models may not come from a simple distribution. On the contrary, it may come from a mixture distribution, like a normal mixture with different means and variances. So we cluster the residuals by a mixture model with parameters estimated from the microarray data automatically. After clustering, we use the second-stage ANOVA models to estimate gene-specific effects and treatment-by-gene interactions. Consequently, gene selection is made by hypothesis tests with the cluster ANOVA with mixture (CANOVAM). We can use traditional F tests when the normality assumption holds. Otherwise, permutation and bootstrap tests can be applied. This approach can be combined with linear models in literature or new linear models. By CANOVAM, we can have more convenient and fast approaches with traditional statistic tools.

In simulation studies, we simulate a simple dye-swap experiment to identify genes with differential expression by CANOVAM under different situations. In empirical studies, we use microarray data from the microarray core laboratory of Dr. YS Lee in the CGM Hospital. There are 24 arrays in a double loop design and we cluster the differentially expressed genes by CANOVAM. In addition, our methodology of CANOVAM can be applied to all different microarray experiment designs, including common reference designs, loop designs (Kerr *et al.*, 2000), split-plot designs (Tsai and Lee, 2004), and other designs. When the microarray data contain the Affymetrix array or other types of microarray data, we can integrate different models to perform CANOVAM. Once the residuals of fitting models have the same character as sparse distributions, we also apply CANOVAM to cluster genes and then select differentially expressed genes by hypothesis tests within every cluster.

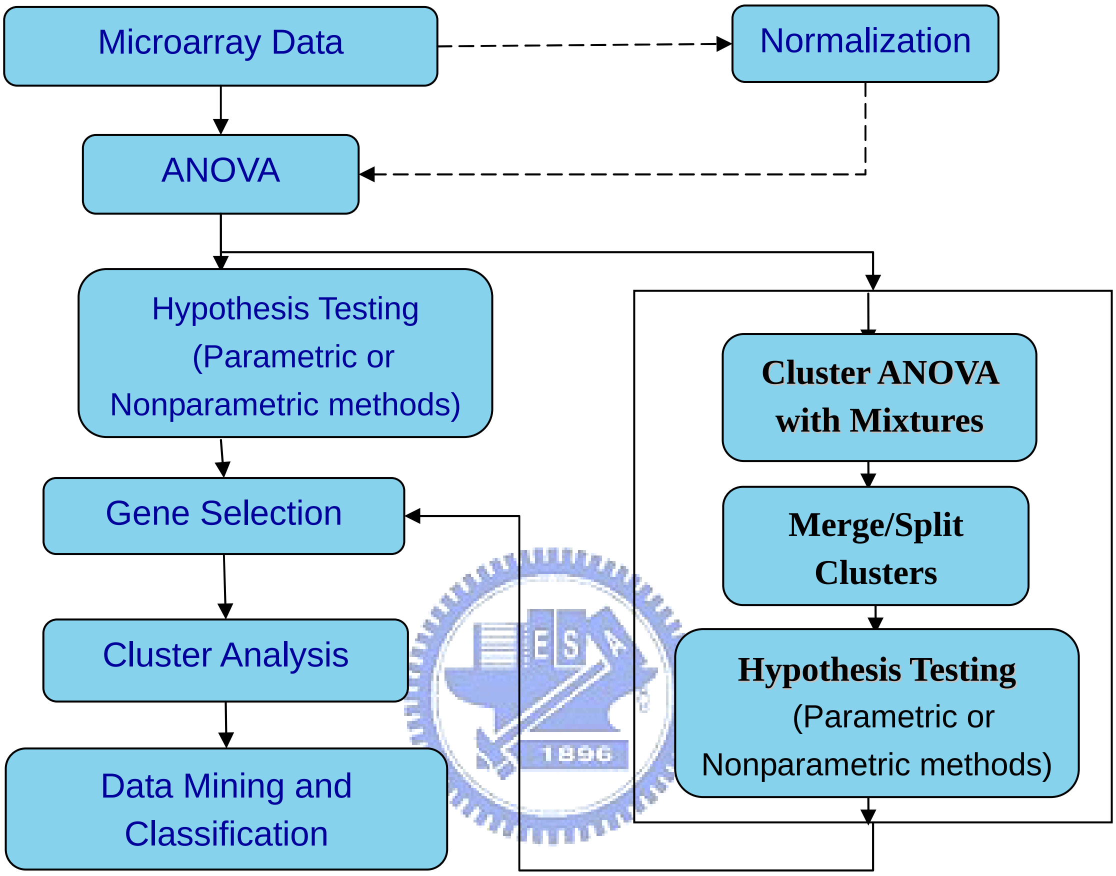
Chapter 2. Methodologies

When the ANOVA model is used to estimate the gene effect, treatment-by-gene interaction and other effects in the analysis of microarray, it is often that the residuals do not fit into a single normal distribution. In particular, the residuals have a sparse distribution with a high peak and two long tails in both sides. Hence, we are motivated to develop a new method that clusters the data with finite mixtures for the residuals. Then, an accurate estimate of variance in every cluster will be obtained and statistical inferences will be made precisely.

Other approaches to model sparse distributions are possible. For instance, the t distribution, the double exponential distribution, Box-Cox transformation, a mixture of a normal distribution and a point mass at zero, and others are proposed in literature (Li *et al.*, 2001, Smyth, 2002, Qiu and Hwang, 2003). However, the estimation and inference procedures become complicated and intractable for high dimensional data. Therefore, we will consider the multivariate normal mixture with the simple EM algorithm for microarray data in high dimension. Consequently, the large amount of genes will be grouped into clusters. The residuals in every cluster have similar variances and they are different between clusters. For every cluster of genes, a simple normal distribution will be used and the statistical inference becomes tractable. In addition, the clustering structure of genes provides biological insights for verification and discovery.

In this chapter, we will use the EM-algorithm to estimate the parameters of multivariate normal distribution for residuals (McLachlan, Bean, and Peel, 2000). The cluster size will be determined by the Bayesian information criterion (BIC). Then, the maximum discriminate rule is used to classify the genes. The flowchart for the cluster ANOVA with mixture (CANOVAM) is illustrated in Figure 2.1.

Figure 2.1: The flow chart of CANOVAM is displayed.


Microarray Data
Normalization
ANOVA
Hypothesis Testing (Parametric or Nonparametric methods)
Gene Selection
Cluster Analysis
Data Mining and Classification
Cluster ANOVA with Mixtures
Merge/Split Clusters
Hypothesis Testing (Parametric or Nonparametric methods)

2.1 The EM Algorithm

The EM algorithm is applied to the mixture model by treating the cluster label of every gene (z) as missing data. The procedure has two steps, E (Expectation) and M (Maximization) steps. Let the observed incomplete data be x and the complete data be $y = (x, z)$. Then the joint density function of complete data y is $P(y;\ \Psi) = P(z \mid x;\ \Psi)\ P(x;\ \Psi)$.

Let $\Psi^{(k)}$ be the old value specified for Ψ. In the E-step, one can evaluate the conditional expectation of the log likelihood of the complete data, $\log(L(\Psi);x)$, given the observed x and $\Psi^{(k)}$. Then the conditional expectation of $L(\Psi)$ is denoted as

$$Q(\Psi;\Psi^{(k)}) = E_{\Psi^{(k)}}[\log L(\Psi) \mid x]. \qquad (2.1.1)$$

In the M-step, one can maximize $Q(\Psi;\Psi^{(k)})$ with respect to Ψ over the parameter space to obtain the updated estimate of $\Psi^{(k+1)}$ such that

$$\Psi^{(k+1)} = \arg\max_{\Psi} Q(\Psi;\Psi^{(k)}). \qquad (2.1.2)$$

The E-step and M-step is repeated until convergence. It is proven that each iteration will increase the log-likelihood of the incomplete data and the EM algorithm will converge to a local maximum monotonically under regular conditions (Dempster, Laird, and Rubin, 1977, Wu, 1983).

2.2 Finite Mixtures

We consider microarray data X={x_1, …, x_N} as a set of multi-dimensional data. Each x_j corresponds to the expression in a variety of arrays for the jth gene. The mixture model for M clusters is defined as

$$p(x;\Psi) = \sum_{m=1}^{M} \pi_m P_m(x;\theta_m), \qquad (2.2.1)$$

where π_m is the mixing proportion of the mth cluster,

$$0 < \pi_m < 1,$$

and

$$\sum_{m=1}^{M} \pi_m = 1. \qquad (2.2.2)$$

The probability density function $P_m(x;\ \theta_m)$ in the mth cluster has the parameter vector θ_m and the entire parameter vector is $\Psi = (\pi_1, \ldots, \pi_M, \theta_1, \ldots, \theta_M)$.

The log-likelihood of incomplete data becomes

$$\log(L(\Psi \mid x)) = \sum_{j=1}^{N} \log(\sum_{m=1}^{M} \pi_m P_m(x_j;\theta_m)) \,. \qquad (2.2.3)$$

The maximum likelihood estimate needs to solve the system of partial differential equations of $\partial \log(L(\Psi \mid x))/\partial\Psi = 0$, which is intractable because the complicated structure of summation inside the log function. Hence, one can introduce the unobserved variable of the cluster label for the observed data x_j as follows:

$$Z_{jm} = \begin{cases} 1, & \textit{when } x_j \textit{ is from the m-th cluster;} \\ 0, & \textit{otherwise.} \end{cases} \tag{2.2.4}$$

Then, the complete data log likelihood for x_j becomes

$$\begin{aligned} \log L(\Psi) &= \sum_{j=1}^{N} \log(\pi_m P_{m_i}(x_j;\theta_{m_i})) \\ &= \sum_{j=1}^{N}\sum_{m=1}^{M} z_{im}(\log \pi_m P_m(x_j;\theta_m)) \\ &= \sum_{j=1}^{N}\sum_{m=1}^{M} z_{im}(\log \pi_m + \log P_m(x_j;\theta_m)). \end{aligned} \tag{2.2.5}$$

E – Step

In E-step, the mixing parameter π_m can be thought as the prior probability of each mixture component. By the Bayes Rule, the posterior probability that x_j belongs to the *m*th cluster of the mixture becomes

$$\begin{aligned} P(m \mid x_j;\Psi^{(k)}) &= \frac{\Pr(x_j, m;\Psi^{(k)})}{\Pr(x_j;\Psi^{(k)})} \\ &= \frac{\pi_m P_m(x_j;\theta_\mathrm{m}{}^{(k)})}{\sum_{m=1}^{M} \pi_m P_m(x_j;\theta_\mathrm{m}{}^{(k)})}. \end{aligned} \tag{2.2.6}$$

Also, the conditional expectation of *logL(Ψ)*|*x* is

$$\begin{aligned} Q(\Psi;\Psi^{(k)}) &= E_{\Psi^{(K)}}[\log L(\Psi) \mid x] \\ &= E_{\Psi^{(K)}}[\sum_{j=1}^{N}\sum_{m=1}^{M} z_{jm}(\log \pi_m + \log P_m(x_j;\theta_m)) \mid x] \\ &= \sum_{j=1}^{N}\sum_{m=1}^{M} E_{\Psi^{(K)}}[z_{jm} \mid x](\log \pi_m + \log P_m(x_j;\theta_m))] \\ &= \sum_{j=1}^{N}\sum_{m=1}^{M} P(m \mid x_j;\Psi^{(k)})(\log \pi_m + \log P_m(x_j;\theta_m))]. \end{aligned} \tag{2.2.7}$$

M-Step

In M-step, one can maximize $Q(\Psi;\Psi^{(k)})$ in (2.2.7). To estimate $\pi_{m,}$ with the constraint that $\sum_{m=1}^{M}\pi_m = 1$, one can differentiate this function and the constraint as follows:

$$Q_\lambda(\Psi;\Psi^{(k)}) = Q\ (\Psi;\Psi^{(k)}) + \lambda(\sum_{m=1}^{M}\pi_m - 1), \qquad (2.2.8)$$

where λ is a Lagrange multiplier. Therefore, the estimates of the mixture proportions turn out to be

$$\pi_m = \frac{1}{N}\sum_{j=1}^{N} P(m \mid x_j;\Psi^{(k)}). \qquad (2.2.9)$$

To estimate Ψ for the new update of $\Psi^{(k+1)}$, one can solve the following equation:

$$\sum_{j=1}^{N}\sum_{m=1}^{M} P(m \mid x_j;\Psi^{(k)})\frac{\partial \log\pi_m + \log P_m(x_j;\theta_m)}{\partial\Psi} = 0. \qquad (2.2.10)$$

By writing down with each proportion in (3.2.9), one can derive the estimate of the parameter Ψ and obtain the new estimate of $\Psi^{(k+1)}$ by solving the (2.2.10) equation.

2.3 Multivariate Normal Mixtures

In the normal mixture, the d-dimensional normal distribution density function and its log transform become

$$P_m(x;\mu_m,\Sigma_m) = (2\pi)^{-d/2}\,|\Sigma|^{-1/2}\,\exp[-\frac{1}{2}(x-\mu_m)^T\Sigma_m^{-1}(x-\mu_m)], \qquad (2.3.1)$$

$$\log P_m(x;\mu_m,\Sigma_m) = -\frac{d}{2}\log(2\pi) - \frac{1}{2}\log(|\Sigma|) - \frac{1}{2}(x-\mu_m)^T\Sigma_m^{-1}(x-\mu_m), \qquad (2.3.2)$$

where $\mu_m = (\mu_m{}^{(1)}, \mu_m{}^{(2)}, \ldots, \mu_m{}^{(d)})^T, m = 1,2,\ldots,M$. We sort the means by the increasing order in each coordinate to avoid the identifiability problem and the equality rarely happens for numerical values of mean estimates. Let Z_{jm} be unobserved data by (2.2.4), then the complete log-likelihood in (2.2.5) can be formulated.

In E-step, one can write down the posterior probability that x_j belongs to the mth cluster of the normal mixture by Bayes Rule,

$$P(m \mid x_j; \Psi^{(k)}) = \frac{\pi_m^{(k)} P_m(x_j\,; \mu_m^{(k)}, \Sigma_m^{(k)})}{\sum_{m=1}^{M} \pi_m^{(k)} P_m(x_j\,; \mu_m^{(k)}, \Sigma_m^{(k)})}, \tag{2.3.3}$$

and the conditional expectation of *logL(Ψ)|x* becomes

$$Q(\Psi;\Psi^{(k)}) \;=\sum_{j=1}^{N}\sum_{m=1}^{M} P(m \mid x_i;\Psi^{(k)})(\log \pi_m^{(k)} + \log P_m(x_j\,; \mu_m^{(k)}, \Sigma_m^{(k)}))]. \tag{2.3.4}$$

In M-step, the mixture proportion turns out to be

$$\pi_m^{(k+1)} = \frac{1}{N}\sum_{j=1}^{N} P(m \mid x_j; \Psi^{(k)}). \tag{2.3.5}$$

Also, the estimate for the mean μ_m is

$$\mu_m^{(k+1)} = \frac{\sum_{j=1}^{N} P(m \mid x_j\,; \Psi^{(k)}) x_j}{\sum_{j=1}^{N} P(m \mid x_j\,; \Psi^{(k)})}, \tag{2.3.6}$$

and the variance-covariance matrix $\sum_m$ is

$$\Sigma_m^{(k+1)} = \frac{\sum_{\mathrm{j=1}}^{\mathrm{N}} P(m \mid x_j; \Psi^{(k)})(x_j - \mu_m^{(k+1)})(x_j - \mu_m^{(k+1)})^T}{\sum_{\mathrm{j=1}}^{\mathrm{N}} P(m \mid x_j; \Psi^{(k)})}. \tag{2.3.7}$$

We iterate the E-step and M-step until $\log L(\Psi^{(k+1)}) \mid x - \log L(\Psi^{(k)}) \mid x < \varepsilon$ for a positive tolerance ε.

2.4 Maximum Likelihood Discriminant Rule

We assume that the expressions of genes come from different clusters and every cluster has the same character of distribution but different parameters. The next question is how should we allocate N genes to M clusters? By the model of finite normal mixtures, we will use the EM algorithm to estimate the parameters in each cluster. The likelihood function for every cluster provides the discrimination function to cluster genes. If the expression of one gene has the largest likelihood in a particular cluster, then this gene shall be clustered into that cluster. We consider

$p(x;\Psi) = \sum_{m=1}^{M} \pi_m P_m(x;\theta_m)$, and π_m is the mixture proportion of *m*th cluster. We allocate the data x_j to the *m*th cluster if

$$P(m^* \mid x_j;\Psi) = \max_m P(m \mid x_j;\Psi), \quad m = 1,\ldots,\mathrm{M}. \qquad (2.4.1)$$

After clustering, we can use classical ANOVA model and F test to select genes in every cluster. After clustering, we assume the residuals in each cluster follow a normal distribution with mean 0 and variance σ. Under the null hypothesis that the gene *g* is not differentially expressed across the treatment conditions and the normality holds, the test statistics of $F^* = MST_g / MSE_m$ follow an *F* distribution. The term of MST_g denotes the mean square for treatment conditions and that of MSE_m denotes the variance of the error term in the *m*th cluster. Besides, we can also fit each cluster with new ANOVA models. In this situation, it is assumed that different gene in the same array has the same array effect and dye effect. Otherwise, we can also re-estimate the ANOVA model of each gene after clustering to rectify the system offsets.

2.5 How Many Clusters?

We now investigate the determination of cluster size. Firstly, is it necessary to cluster the genes? After the first-stage of ANOVA, we can test the normality of residuals by Kolmogorov-Smirnov tests or other tests. If the normality test is passed, then it is not necessary to perform cluster ANOVA. In most cases, the normality test of the residuals after the first-stage ANOVA fails. Then, we need to consider the approach of CANOVAM. We can use the Bayesian information criterion (BIC, Schwartz, 1978) to choose a cluster number.

For a specific cluster size, the likelihood of the maximum likelihood estimates of parameters measures the goodness-of-fit for the residuals. This goodness-of-fit will increase as the number of parameters increases. In order to avoid the problem of over-fitting, a penalty for model complexity related to the number of parameters shall be included to balance these two factors in a criterion. Then, a suitable model with a proper number of parameters can be selected based on the criterion.

Both the Akaike information criterion (AIC) and Bayesian Information criterion (BIC) are common used for model selection in literature (Burnham and Anderson,1998). With a minus sign, the maximization of the penalized likelihood is equivalent to the minimization of AIC and BIC as follows:

$$AIC = -2\log(L_{ML}) + 2K_a, \tag{2.5.1}$$

$$BIC = -2\log(L_{ML}) + K_a \log N, \tag{2.5.2}$$

where N is the total number of observations, and K_a is the total number of free parameters in the finite mixture model.

For pair-wise comparisons of two nested models, AIC and BIC are equivalent to the likelihood–ratio test (Akaike, 1973). That is, we consider the null and alternative hypotheses as follows: H_0: a small model of M_1 is sufficient *vs.* H_1: a large model of M_2 that contains M_1 is sufficient. The significance level of BIC, P(accept M_2 | M_1 is true), is approaching 0 as $N\rightarrow\infty$. But the significance level of AIC does not approach 0 asymptotically. Therefore, BIC is a better method asymptotically. When the sample size is small (*N=7.389*), AIC and BIC are the same. In microarray data analysis, large sample size is very large because of a large number of genes is used. Then, the penalty of BIC is larger than that of AIC. Hence, model selection by BIC for microarray data will select a simpler model that by AIC.

Chapter 3. Simulation and Empirical Studies

3.1 Simulation Studies

In this chapter, we simulate that the residuals of microarray data have two possible kinds of distributions. Firstly, we use CANOVAM to perform the gene selection when the residuals follow a normal mixture model. We will investigate the improvements of CANOVAM in comparison to those of ANOVA. Secondly, we simulate the cases that the residuals follow a heavy distribution like the t distribution. We will study the performance of CANOVAM in this situation.

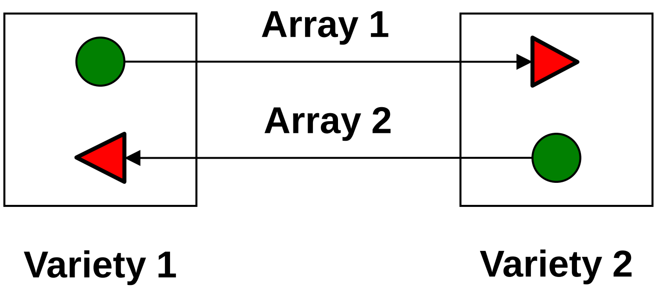
3.1.1 Residuals with Normal Mixture Distributions

A simple dye-swap experiment is illustrated in Figure 3.1.1. Every arrow in the figure represents a microarray chip. The variety in the dotted side of an array is labeled as Cy3 or G and the variety in the arrowhead side of an array is labeled as Cy5 or R. A simple ANOVA model is used for the gene expressions of *1000* genes, $g = 1, \ldots, 1000$, with samples coming from two treatments, $k = 1, 2$, and two dyes, $j = 1, 2$, on two arrays, $i = 1, 2,$ as follows:

$$y_{ijkg} = \mu + A_i + D_j + AD_{ij} + G_g + VG_{kg} + \varepsilon_{ijkg}\,. \qquad (3.1.1)$$

The constraints are that the sum of every effect is zero to avoid the problem of identification as in Appendix 2. Here, the treatment effects are replaced by the array and dye interaction terms (Wolfinger *et al.,* 2001) as the simulated values in Table 3.1.1 of the Appendix 1.

Figure 3.1.1: A simple dye-swap experiment is illustrated.



Firstly, we simulate the case that the residuals follow a standard normal distribution with mean 0 and variance 1. The BIC will select the cluster size of 1 and the CANOVAM is equivalent to the classical ANOVA. Then, we will simulate the case that the residual ε comes from a distribution of two normal mixtures:

$$f(\varepsilon \mid \Psi) = \pi_1 \mathrm{N}(\mu_1, \sigma_1^2) + \pi_2 \mathrm{N}(\mu_2, \sigma_2^2)\,. \tag{3.1.2}$$

That is, we assume that the residuals have three clusters with small and large variations which mimic the sparse distribution for microarray data in practice. The simulated values of parameters are reported in Table 3.1.2 of the Appendix 1.

Moreover, we will simulate the case that the residual ε comes from a distribution of three normal mixtures:

$$f(\varepsilon \mid \Psi) = \pi_1 \mathrm{N}(\mu_1, \sigma_1^2) + \pi_2 \mathrm{N}(\mu_2, \sigma_2^2) + \pi_3 \mathrm{N}(\mu_3, \sigma_3^2)\,. \tag{3.1.3}$$

That is, we assume that the residuals have three clusters with small, medium, and large variations which mimic the sparse distribution for microarray data in practice. The simulated values of parameters are reported in Table 3.1.4 of the Appendix 1.

We demonstrate the difference in density plots of the three normal mixture and a normal distribution with the same mean and variance in Figure 3.1.2. From the density plots, the mixture distribution has a sparse distribution with a high peak in the middle and two long tails in both sides that mimics the distribution of ANOVA residuals for microarray data. The density plots of three normal distributions in the normal mixtures in Table 3.1.4 are illustrated in Figure 3.1.3 of the Appendix 1.

The simulation studies generate different data sets with various percentages of significant genes. If there are only a few or a half of genes are significant, we can simulate these cases with only 5% or 50% significant genes respectively. Then we can use different statistics to select significant genes by ANOVA or CANOVAM. The match percentages of the selected genes are evaluated for comparison studies.

In the process of gene selection, we consider the classic *F* statistic and *F*-like statistics (Cui X *et al.,* 2003). For split plot designs, we also consider the interquartile range method for gene selection (Tsai and Lee, 2004). The following notations will be used.

Let MST_g denote the mean squares of relative expression levels of one gene in multiple samples. Variance components $\hat{\sigma}_g^2$ in F_1 are estimated form the expressions of one gene. F_2 and F_3 statistics are proposed by Cui and Churchill (2003). The statistics F_3 uses the pooled variance estimator, $\hat{\sigma}_{pool}^2$, for each variance component and F_2 uses the average of $\hat{\sigma}_{pool}^2$ and $\hat{\sigma}_g^2$ for each component. The statistics F_S uses the shrinkage estimator based on $\tilde{\sigma}_g^2$. The statistics C_g is based on interquartile range method in Tsai and Lee (2004). Under the null hypothesis that gene *g* is not differentially expressed among the treatment conditions, this statistics should be distributed approximately as chi-square distribution with df_T and df_T denoted the degrees of freedom. Let *Median(MST)* denoted the median of the MST_g values. Then, the statistics are defined as:

$$
\begin{aligned}
&F_1 = MST_g / \hat{\sigma}_g^2, \\
&F_2 = MST_g / \frac{1}{2}(\hat{\sigma}_g^2 + \hat{\sigma}_{pool}^2), \\
&F_3 = MST_g / \hat{\sigma}_{pool}^2, \qquad\qquad\qquad\qquad\qquad\qquad (3.1.3) \\
&F_S = MST_g / \tilde{\sigma}_g^2, \\
&C_g = \chi_{df_T}^2(0.5)\frac{MST_g}{Median(MST)}.
\end{aligned}
$$

In Table 3.1.3 and 3.1.5, the results of ANOVA and CANOVAM with five different statistics are reported. In this table, the match number represents the number of truly significant genes that are selected. If the match number is bigger, then the correctness of gene selection is higher. We will consider the top 5%, 10% and 50% selected genes which are chosen by the five statistics with ANOVA or CANOVAM to evaluate the correctness of these approaches.

It is noted that the *F1* statistics has no any difference before and after clustering. Hence, we evaluate the capability of the other four statistics only for CANOVAM. From the results of Table 3.1.3 and 3.1.5, it is found that the correctness of these four statistics increases under the condition of choosing the same amount of genes in CANOVAM with the correct cluster size when compared with the results of ANOVA.

We have to decide the number of clusters by BIC before clustering. If we select a wrong cluster number, then the selected number of clusters may be more or less than the correct size. Under this situation, can we get better results by CANOVAM than ANOVA? In stead of the correct three clusters, we use the number of clusters of two clusters and four clusters in Table 3.1.6. The results of matched genes in CANOVAM with incorrect cluster sizes are still better than those of ANOVA without clustering in these studies. That is, the estimate of variances from a group of genes from neighboring clusters can still improve the statistical inferences by borrowing the strength from the expressions of similar genes.

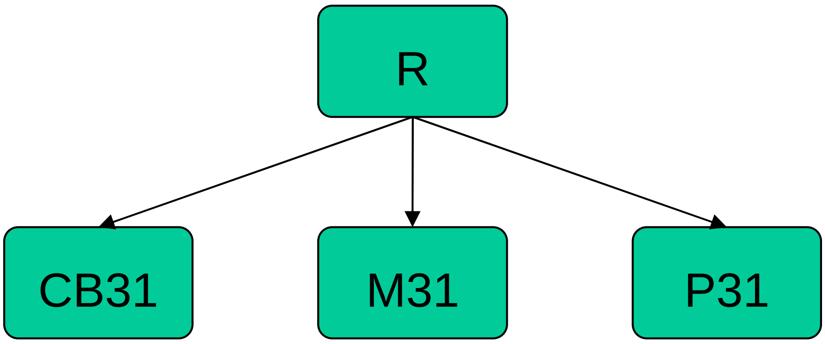
3.1.2 Residuals with t Distributions

When the distributions of residuals are other types of distributions that are not mixture distributions, we will investigate the performances of CANOVAM in these situations by simulation studies. In simulations, we will assume the residuals come from a student's *t* distribution with *df* = 5. We use the BIC to choose the number of cluster and select two clusters as a result. Then, we separate genes into two clusters by means of CANOVAM. In Table 3.1.7, the results of CANOVAM are reported. The performances of CANOVAM are better than those by ANOVA in this case, which indicates the robustness of CANOVAM when the distributions of residuals are not normal mixtures.

3.2 Empirical Studies with Spike Genes

In empirical studies, we use the microarray data with spike genes generated in the microarray core laboratory by Dr. Yu-Shien Lee at the CGM Hospital in Taiwan. This is a reference design which contains three arrays and three treatments as displayed in Figure 3.2.1.

Figure 3.2.1: The reference design with 3 arrays and 3 treatments is displayed.



In this experiment, there are 256 spike genes in a total of 14924 genes.

Different spike genes have different spotted ratios of Cy5/Cy3 in every array. In Table 3.2.1, there are 8 kinds of spike genes and each kind has 32 replications in an array. The spotted ratios have four levels as summarized in Table 3.2.1.

Table 3.2.1: The spotted ratios of spike genes are reported.

Gene name		**Cy5/Cy3 Ratio**
Spike 1	**Spike 2**	**10:1**
Spike 3	**Spike 4**	**5:1**
Spike 5	**Spike 6**	**2.5:1**
Spike 7	**Spike 8**	**1:1**

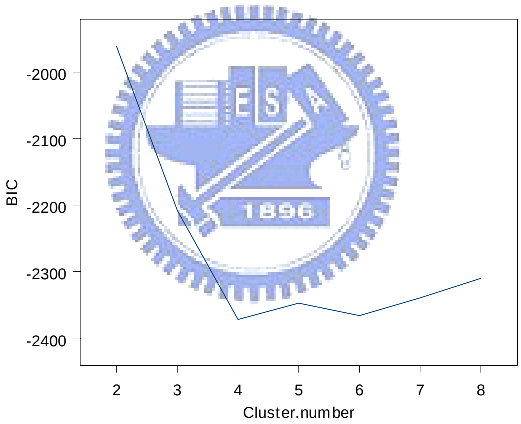
Firstly, we use the log transform of expressions and the following ANOVA model for the gene expressions of 14924 genes, $g = 1, \ldots, 14924$, with samples coming from three treatments, $k = 1, 2, 3,$ and two dyes, $j = 1, 2$, on three arrays $i = 1, 2, 3,$ as follows:

$$log(y_{ijg}) = \mu + A_i + AD_{ij} + D_j + G_g + (VG)_{ijg} + \varepsilon_{ijg}\,. \qquad (3.2.1)$$

In this classical ANOVA, we assume that residuals follow a normal distribution, $\varepsilon_{ijg} \sim N(0, \sigma^2)$. The least square estimate of (VG) effect is derived in Appendix 2, which is also the maximum likelihood estimate under the assumption of normality. We can check the normality assumption of residuals in the data by normality tests like the Kolmogorov-Smirnov test or the chi-square tests (Ross 1997). The null hypothesis is H_0: residuals have the normal distribution; whereas the alterative hypothesis is the opposite of the null hypothesis. The p-value of the Kolmogorov-Smirnov for the residuals of ANOVA in this data is very close to 0 and the null hypothesis is rejected. Therefore, the normality assumption of residuals is rejected in this data and we will also analysis this data by CANOVAM for comparison.

In Figure 3.2.2, it is observed the BIC value of normal mixtures for the residuals in this data has the smallest value when the cluster number is 4. Hence, we cluster the genes into four clusters according the model of normal mixture in this study.

Figure 3.2.2: The BIC values of normal mixtures for the residuals in this data are plotted against the cluster number.



After clustering, we can check the normality of residuals in four clusters by the density plots and normal QQ plots in Figure 3.2.3. The residuals in cluster 1, 3, and 4 in Figure 3.2.3 can be fitted by normal distributions with different parameters. The residuals in cluster 2 have a longer tails in both sides than a normal distribution. Normality tests like Kolmogorov-Smirnov tests can be applied to confirm these observations. We can either cluster the genes in cluster 2 to more small sub-clusters or increase the size of clusters from four to more so that the residuals in every cluster follow a normal distribution. For instance, we can use twelve clusters for this data and the residuals in every cluster follow a normal distribution. But, we will need to estimate more parameters and the model complexity increase. Hence, we will consider four clusters suggested by BIC to balance the effects of model fitting and complexity. Finally, we can check whether the variances in four clusters are the same or not. The result of Bartlett test (Snedecor and Cochran, 1983) is reported in Table 3.2.2. As the p-value of Bartlett test is very small, we can reject the null hypothesis that the variances in four clusters are the same. Hence, we do not merge these four clusters into smaller sizes of clusters. More robust tests, like the Levene tests (Levene 1960), can be applied to test the equality of variances.

Figure3.2.3: The density plots and normal QQ plots for normality checking in four clusters are displayed.

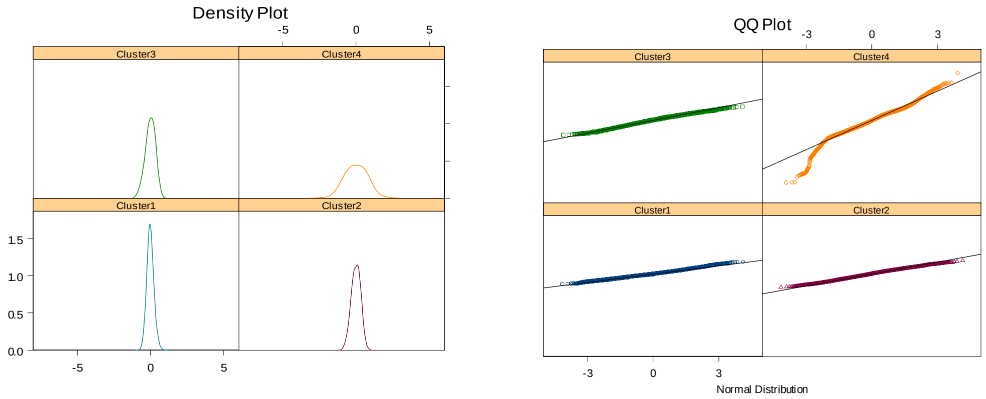


Table 3.2.2: The results of Bartlett test for four clusters are summarized.

Bartlett test			
s_1	0.057844	N_1	27042
s_2	0.101582	N_2	29490
s_3	0.12412	N_3	22416
s_4	0.669071	N_4	10848
s_{pool}	0.16259	N	89796
T	32519.48	P-value	0

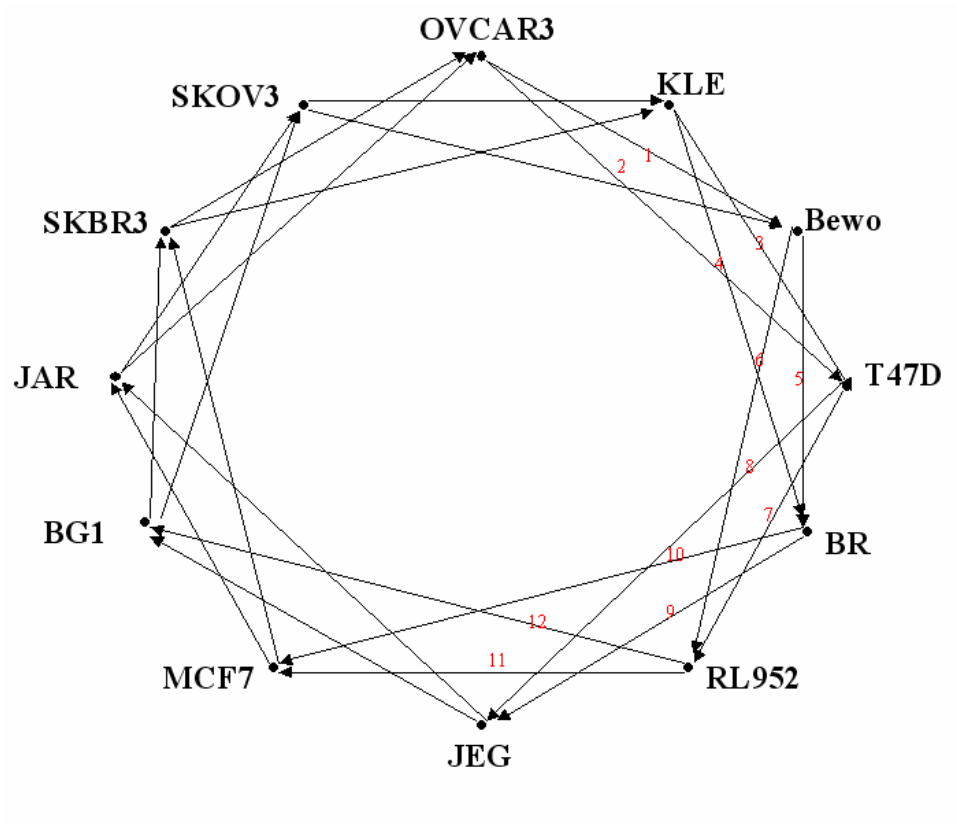
Now, we investigate the performances of ANOVA and CANOVAM in spike genes. By Table 3.2.1, there are two kinds of spike genes, Spike 7 and 8, has the spotted ratios of 1:1 and they are designed to represent the non-differential expressed (or insignificant) genes. The other six kinds of spike genes have spotted ratios that are different from 1:1, which represents the differential expressed (or significant) genes.

The match numbers and percentages of spike genes for ANOVA and CANOVAM are reported in the Appendix 1. In Table 3.2.3, it is observed that the match numbers and percentages for significant and insignificant genes in spike genes are both higher in CANOVAM than those in ANOVA.

3.3 Empirical Studies with 24 Microarrays

One double loop design is shown in Figure 3.3.1 to demonstrate the flexibility of CANOVAM for complicated designs. There are 12 varieties and 24 arrays in this experiment. The capitals of “OVCAR3,” “KLE,” …, and ”SKOV3” denote the types of varieties. The graphic display of arrow is the same as that in Figure 3.3.1.

Figure 3.3.1: The experiment design of one double loop microarray experiment is displayed.



Firstly, we use the log transform of expressions and the following ANOVA model for the gene expressions of 7334 genes, $g = 1, \ldots, 7334$, with samples coming from 12 treatments, $k = 1, \ldots, 12,$ and two dyes, $j = 1, 2$, on 24 arrays $i = 1, \ldots, 24,$ as follows:

$$log(\, y_{ijkg}) = \mu + A_i + D_j + AD_{ij} + G_g + (AG)_{ig} + (DG)_{ig} + (VG)_{kg} + e_{ijkg.}$$

(3.4.1)

In Figure 3.3.2, the histogram in the center is plotted from the fitted residuals and the smooth curve is the density plot with a simple normal distribution with the same mean and variance. It is clear that the residuals have a sparse distribution and we will consider the analysis of CANOVAM. The normal QQ plot of residuals in Figure 3.3.3 confirms this phenomenon as well.

Figure 3.3.2: The iduals are plotted.

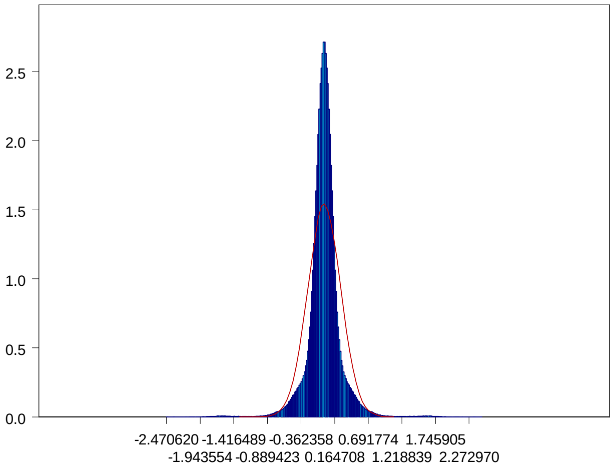
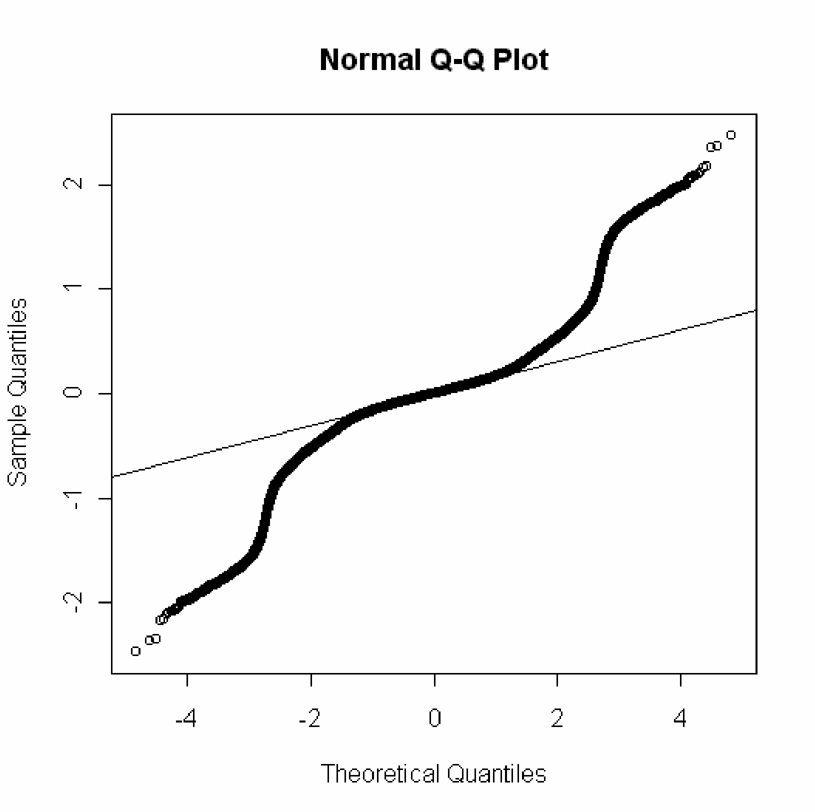


Figure 3.3.3: The normal QQ plot of residuals is displayed.



The BIC values for normal mixtures of these residuals are reported in Table 3.3.1. The minimum of BIC occurs at the cluster size of 15 in this study.

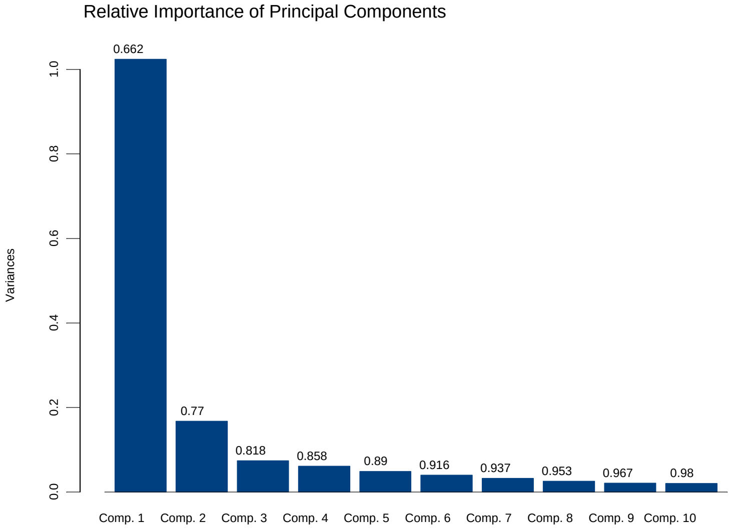
Table 3.3.1: The BIC values for the residuals of 24 microarrays are reported.

cluster	log(L)	BIC
6	1182224	-1179667
7	1190841	-1187819
…	…	…
14	1268331	-1262052
15	**1269584**	**-1262840**
16	1270008	-1262799

For every gene, there are expressions in 24 arrays with 12treatment and 2 dyes in this case. As a result, the dimension of residuals for every gene is 48. For this kind of high dimensional data, dimension reduction techniques are useful to reduce the dimension to search for a start cluster size of normal mixtures by BIC. Principal component analysis (PCA) is a dimension reduction tool that transforms a set of correlated response variable into a small set of uncorrelated variables, which are called principal components (Hotelling, Harold, 1933).

In this study, PCA can transfer the dimension of residuals from 24 to 2 in this study that can explain the most part of variations between varieties and dyes. In Figure 3.3.4, one finds the first principal component explains 66.2% of total variation and the first two principle components explain 77% of total variance. Since the including of the third principal component does not improve too much percentage in explaining total variation, we use the first two principal components to search for a start cluster size of normal mixtures by BIC.

Figure 3.3.4: Relative importance of principal components for 24 microarrays is displayed.



The BIC values for normal mixtures of the first two principal components in these 24 microarrays are listed in Table 3.3.2. The minimum value of BIC occurs at the cluster size of 8 in this study. Because the variation explained by the leading principal components is smaller than 100%, the cluster size selected by the minimum of BIC with the leading principal components is usually smaller than that selected by BIC with the entire data that include all principal components. However, the cluster size selected BIC with PCA provides a good initial start point. We can search the minimum of BIC values for the original data with a high dimension by increasing the cluster size from the start point. Because the computation cost of BIC is less by PCA in low dimension, PCA can be used to provide a good start point for BIC with fast computation time when the original data in high dimensional.

Table 3.3.2: The BIC values for normal mixtures of the first two principal components in 24 microarrays are listed.

	log(L)	BIC
cluster		
6	5362.739	-5281.19
7	5399.468	-5303.53
8	**5423.552**	**-5313.23**
9	5432.17	-5307.46
10	5422.666	-5283.56
11	5427.42	-5273.93

Chapter 4. Conclusion and Discussion

Because of many sources of experiment errors, the residuals of ANOVA models are usually sparse for microarray data. We have proposed the CANOVAM to cluster the residuals of ANOVA by normal mixtures so that the expressions of genes in every cluster can be modeled with a simple ANOVA model with a normal distribution. The selection of significant genes and statistical inferences become tractable with CANOVAM.

The BIC is used to select the cluster size of normal mixtures for residuals. Even the cluster size is selected incorrectly by the BIC, the CANOVAM still outperforms the ANOVA in simulation and empiric studies because the information of similar gene expressions are polled together.

When the residuals is high dimensional for experiments with many arrays, PCA can be applied to reduce the dimension and the computation cost. The computation cost of the normal mixtures with the EM algorithm can be further reduced by the fast versions of generalized EM algorithms that improve the convergence rate of the EM algorithm (Demester, Laird, Rubin 1977).

Other methods of clustering besides normal mixtures can be applied to the model of cluster ANOVA as well. Integration of cluster ANOVA with different normalization methods is also feasible. In addition, we are highly interesting in applying CANOVAM to Affymetrix microarrays in future studies.

Appendix 1

Table 3.1.1: Simulated values of array and dye effects in the ANOVA model are listed.

	High level	Low level
Array effect	1	-1
Dye effect	0.5	-0.5
AD1 effect	1	0
AD2 effect	0	-1

Table 3.1.2: Simulated values of two normal mixtures are reported.

Parameter	Cluster 1	Cluster 2
Mixture proportion	0.5	0.5
Mean	0	0
Variance	1	100

Table 3.1.3: Simulation results of ANOVA and CANOVAM for two normal mixtures with the correct cluster size are reported. Here, $CV =| VG_{ig} - VG_{jg} | / \sqrt{\sigma_1^2 + \sigma_2^2}$. The number of differentially expressed (or significant) genes is 5% of all 1000 genes

in simulations. The match number of significant genes in the top 5%, 10%, and 50% selected genes by different statistics of *F1, F2, F3, Fs,* and *Cg*. Note that the results of *F1* are the same for ANOVA and CANOVAM since the individual variance for every gene remains the same after clustering.

Significant genes: 5%		ANOVA			
CV=2	F1	F2	F3	Fs	Cg
Match number in top 5%	13	23	23	18	23
Match percentage (%)	26.00%	46.00%	46.00%	36.00%	46.00%
Match number in top 10%	29	27	25	23	25
Match percentage (%)	58.00%	54.00%	50.00%	46.00%	50.00%
Match number in top 50%	49	49	49	45	49
Match percentage (%)	98.00%	98.00%	98.00%	90.00%	98.00%

Significant genes = 5%		CANOVAM (Cluster size = 2)		
CV=2	F2	F3	Fs	Cg
Significant gnens (5%)	34	34	30	33
Percentage (%)	68.00%	68.00%	60.00%	66.00%
Significant genes (10%)	45	43	41	45
Percentage (%)	90.00%	86.00%	82.00%	90.00%
Significant genes (50%)	50	50	50	50
Percentage (%)	100.00%	100.00%	100.00%	100.00%

Table 3.1.4: Simulated values of three normal mixtures are reported.

Parameter	Cluster 1	Cluster 2	Cluster 3
Mixture proportion	0.5	0.25	0.25
Mean	0	0	0
Variance	1	25	100

Figure 3.1.2: The density plot of a sparse distribution by three normal mixtures with the simulated values in Table 3.1.2 and a normal distribution with the same mean as well as variance are displayed for comparison.

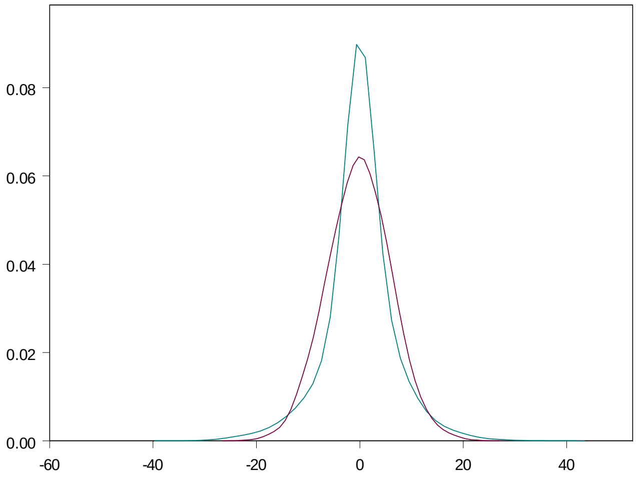


Figure 3.1.3: Density plots of three normal distributions in a normal mixture are illustrated with the simulated values in Table 3.1.4.

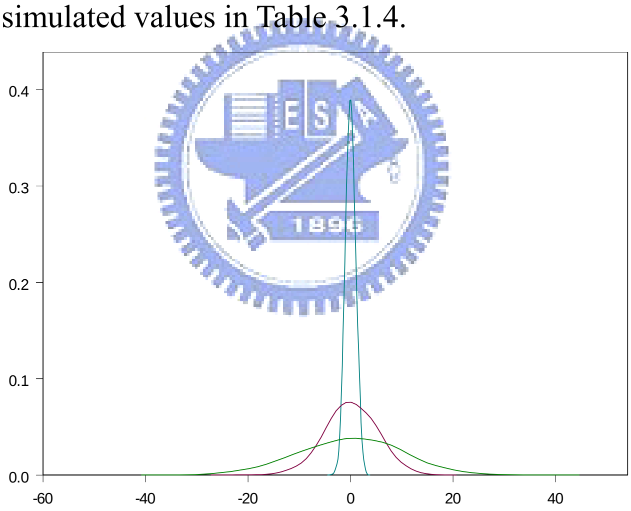


Table 3.1.5: Simulation results of ANOVA and CANOVAM for three normal mixtures with the correct cluster size are reported. The number of differentially expressed (or significant) genes is 5% or 50% of all 1000 genes in simulations.

(a) Significant genes: 5%

Significant genes = 5%		ANOVA			
CV=2	F1	F2	F3	Fs	Cg
Match number in top 5%	18	22	22	19	22
Match percentage (%)	36.00%	44.00%	44.00%	38.00%	44.00%
Match number in top 10%	24	25	24	20	24
Match percentage (%)	48.00%	50.00%	48.00%	40.00%	48.00%
Match number in top 50%	50	50	50	43	50
Match percentage (%)	100.00%	100.00%	100.00%	86.00%	100.00%

Significant genes = 5%		CANOVAM (Cluster size = 3)		
CV=2	F2	F3	Fs	Cg
Match number in top 5%	26	27	22	27
Match percentage (%)	52.00%	54.00%	44.00%	54.00%
Match number in top 10%	42	43	39	43
Match percentage (%)	84.00%	86.00%	78.00%	86.00%
Match number in top 50%	50	50	50	50
Match percentage (%)	100.00%	100.00%	100.00%	100.00%

(b) Significant genes: 50%

Significant genes = 50%		ANOVA			
CV=2	F1	F2	F3	Fs	Cg
Match number in top 5%	29	38	44	47	44
Match percentage (%)	5.80%	7.60%	8.80%	9.40%	8.80%
Match number in top 10%	65	70	79	80	79
Match percentage (%)	13.00%	14.00%	15.80%	16.00%	15.80%
Match number in top 50%	337	319	301	268	301
Match percentage (%)	67.40%	63.80%	60.20%	53.60%	60.20%

Significant genes = 50%		CANOVAM (Cluster size = 3)		
CV=2	F2	F3	Fs	Cg
Match number in top 5%	44	44	50	45
Match percentage (%)	8.80%	8.80%	10.00%	9.00%
Match number in top 10%	92	90	100	91
Match percentage (%)	18.40%	18.00%	20.00%	18.20%
Match number in top 50%	456	457	347	457
Match percentage (%)	91.20%	91.40%	69.40%	91.40%

Table 3.1.6: Simulation results of ANOVA and CANOVAM for three normal mixtures with incorrect cluster sizes are reported.

(a) Significant genes: 5%

Significant genes = 5%		ANOVA			
CV=1	F1	F2	F3	Fs	Cg
Match number in top 5%	5	11	10	10	10
Match percentage (%)	10.00%	22.00%	20.00%	20.00%	20.00%
Match number in top 10%	11	16	15	14	15
Match percentage (%)	22.00%	32.00%	30.00%	28.00%	30.00%
Match number in top 50%	42	44	42	34	42
Match percentage (%)	84.00%	88.00%	84.00%	68.00%	84.00%

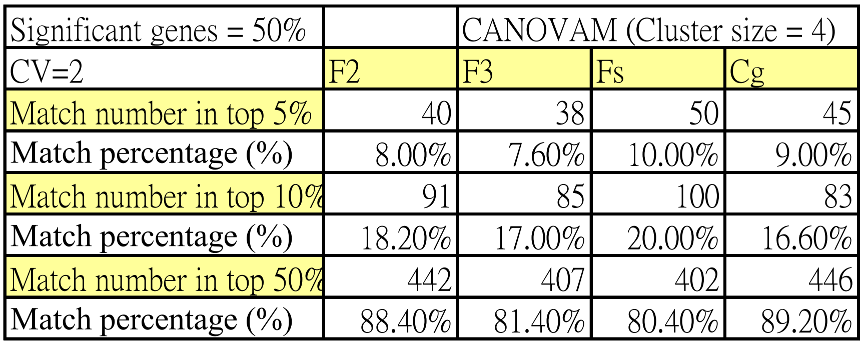
Significant genes = 5%		CANOVAM (Cluster size = 2)		
CV=1	F2	F3	Fs	Cg
Match number in top 5%	21	20	23	22
Match percentage (%)	42.00%	40.00%	46.00%	44.00%
Match number in top 10%	41	42	25	43
Match percentage (%)	82.00%	84.00%	50.00%	86.00%
Match number in top 50%	50	50	50	50
Match percentage (%)	100.00%	100.00%	100.00%	100.00%

Significant genes = 5%		CANOVAM (Cluster size = 4)		
CV=1	F2	F3	Fs	Cg
Match number in top 5%	20	19	25	23
Match percentage (%)	40.00%	38.00%	50.00%	46.00%
Match number in top 10%	39	40	29	43
Match percentage (%)	78.00%	80.00%	58.00%	86.00%
Match number in top 50%	50	50	50	50
Match percentage (%)	100.00%	100.00%	100.00%	100.00%

(b) Significant genes: 50%

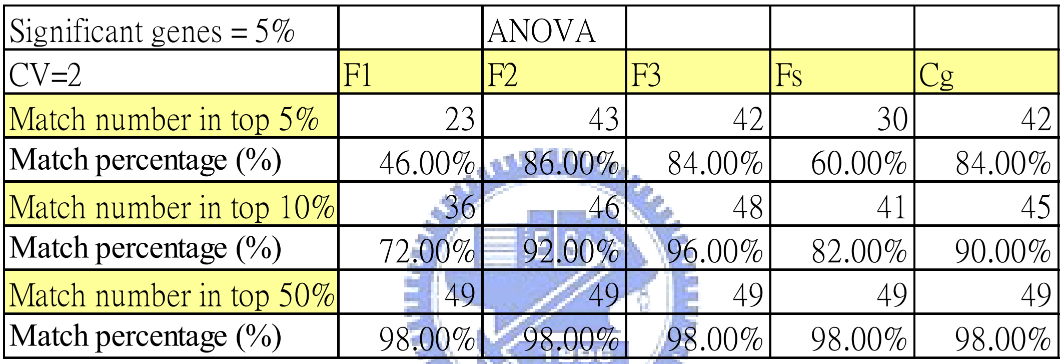
Significant genes = 50%		ANOVA			
CV=2	F1	F2	F3	Fs	Cg
Match number in top 5%	47	50	50	50	50
Match percentage (%)	9.40%	10.00%	10.00%	10.00%	10.00%
Match number in top 10%	91	100	100	98	99
Match percentage (%)	18.20%	20.00%	20.00%	19.60%	19.80%
Match number in top 50%	403	433	369	361	428
Match percentage (%)	80.60%	86.60%	73.80%	72.20%	85.60%

Significant genes = 50%		CANOVAM (Cluster size = 2)		
CV=2	F2	F3	Fs	Cg
Match number in top 5%	44	44	50	45
Match percentage (%)	8.80%	8.80%	10.00%	9.00%
Match number in top 10%	92	90	100	91
Match percentage (%)	18.40%	18.00%	20.00%	18.20%
Match number in top 50%	458	457	397	457
Match percentage (%)	91.60%	91.40%	79.40%	91.40%

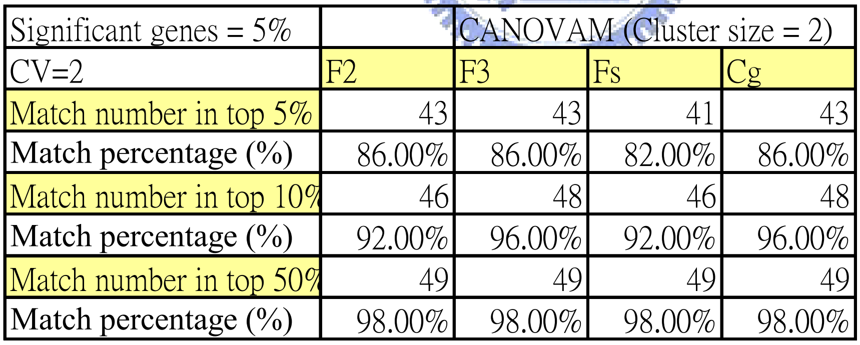


Significant genes = 50%		CANOVAM (Cluster size = 4)		
CV=2	F2	F3	Fs	Cg
Match number in top 5%	40	38	50	45
Match percentage (%)	8.00%	7.60%	10.00%	9.00%
Match number in top 10%	91	85	100	83
Match percentage (%)	18.20%	17.00%	20.00%	16.60%
Match number in top 50%	442	407	402	446
Match percentage (%)	88.40%	81.40%	80.40%	89.20%

Table 3.1.7: Simulation results of ANOVA and CANOVAM for a *t* distribution are reported.



Significant genes = 5%		ANOVA			
CV=2	F1	F2	F3	Fs	Cg
Match number in top 5%	23	43	42	30	42
Match percentage (%)	46.00%	86.00%	84.00%	60.00%	84.00%
Match number in top 10%	36	46	48	41	45
Match percentage (%)	72.00%	92.00%	96.00%	82.00%	90.00%
Match number in top 50%	49	49	49	49	49
Match percentage (%)	98.00%	98.00%	98.00%	98.00%	98.00%



Significant genes = 5%		CANOVAM (Cluster size = 2)		
CV=2	F2	F3	Fs	Cg
Match number in top 5%	43	43	41	43
Match percentage (%)	86.00%	86.00%	82.00%	86.00%
Match number in top 10%	46	48	46	48
Match percentage (%)	92.00%	96.00%	92.00%	96.00%
Match number in top 50%	49	49	49	49
Match percentage (%)	98.00%	98.00%	98.00%	98.00%

Table 3.2.3: The results for spike genes are reported. The number of differentially expressed (i.e., significant or unexpressed) genes is 192 and that of non-differentially expressed (i.e., insignificant or unexpressed) genes is 64. The hypotheses are H0: non-differential expressed (i.e., insignificant or unexpressed) *vs.* H1: differentially expressed (i.e., significant or unexpressed). The top 192 genes with highest ranks of

F statistics are selected as significant genes. The correctly classification and misclassification numbers are reported. The percentages are the number divided by the total number of 256.

ANOVA Significant gene=192	Test declaration:		Number of genes
	Unexpressed	Expressed	
Unexpressed H0	13(5.08%)	51(19.92%)	64
Expressed H1	51(19.92%)	141(55.08%)	192
Total	64	192	

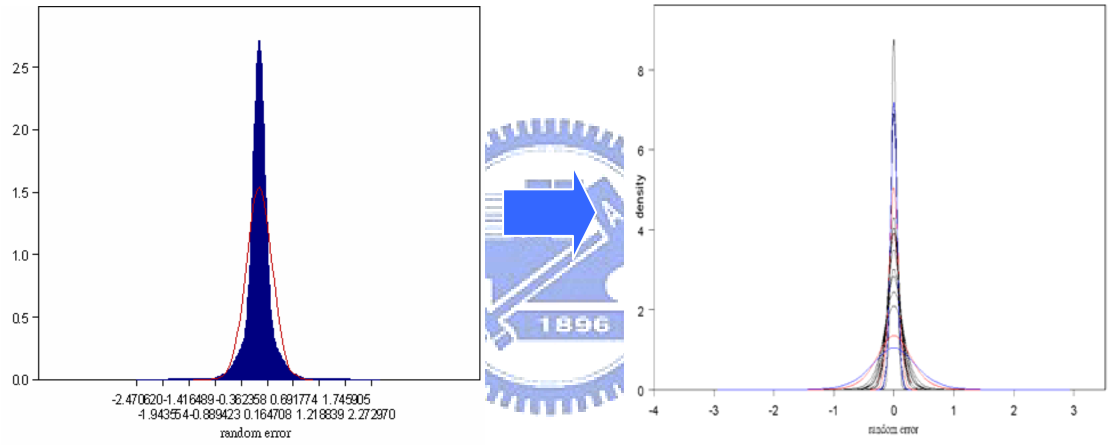
CANOVAM (Cluster size = 4) Significant gene=192	Test declaration:		Number of genes
	Unexpressed	Expressed	
Unexpressed H0	19(7.42%)	45(17.58%)	64
Expressed H1	45(17.58%)	147(57.42%)	192
Total	64	192	

CANOVAM (Cluster size = 12) Significant gene=192	Test declaration:		Number of genes
	Unexpressed	Expressed	

Unexpressed H0	13(5.08%)	51(19.92%)	64
Expressed H1	51(19.92%)	141(55.08%)	192
Total	64	192	

Figure 3.3.5: The histogram and density plot of a normal distribution with the same and variance of all residuals in 24 microarrays are displayed in part (a). The density plot in every cluster is illustrated in part (b) and (c).

(a) (b)



(c)

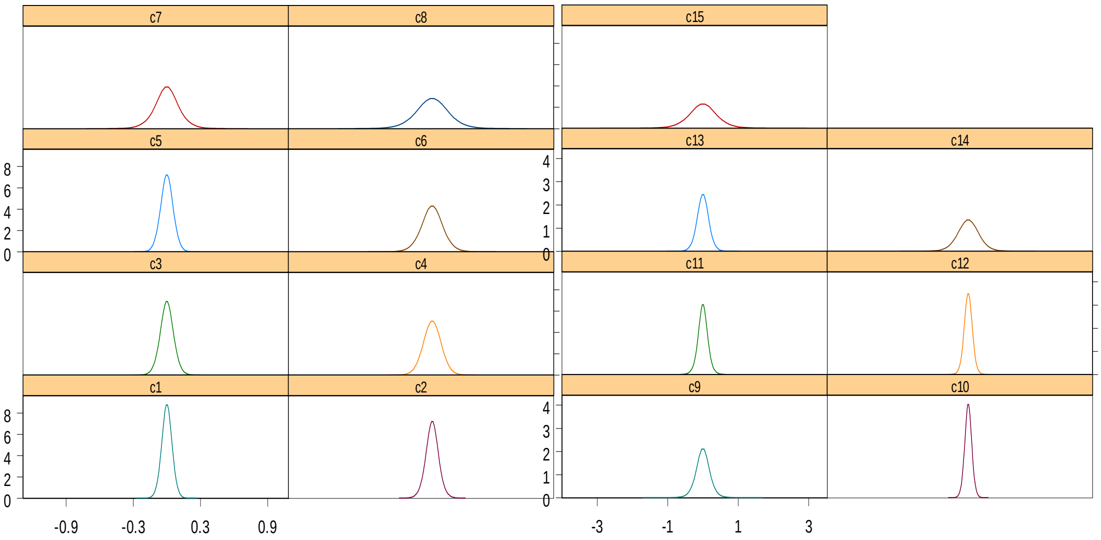
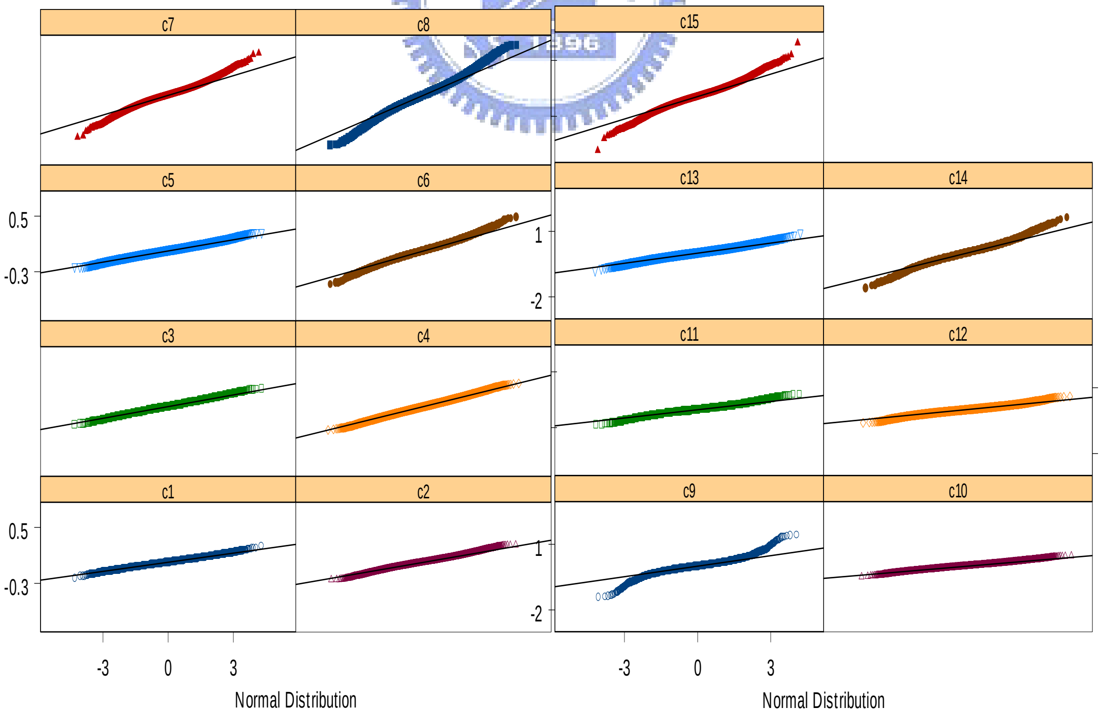


Figure 3.3.6: The normal QQ plot of residuals of 24 microarrays in every cluster is illustrated.



Appendix 2

A2.1 Least Square Estimators for Reference Designs

Assume there are variety *V_k*, $k = 1, \ldots, v$, and a common reference variety V_0 in a reference design without dye swap. Then the ANOVA model becomes

$$y_{ikg}^{(c)} = \mu \ + A_i \ + V_k \ + G_g + (AG)_{ig} + (VG)_{kg} + \varepsilon_{ikg}^{(c)} \qquad \text{(A2.1)}$$

with linear constraints that

$$\sum_i A_i = \sum_g G_g = \sum_g (AG)_{ig} = \sum_g (VG)_{kg} = vV_0 \ + V_1 \ +,\ldots,+V_v \ = v(VG)_{0g} + (VG)_{1g} +,\ldots,+(VG)_{vg} = 0. \qquad \text{(A2.2)}$$

In the reference design without dye swap, there will be no dye effect used in the ANOVA model because the dye effect is completely confounded with the variety effect.

Let $\theta=(\mu, A_i, D_j, V_k, G_g, AG_{ig}, VG_{kg})$. Then, we can take partial derivatives of RSS with respect to the parameters θ for ANOVA models in (A2.1) and (A2.2), where

$$RSS = \sum_{ikg} \ [y_{ijkg}^{(c)} - \mu \ + A_i \ + V_k \ + G_g + (AG)_{ig} + (VG)_{kg}]^2. \qquad \text{(A2.3)}$$

With the constraints in (A2.2), the LSEs of main effects turn out to be:

$$\begin{aligned}
\hat{\mu} \ &= \bar{y}_{\bullet\bullet\bullet}, \\
\widehat{A}_i \ &= \frac{1}{2}(2\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet 0\bullet} - y_{kk\bullet}), \\
\hat{V}_k^{(c)} &= \{ \begin{array}{ll} \bar{y}_{\bullet 0\bullet} - \hat{\mu} & \text{if k} = 0 \\ \bar{y}_{kk\bullet} - \widehat{A}_k - \hat{\mu} & \text{if k} \neq 0 \end{array}, \\
\hat{G}_g^{(c)} &= \bar{y}_{\bullet\bullet g} - \hat{\mu} \ ,
\end{aligned} \qquad \text{(A2.5)}$$

and the LSEs of the interaction terms of array-by-gene and variety-by-gene effects become

$$\begin{aligned}
\widehat{(AG)}_{ig} &= \frac{1}{2}(\bar{y}_{i0g} + \bar{y}_{iig} - 2(\hat{\mu} + \hat{A}_i + \hat{G}_g) \\
&\qquad - \hat{V}_0 - \hat{V}_i - \widehat{(VG)}_{0g} - \widehat{(VG)}_{ig}), \\
\widehat{(VG)}_{0g} &= \bar{y}_{\bullet 0g} - \hat{\mu} - \hat{V}_0 - \hat{G}_g, \\
\widehat{(VG)}_{kg} &= y_{kkg} - \hat{\mu} - \hat{A}_k - \hat{V}_k - \hat{G}_g - \widehat{(AG)}_{kg} \\
&= 2(y_{kkg} - \bar{y}_{kk\bullet} + \bar{y}_{k\bullet\bullet} - \bar{y}_{k\bullet g}) + y_{\bullet 0g} - \bar{y}_{\bullet 0\bullet} + \bar{y}_{\bullet\bullet g} - \bar{y}_{\bullet\bullet\bullet}.
\end{aligned} \qquad \text{(A2.6)}$$

A2.2 Bartlett Test

The Bartlett test is designed to test the equality of variances with the following hypotheses for multiple normal distributions of $\mathrm{N}(\mu_i, \sigma_i^2), i = 1, \ldots, M$:

$\mathrm{H}_0 : \sigma_1 = \sigma_2 = \ldots = \sigma_M$;
$\mathrm{H}_1 : \sigma_i \neq \sigma_j$ for at least one pair of (*i,j*).

The Bartlett test statistics is

$$T = \frac{(N - M)\ln s_{pool}^2 - \sum_{i=1}^{M}(N_i - 1)\ln s_i^2}{1 + (\frac{1}{3(M-1)})((\sum_{i=1}^{M}\frac{1}{N_i}) - \frac{1}{N - M})}, \qquad \text{(A2.13)}$$

where $\boldsymbol{s_i^2}$ is the variance of the *i*th group, *N* is the total sample size, N_i is the sample size of the *i*th cluster, *M* is the number of cluster, and $\boldsymbol{s_p^2}$ is the pooled variance. The pooled variance is a weighted average of the group variances that is defined as

$$s_{pool}^2 = \sum_{i=1}^{M}(N_i - 1)s_i^2 / (N - M)\,. \qquad \text{(A2.14)}$$

If $T > \chi^2_{(\alpha, M-1)}$, then the null hypothesis is rejected, where $\chi^2_{(\alpha, M-1)}$ is the upper critical value of the chi-square distribution with *M*-1 degrees of freedom and a significance level *α*.

A2.3 Levene Test

Levene test is an alternative to the Bartlett's test. The Levene test is less sensitive than the Bartlett test to departures from normality. The Levene test statistics is defined as

$$W = \frac{(N-M)\sum_{i=1}^{M} N_i (\overline{Z}_{i\bullet} - \overline{Z}_{\bullet\bullet})^2}{(M-1)\sum_{i=1}^{M}\sum_{j=1}^{N_i} (Z_{ij} - \overline{Z}_{i\bullet})^2}, \qquad \text{(A2.15)}$$

where Z_{ij} can have one of the following three definitions:

1. $Z_{ij} = | Y_{IJ} - \overline{Y}_i |$, where $\overline{Y}_i$ is the mean of the *i*th subgroup.

2. $Z_{ij} = | Y_{IJ} - \widetilde{Y}_i |$, where $\widetilde{Y}_i$ is the median of the *i*th subgroup.

3. $Z_{ij} = | Y_{IJ} - \widetilde{Y}_i' |$, where $\widetilde{Y}_i'$ is the 10% trimmed mean of the *i*th subgroup.

Note that $\overline{Z}_{i\bullet}$ is the group mean of Z_{ij} and $\overline{Z}_{\bullet\bullet}$ is the overall mean of Z_{ij}. The Levene test rejects the hypothesis that the variances are equal if $W > F_{(\alpha, k-1, N-M)}$, where $F_{(\alpha, k-1, N-M)}$ is the upper critical value of the *F* distribution with *k-1* and *N-k* degrees of freedom at a significance level of *α*.

References

1. Akaike H. Fitting autoregressive models for prediction. Ann. Inst. Stat. Math 1969. 21:243-247.
2. Aitkin M. A general maximum likelihood analysis of variance components in generalized linear models. Biometrics. 1999 Mar;55(1):117-28.
3. Aitkin M, Rubin DB. Estimation and hypothesis test in finite mixture models. JRSS Series B. 1985 Vol. 47 No. 1 67-75.
4. Bhattacharya S, Long D, Lyons-Weiler J. Overcoming confounded controls in the analysis of gene expression data from microarray experiments. Appl Bioinformatics. 2003;2(4):197-208.
5. Cui X, Churchill GA. . Statistical tests for differential expression in cDNA microarray experiments. Genome Biol. 2003;4(4):210. Epub 2003 Mar 17. Review.
6. Cui X, Hwang JTG, Qiu J, Blades, NJ and Churchill GA. Improved statistical tests for differential gene expression by shrinking variance components estimates. Technical Report 2003 [http://www.jax.org/staff/churchill/labsite/pubs/index.html]
7. Tsai CA, Hsueh HM, Chen JJ. Estimation of false discovery rates in multiple testing: application to gene microarray data. Biometrics. 2003 Dec; 59(4):1071-81.
8. Dempster AP, Laird NM, Rubin DB. Maximum likelihood form incomplete data via the EM algorithm (with discussion). J.R Statist 1977. Soc B,39,1-38.
9. Efron B, Tibshirani RJ. An introduction to the bootstrap. Chapman & Hall 1993.
10. Ghosh D. Mixture models for assessing differential expression in complex tissues using microarray data. Bioinformatics. 2004 Feb 26 [Epub ahead of print]

11. Good P. Permutation Tests. Springer 2000.

12. Hotelling, Harold. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology 1933, 24, 471-411, 498-520.

13. Kerr MK, Martin M, Churchill GA. Analysis of variance for gene expression microarray data. J Comput Biol. 2000;7(6):819-37..

14. Kerr MK, Churchill GA.. Experimental design for gene expression microarrays. Biostatistics. 2001 Jun;2(2):183-201.

15. Kerr MK and Churchill GA. Statistical design and analysis of gene expression microarray. Genetical Research 2001b; 77:123-128.

16. Kerr MK, Leiter E, Picard L and Churchill GA. Analysis of a designed microarray experiment. Proceedings of the IEEE-Eurasip Nonlinear Signal and Inage Processing Workshop. June 3-6 2001.

17. Kerr MK. Design considerations for efficient and effective microarray studies. Biometrics. 2003 Dec; 59(4):822-8.

18. K. P Burnham and DR Anderson (1998). Model selection and inference: A practical information-theoretic approach. Springer.

19. Levene H. In Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling, I. Olkin *et al*. eds. Stanford University Press 1960 pp278-292.

20. Li C, Wong WH.. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci U S A. 2001 Jan 2;98(1):31-6.

21. McLachlan GJ, Bean RW, Peel D. A mixture model-based approach to the clustering of microarray expression data. Bioinformatics. 2002 Mar; 18(3):413-22.

22. Qiu J and Hwang J.T.G. Sharp simulations intervals for the means of selected populations with application to microarray data analysis. Research paper 2003.

[http://www.math.cornell.edu/~hwang/publications.html]

23. Reverter A, Byrne KA, Brucet HL, Wang YH, Dalrymple BP, Lehnert SA. A mixture model-based cluster analysis of DNA microarray gene expression data on Brahman and Brahman composite steers fed high-, medium-, and low-quality diets. J Anim Sci. 2003 Aug;81(8):1900-10.

24. Render RA, Walker HF. Mixture densities, maximum likelihood and the EM algorithm. SIAM Review. 1984. Vol. 26, NO. 2: 195-239.

25. Ross SM. Simulation 2nd edition. Academic Press, Boston. 1997.

26. Schwartz. G. Estimating the dimensions of a model. Ann. Stat.6:461-464 1978.

27. Smyth GK, Yang YH, Speed T. Statistical issues in cDNA microarray data analysis.

    Methods Mol Biol. 2003; 224:111-36. No abstract available.

28. Snedecor, George W. and Cochran, William G. Statistical Methods, Eighth Edition, Iowa State University Press. 1989

29. Tsai PW, Lee MLT. Split-plot microarray experiments: Issues of design, power and sample size. Technical Report 2004.

30. Wu CFJ. On the convergence properties of the EM algorithm. Annals of Statistics 1983, 11, 95-103.

31. Wu H. R/ maanova. 2003.

    [http://www.jax.org/staff/Churchill/labsite/software/anova/rmaanova

32. Wu H, Kerr MK, Cui X, Churchill GA. MAANOVA: a software package for the analysis of spotted cDNA microarray experiments. 2003.

    [http://www.jax.org/staff/churchill/labsite/pubs/Wu_maanova.pdf]

33. Wang Y, Luo L, Freedman MT and Kung SY. Probabilistic principal component subspaces: A hierarchical finite mixture model for data visualization. IEEE Transactions on neural networks, Vol.11, No.3, May 2000

34. Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules RS. Assessing gene significance from cDNA microarray expression data via mixed models. J Comput Biol. 2001;8(6):625-37.