

Model-Based Clustering in Gene Expression Microarrays: An Application to Breast Cancer Data

J.C. Mar and G.J. McLachlan

Department of Mathematics,
University of Queensland Q 4072

Contact: Geoff McLachlan, gjm@maths.uq.edu.au

Abstract

In microarray studies, the application of clustering techniques is often used to derive meaningful insights into the data. In the past, hierarchical methods have been the primary clustering tool employed to perform this task. However attention is now turning to model-based clustering approaches. The hierarchical algorithms have been mainly applied heuristically to these cluster analysis problems. Further, a major limitation of these methods is their inability to determine the number of clusters. Thus there is a need for a model-based approach to these clustering problems. To this end, McLachlan et al. (2002) developed a mixture model-based algorithm (EMMIX-GENE) for the clustering of tissue samples. To further investigate the EMMIX-GENE procedure as a model-based approach, we present a case study involving the application of EMMIX-GENE to the breast cancer data as studied recently in van't Veer et al. (2002). Our analysis considers the problem of clustering the tissue samples on the basis of the genes which is a non-standard problem because the number of genes greatly exceed the number of tissue samples in a typical study. We demonstrate how EMMIX-GENE can be useful in reducing the initial set of genes down to a more computationally manageable size. The results from this analysis also emphasise the difficulty associated with the task of separating two tissue groups on the basis of a particular subset of genes. These results also shed light on why supervised methods have such a high misallocation error rate for the breast cancer data.

Keywords: microarray, mixture modelling, cluster analysis.

1 Introduction

The complexity and magnitude of DNA microarray data have inundated researchers with a flood of new bioinformatic challenges. The data generated by these experiments necessitate the use of specialised statistical tools in order to make reliable inferences about the data. In this paper we discuss the application of a model-based approach to the cluster analysis for gene expression microarrays.

Cluster analyses have previously demonstrated their utility in the elucidation of unknown gene function, the validation of gene discoveries, and the interpretation of biological processes; see Alizadeh et al. (2000), Eisen et al. (1998), Iyer et al. (1999) for examples. The aim of a

typical cluster analysis is to organise genes or tissue samples (data produced by separate hybridisations) into groups or clusters displaying similar patterns of gene expression. In the past, mainly hierarchical methods have been applied to cluster analysis problems. However attention is now turning to model-based approaches; see Ghosh and Chinnaiyan (2002), McLachlan et al. (2002), Pan et al. (2002), Yeung et al. (2001).

The majority of clustering algorithms that have been employed in the literature are largely heuristically motivated. There exist a number of unresolved issues associated with the use of these algorithms, including how to determine the number of clusters.

As commented by Yeung et al. (2001), "in the absence of a well-grounded statistical model, it seems difficult to define what is meant by a 'good' clustering algorithm or the 'right' number of clusters." They have advocated a model-based approach to clustering by adopting a finite mixture model for the distribution of each observation. In their study, they were concerned with the clustering of the genes on the basis of the tissue samples. Here we consider the problem of clustering the tissues on the basis of the genes, which is a more challenging problem to consider in a mixture model framework, since the number of observations to be clustered (the tissue samples) is typically small relative to the number of genes in each tissue sample.

A recent application of microarray technology involves its use in the development of patient-tailored therapies to target complex, highly heterogeneous diseases. The work of van't Veer et al. (2002) used microarray experiments on three patient groups who had different classes of breast cancer tumours. The goal of the overall experiment was to identify a set of genes that could distinguish between the different tumour groups based on their gene expression information for a given tumour sample. We use the data set produced by this group as the basis for our analysis using a mixture model-based clustering algorithm called EMMIX-GENE (McLachlan et al., 2002).

The results of this analysis shed light on why it is such a difficult problem to distinguish between the two tissue groups (disease-free and metastases) and consequently why supervised methods have such a high error rate for this data set, as noted by Tibshirani and Efron (2002).

2 EMMIX-GENE: A Mixture Model-based Clustering Algorithm

The EMMIX-GENE algorithm consists of three stages (see McLachlan et al., 2002 for more specific details). The first is a filtering step designed to isolate the most informative genes to be considered in the cluster analysis. For all genes in the original data set mixtures of t distributions are fitted, and each gene is assigned a value of the test statistic $-2\log\lambda$ that tests for the presence of a single component versus two components in the fitted mixture model (λ is the likelihood ratio). Genes displaying strong differential expression across different tumour groups will yield a significantly large value of the test statistic, whereas those genes bearing little change will receive a lower score. Hence genes that have test scores that fall above a user-specified threshold are retained, while all other genes are discarded from the analysis.

The second stage involves grouping the retained set of genes into a user-specified number of clusters. The genes have been clustered into groups using Euclidean distance with a view to representing the genes within a group by their mean for each tissue.

If a clustering is sought on the basis of the totality of the genes, then it can be obtained by fitting a mixture model to the group means. If the number of group means N is too large to fit a normal mixture model with unrestricted component-covariance matrices, then EMMIX-GENE has the facility for the fitting of mixtures of factor analyzers. The use of the latter reduces the number of parameters imposing the assumption that the correlations between the genes can be expressed in a lower space by the dependence of the tissues on q ($q < N$) unobservable factors.

In addition to clustering the tissues on the basis of all of the genes, there may be interest in seeing if the different groups of genes separately lead to different clustering of the tissues when each is considered separately. For example, only a subset of the genes may be useful in identifying certain subtypes of the cancer being studied.

3 Description of the Experimental Data Set

In van't Veer et al. (2002), microarray experiments were performed on 98 primary breast cancers acquired from three groups of patients: 44 representing a good prognosis group, (ie. those who remained metastasis free after a period of more than 5 years), 34 from a poor prognosis group (those who developed distant metastases within 5 years), and 20 representing a hereditary form of cancer, due to a BRCA1 (18 tumours) or BRCA2 (2 tumours) germline mutation.

Each microarray experiment involved an initial set of 24,881 genes. To reduce the number of genes to something more computationally manageable, the same pre-processing filter used by van't Veer et al. (2002) was applied to the data at the outset of our analysis. The selection criteria of this filter required a gene to have both a P -value of less than 0.01 and at least a two fold difference in more than five out of the ninety-eight tissues

for the gene to be retained. The initial set of genes was effectively reduced to 4,869 genes using this criteria.

The focus of van't Veer's study was to identify a subset of genes that would be useful in predicting the disease outcome of any given tissue sample. They anticipated that this gene signature could be applied as a diagnostic screen to select patients that would benefit from certain therapies over others.

4 An Unsupervised Classification Analysis Using EMMIX-GENE

The first step of the EMMIX-GENE algorithm was used to select the most relevant genes from this filtered set of 4,869 genes, reducing this number further to 1,867. The retained genes were clustered into forty groups using the second step of the EMMIX-GENE algorithm, and the majority of gene groups produced were reasonably cohesive and distinct. Based upon these forty group means, the tissue samples were clustered into two and three components using a mixture of factor analyzers model with $q = 4$ factors.

5 Investigating the Usefulness of the Selection of Relevant Genes

In clustering the genes, van't Veer et al. (2002) relied upon an agglomerative hierarchical algorithm to organise the genes into dominant genes groups. Two of these clusters were highlighted in the paper and the genes contained in these two groups correspond to biologically significant features. We denote Cluster A as the group of genes van't Veer et al. have identified as containing genes co-regulated with the ER- α gene (ESR1) and Cluster B as the group containing "co-regulated genes that are the molecular reflection of extensive lymphocytic infiltrate, and comprise a set of genes expressed in T and B cells". Both of these clusters contain 40 genes.

	Cluster Index	Number of Genes Matched	Percentage Matched (%)
Cluster A	2	21	87.5
	3	2	8.33
	14	1	4.17
Cluster B	17	18	78.3
	19	1	4.35
	21	4	17.4

Table 1: Comparing Clusters Constructed by an Hierarchical Algorithm with those Produced by the EMMIX-GENE Algorithm

Of these 80 genes, the first step of the EMMIX-GENE algorithm `select-genes` retains only 47 genes (24 from Cluster A, 23 from Cluster B). When compared to the 40 groups that the `cluster-genes` step of the EMMIX-GENE algorithm produces, subsets of these 47

genes appeared inside several of these 40 groups (see Table 1 above).

The motivation behind `select-genes` is to isolate the most informative genes to be used for the cluster analysis. For any clustering algorithm, genes which lack distinctive expression pattern changes across different tumour groups only serve to confuse the clustering algorithm and increase the number of misallocation errors made.

The 21 genes which appear in Cluster A, have been grouped in the second cluster constructed by EMMIX-GENE. In Figure 1 (below), these genes demonstrate clear expression changes for the three groups of tumours (indicated by the horizontal blue lines).

For the remaining sixteen genes that were rejected by `select-genes` but belong to Cluster A, it is evident from Figure 2 that these genes bear very little information in distinguishing between the tumour groups.

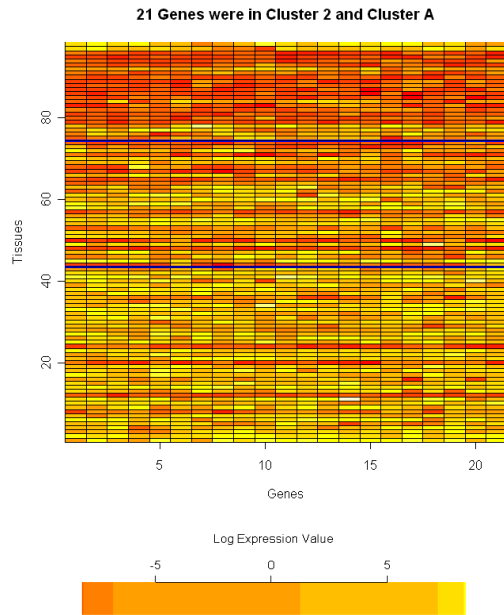


Figure 1: Genes Retained by EMMIX-GENE Appearing in Cluster A

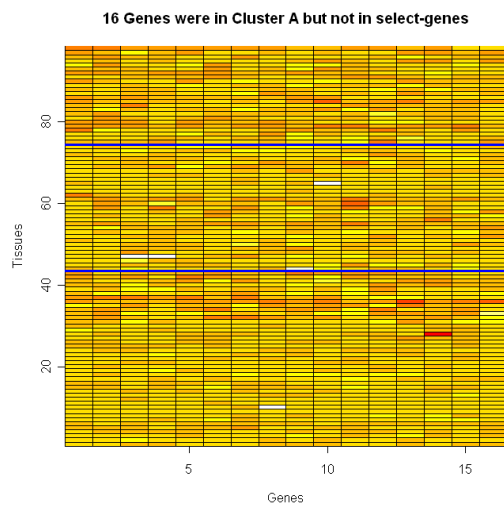


Figure 2: Genes Rejected by EMMIX-GENE Appearing in Cluster A

The heat map displays almost a constant level of expression for these genes across all tumours, the same observations apply to genes in Cluster B (see Figure 3 and 4). The expression profile of the gene which received the highest $-2\log\lambda$ value is shown in Figure 5. This gene is notably up-regulated for the disease-free tumour group and the metastases tumour group, and down-regulated in the hereditary tumour group.

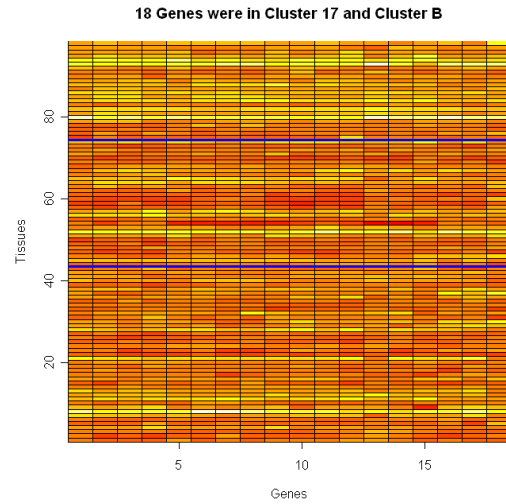


Figure 3: Genes Retained by EMMIX-GENE Appearing in Cluster B



Figure 4: Genes Rejected by EMMIX-GENE Appearing in Cluster B

An expression profile is shown in Figure 6 for a gene which appeared in Cluster A but whose $-2\log\lambda$ value was not high enough to be retained by the `select-genes` step.

The overall expression of the gene is essentially unchanging, however excessively large values for the seventeenth disease-free patient in the first tumour group and the sixth BRCA patient in the third tumour group appear to dominate the expression profile. These outliers seem to account for this gene's inclusion in Cluster A.

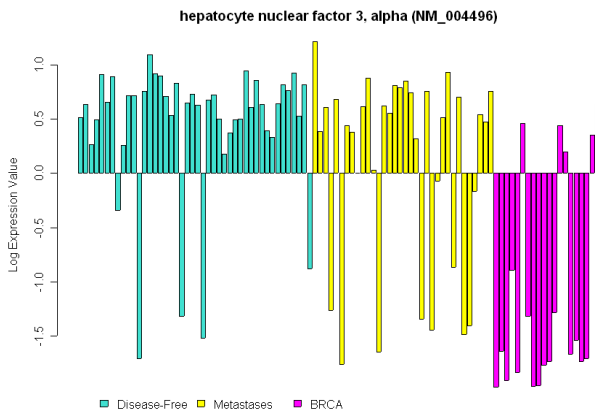


Figure 5: Expression Profile for the Gene with the Highest $-2\log\lambda$ Value

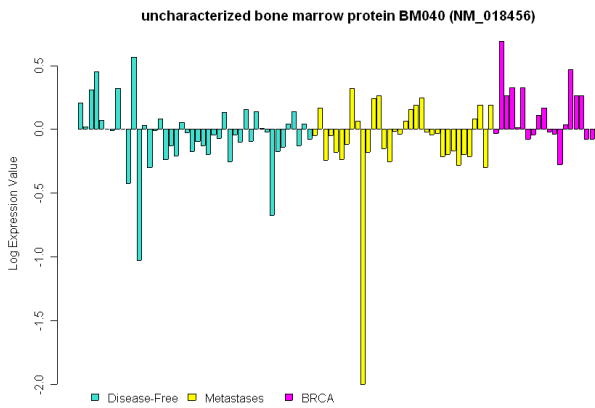


Figure 6: Example of A Gene Rejected by *select-genes* But Retained by Cluster A

6 Clustering Genes on the Basis of Tissue Samples Using EMMIX-GENE

As can be seen by the heat map displayed in Figure 7 (below), the task of discerning an underlying class

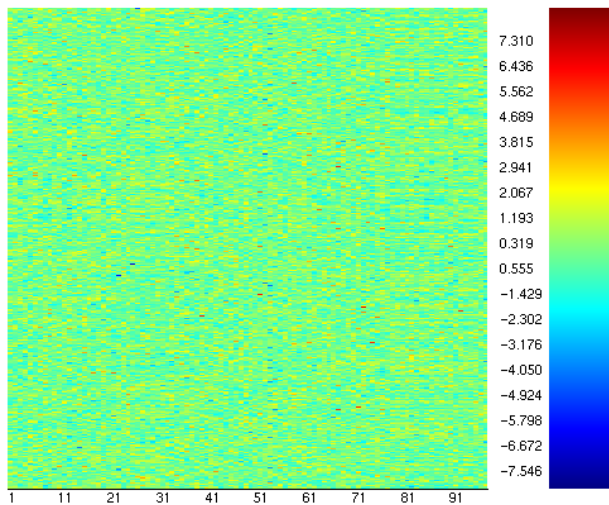


Figure 7: Heat Map Displaying the Initial Set of 4,869 Genes in the Breast Cancer Data

structure in the data on the basis of the full set of 4,869 genes is extremely difficult.

For the present breast cancer data set, the heat maps of the genes in a group tend to mainly support the same breakup of the 98 tissues. To illustrate this, we list in Figures 8 to 10 the heat maps for the top three groups G_1 , G_2 , and G_3 , which contain 146, 93, and 61 genes, respectively. Important features to note from these heat maps are that they each indicate a change in gene expression is apparent between the sporadic (first 78 tissue samples) and hereditary (last 20 tissue samples) tumours. For instance, in Figure 8, the genes in this cluster are generally down-regulated for the former group of tumours, and up-regulated in the latter. Genes in G_2 were largely constant in expression across the sporadic tumours but notably down-regulated for the hereditary tumours. Additionally, the final two tissue samples, which represent the two BRCA2 tumours show consistent patterns of expression in each of the clusters that are different from those exhibited by the set of BRCA1 tumours.

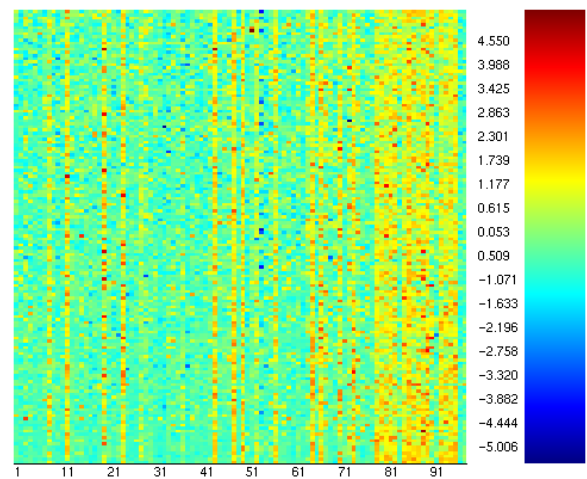


Figure 8: Heat Map of Genes in Group G_1

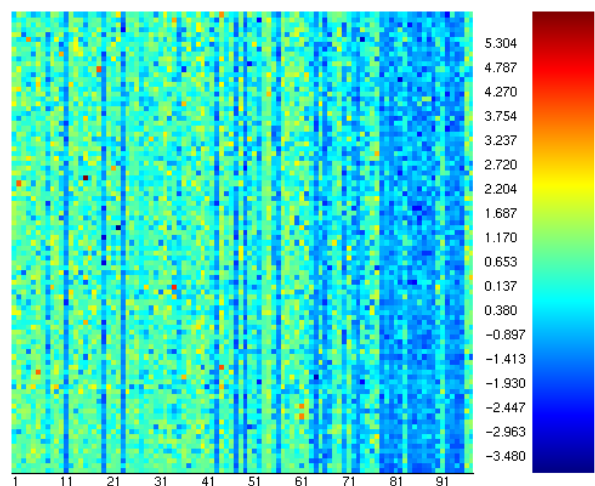


Figure 9: Heat Map of Genes in Group G_2

It can be seen from these groups that the problem of trying to distinguish between the two classes, patients who were disease-free after 5 years Π_1 and those with

metastases within 5 years Π_2 , is not straightforward on the basis of the gene expressions.

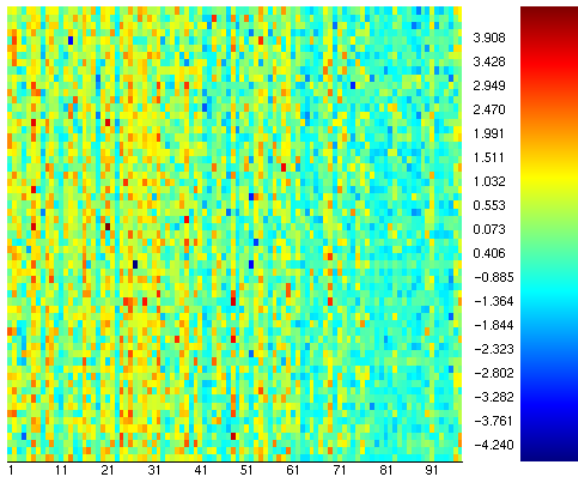


Figure 10: Heat Map of Genes in Group G_3

7 Clustering Tissue Samples on the Basis of Gene Groups Using EMMIX-GENE

Turning now to the problem of clustering tissues on the basis of gene expression, we investigate the clusters constructed by the EMMIX-GENE algorithm in light of the genuine tissue grouping.

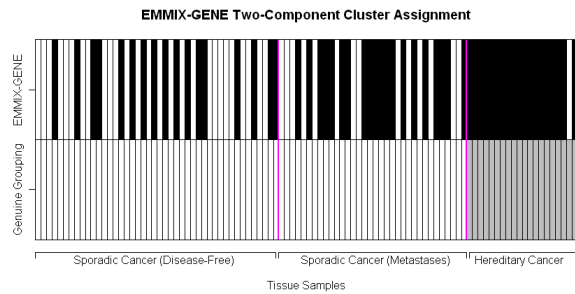


Figure 11: Comparing EMMIX-GENE Cluster Assignments with the Genuine Two Group Structure

The tissue samples can be subdivided into two groups corresponding to the 78 sporadic tumours and 20 hereditary tumours. Figure 11 shows these two clusters with respect to the genuine grouping. EMMIX-GENE has correctly clustered the majority of the hereditary tumours (misallocation error of $1/20$), although 37 of the sporadic tumours were incorrectly assigned to the cluster of hereditary tumours. (Pink vertical lines denote the three tumour groups; black denotes the hereditary tumour cluster, white denotes the sporadic tumour cluster; grey distinguishes the genuine grouping).

The set of sporadic tumours can be further divided into good and poor prognosis groups, ie. 44 patients who continued to be disease-free after 5 years, and 34 patients who developed metastases within 5 years, respectively.

Figure 12 shows the tissue samples rearranged according to the three cluster assignments allocated by EMMIX-GENE when a mixture of factor analyzers model with $q = 4$.

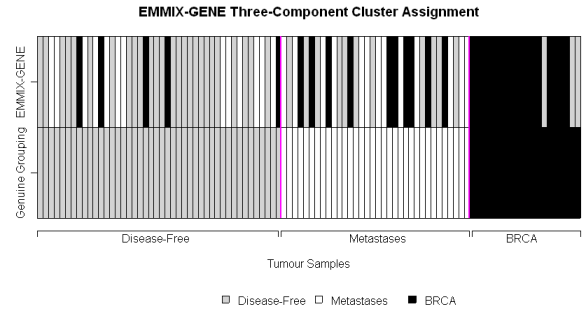


Figure 12: Comparing EMMIX-GENE Cluster Assignments with a Genuine Three-Group Structure

Using a mixture of factor analyzers model with $q = 8$ factors, we would misallocate 7 out of the 44 members of Π_1 and 24 out of the 34 members of Π_2 ; one member of the 18 BRCA1 samples would be misallocated.

The misallocation rate of $27/34$ for the second class Π_2 is not surprising given the genes expressions as summarized in the groups of genes (see Figures 8 to 10). Also, one has to bear in mind that we are classifying the tissues in an unsupervised manner without using the knowledge of their true classification. But even when such knowledge was used (supervised classification) in van't Veer et al. (2002), the reported error rate was approximately 50% for members of Π_2 when allowance was made for the selection bias in forming a classifier on the basis of an optimal subset of the genes (Ambroise and McLachlan 2002). Further analysis of this data set in a supervised context by Tibshirani and Efron (2002) confirms the difficulty in trying to discriminate between the disease-free class Π_1 and the metastases class Π_2 .

8 Assessing the Number of Tissue Groups

We also considered the choice of the number of components g to be used in our normal mixture. The likelihood ratio test statistic $-2\log\lambda$ was adopted for this purpose, and we used the resampling approach of McLachlan (1987) to assess the P -value. This is because the usual chi-squared approximation to the null distribution of $-2\log\lambda$ is not valid for this problem, due to the breakdown in regularity conditions. We proceeded sequentially, testing the null hypothesis $H_0: g = g_0$ versus the alternative hypothesis $H_1: g = g_0 + 1$, starting with $g_0 = 1$ and continuing until a non-significant result was obtained. We concluded from these tests that $g = 3$ components were adequate for this data set.

9 Investigating Underlying Signatures with Other Clinical Indicators

For each of the tumour samples in this data set, additional clinical predictors containing information about histological grade, angioinvasion and lymphocytic infiltrate was included. We investigated whether the three clusters constructed by EMMIX-GENE followed patterns according to these biological indicators. The tumour samples have been ordered in Figure 13 according to the three clustered groups.

Tumours assigned to Cluster 3 appear to match tumours labelled ER positive, while the majority of tumours in Clusters 1 and 2 were ER negative. A close association was also noted between tumours assigned to Cluster 1 and a histological grade of 3, while the tumours in Clusters 2 and 3 were more likely to have a histological grade of 1 or 2. Some association was visible between Clusters 1 and 2 and the lymphocytic infiltrate score, where the majority of tumours in these clusters had scores of 0, and tumours in Cluster 3 had scores of 1. Indicators related to angiogenesis did not bear a strong association with the EMMIX-GENE clusters. These observations were consistent with those reported by van't Veer et al. (2002).

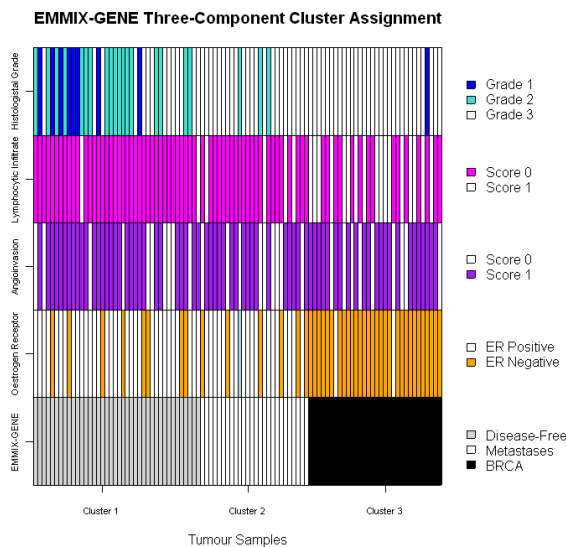


Figure 13: Comparing EMMIX-GENE Cluster Assignments with Other Clinical Indicators

10 Discussion

When we identified the clusters produced by EMMIX-GENE with the externally existing classes Π_1 (disease-free group), Π_2 (metastases group), and Π_3 (BRCA), the error rate of this rule is not small.

However, it is consistent with the gene expressions as displayed in the heat maps for the 40 groups of similar genes. For example, in the first three groups given in Figures 8 to 10, it can be seen that tissues of class Π_2 have similar gene expression patterns to those of the majority of the tissues in class Π_1 (Π_3) that have similar gene expression patterns to those of the majority of the tissues in class Π_1 (Π_3). Likewise, the misallocated tissues of the class Π_1 have similar gene expression patterns to those of the majority of the tissues in class Π_2 .

11 References

- ALIZADEH, A., EISEN, M.B., DAVIS, R.E., MA, C., LOSSOS, I.S., ROSENWALD, A., BOLDRICK, J.C., SABET, H., TRAN, T., YU, X., et al. (2000): Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**:503-511.
- AMBROISE, C., and MCLACHLAN, G.J. (2002): Selection bias in gene extraction on basis of microarray

gene expression data. *Proceedings of the National Academy of Sciences USA* **99**:6562-6566.

EISEN, M.B., SPELLMAN, P.T., BROWN, P.O., and BOTSTEIN, D. (1998): Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences USA* **95**: 14863-14868.

GHOSH, D., and CHINNAIYAN, A.M. (2002): Mixture modelling of gene expression data from microarray experiments. *Bioinformatics* **18**:275-286.

IYER, V.R., EISEN, M.B., ROSS, D.T., SCHULER, G., MOORE, T., LEE, J.C.F., TRENT, J.M., STAUDT, L.M., HUDSON, J., BOGUSKI, M.S., LASHKARI, D., SHALON, D., BOTSTEIN, D., and BROWN, P.O. (1999): The transcriptional program in the response of human fibroblasts to serum. *Science* **283**:83-87.

MCLACHLAN, G.J. (1987): On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics* **36**:318-324.

MCLACHLAN, G.J., BEAN, R.W., and PEEL, D. (2002): A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics* **18**:413-422.

PAN, W., LIN, J., and LE, C.T. (2002): Model-based cluster analysis of microarray gene expression data. *Genome Biology* **3**(2):research0009.1-0009.8.

TIBSHIRANI, R.J., and EFRON, B. (2002): Pre-validation and inference in microarrays. *Statistical Applications in Genetics and Molecular Biology* **1**:1.

VAN 'T VEER, L.J., DAI, H., VAN DE VIJVER, M., HE, Y.D., HART, A.M., MAO, M., PETERSE, H.L., VAN DER KOOY, K., MARTON, M.J., WITTEVEEN, A.T., SCHREIBER, G.J., KERKHOVEN, R.M., ROBERTS, C., LINSLEY, P.S., BERNARDS, R., and FRIEND, S.H. (2002): Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**:530-536.

YEUNG, K.Y., FRALEY, C., MURUA, A., RAFTERY, A.E., and RUZZO, W.L. (2001): Model-based clustering and data transformations for gene expression data. *Bioinformatics* **17**:977-987.