# Robust Model-Based Clustering of Genes in Microarray Data: Are there Gene Clusters?

Johanna Hardin

Southwest Oncology Group

Fred Hutchinson Cancer Research Center

johannah@swog.fhcrc.org

David M. Rocke

Center for Image Processing and Integrated Computing

University of California at Davis

dmrocke@ucdavis.edu

David L. Woodruff

Graduate School of Management

University of California at Davis

dlwoodruff@ucdavis.edu

December 2000

**Abstract**

Microarray technology has produced the capability to collect data on large amounts of genetic information. We provide a method of refinement to the clustering of genes based on such technology with the idea that these clusters might lead to the discovery of genetic pathways that cause various diseases. Our method uses Rousseeuw's Minimum Covariance Determinant (MCD) (Lopuhaä and Rousseeuw, 1991; Rousseeuw and Leroy, 1987) as a robust measure of location and shape. We use the MCD of each cluster as a robust, data-dependent, measure of each of the different clusters. We then apply F-distribution quantiles to the Mahalanobis squared distances of the data, based on the MCD parameters, to find points which may in fact belong to more than one cluster or may be outliers which do not seem to belong to any cluster (Hardin and Rocke, 2000a; Hardin and Rocke, 2000b). The results are applied to the leukemia data of Golub et al. (Golub et al., 1999) after subsetting for genes that are primarily expressed as present across the samples. The clusters are then analyzed as to their validity and usefulness.

# 1 Introduction

With the advent of microarray technology has come the importance of developing new statistical methods to use on microarray data. Some new methods are designed to answer questions that look at the difference in genetic make up of two different groups with respect to their classification (e.g., a treatment group and a control group) (Golub et al., 1999; Bittner et al., 2000; Dudoit et al., 2000a). Our work is aligned with methods that are designed to cluster genes with respect to expression across many samples and may or may not take into account the classification of the samples (Hastie et al., 2000; Bittner et al., 2000).

Microarray data exists in high dimension (in our work, dimension is sample size, typically 10-50 samples), which makes it difficult or impossible to view the data and the clusterings graphically. Because of this restriction, overlapping clusters and outlying points can go undetected. Also, methods which force all

points to belong to one or another cluster can become distorted if there exists one or more severely outlying points. Outlier detection methods applied to clustered data can give new insight to the structure of the data and to which points belong to no clusters or multiple clusters. After outlying points are identified, clustering methods can be reapplied to give more accurate results.

Various method for detecting outliers in the one cluster multiple dimensional setting have been studied (Atkinson, 1994; Barnett and Lewis, 1994; Gnanadesikan and Kettenring, 1972; Hadi, 1992; Hawkins, 1980; Maronna and Yohai, 1995; Penny, 1995; Rocke and Woodruff, 1996; Rousseeuw and VanZomeren, 1990). One way to identify possible multivariate outliers is to calculate a distance from each point to a "center" of the data. An outlier would then be a point with a distance larger than some predetermined value. A conventional measurement of quadratic distance from a point X to a location Y given a shape S, in the multivariate setting is:

$$d_S^2(X,Y) = (X-Y)^T S^{-1}(X-Y)$$

This quadratic form is often called the Mahalanobis Squared Distance (MSD).

In the clustering context, an outlier can be thought of as a point with a large MSD from the center of each and every one of the clusters. (A point with a small MSD from multiple clusters can be thought of as a borderline point or a point belonging to multiple clusters.) After a robust clustering algorithm is applied to the data, each of the clusters can be thought of as coming from a unique population, and outlier identification methods can be used individually on each cluster.

For data that come from one population, the distribution of the MSD with both the true location and shape parameters and the conventional location and shape parameters is well known (Gnanadesikan and Kettenring, 1972). However, the conventional location and shape parameters are not robust to outliers, and the distributional fit breaks down when robust measures of location and shape are used in the MSD (Rousseeuw and VanZomeren, 1990). Hardin and Rocke (2001a) developed a distributional fit to Mahalanobis distances which use a robust shape and location estimate, namely the Minimum Covariance Determinant (MCD).

Given $n$ data points, the MCD of those data is the mean and covariance matrix based on the sample of size $h$ ($h \leq n$) that minimizes the determinant of the covariance matrix (Rousseeuw, 1984).

$$\begin{aligned} MCD &= (\overline{X}^*_J, S^*_J) \\ \text{where} \quad J &= \{ \text{ set of } h \text{ points:} |S^*_J| \leq |S^*_K| \quad \forall \text{ sets } K \text{ s.t. } |K| = h\} \\ \overline{X}^*_J &= \frac{1}{h}\sum_{i \in J} x_i \\ S^*_J &= \frac{1}{h}\sum_{i \in J}(x_i - \overline{X}^*_J)(x_i - \overline{X}^*_J)^t \end{aligned}$$

The value $h$ can be thought of as the minimum number of points which must not be outlying. We will use $h = \lfloor \frac{(n+p+1)}{2} \rfloor$ in our calculations and refer to a sample of size $h$ as a half sample. The MCD is computed from the "closest" half sample, and therefore, the outlying points will not skew the MCD location or shape estimates. The concept of the MCD can be modified easily to fit the multiple cluster setting. With a good initialization and a known number of clusters, $g$, the MCD can be found separately for each of the clusters. The size of each cluster is determined by the number of points which are closer to that cluster's center than to any other cluster center. The sizes of the clusters and the MCD samples will be $n_i$ and $h_i = \lfloor \frac{(n_i+p+1)}{2} \rfloor \quad i = 1, \ldots, g$, respectively.

Using the MCD estimates in the MSD leads to robust distances from each of the cluster centers for each of the data points. A datum which is outlying will have a large robust distance from each of the cluster centers. However, not every data set will give rise to an obvious separation between extreme points which belong to the data set (i.e. are not outliers) and those which do not (i.e. are outliers.) In order to distinguish between these two types of extrema, outliers can be identified using the quantiles of an F distribution (Hardin and Rocke, 2000a) on a cluster by cluster basis. An outlying point will be labeled as such only if it is outlying with respect to all of the clusters. Additionally, points that have small distances (according to the F-cutoffs) to more than one cluster can been seen as borderline points, points which may belong to more than one cluster.

We analyze Golub's Leukemia data (Golub et al., 1999) after standardizing the data and subsetting the data to include only those genes which are primarily

turned on across samples. For a robust cluster initialization, we use a model based clustering program, EMMIX (McLachlan et al., 1998). Based on these clusterings (at different values for number of clusters) we find the MCD for the given clusters and evaluate which points may be outlying and which may be part of a more than one cluster.

We discuss the merits and pitfalls associated with model based clustering methods applied to microarray data. Specifically, normalization and lack of independence can lead to misleading results. Microarray data is not well understood and needs much work on issues of normalization and standardization.

# 2 Data Thresholding and Standardization

The data we analyzed were oligonucleotides measured with Affymetrix software. The data are 38 samples of 7129 genes; some of the 38 samples have acute lymphoblastic leukemia (ALL) and some have acute myeloid leukemia (AML). In this work, we do not distinguish between the two types of leukemia because we are interested in clustering genes not in clustering samples. Because the microarray technology is still so new, there is not yet a single well tested method for standardizing the data before analysis. However, it is agreed that some type of standardization should be used on the data prior to analysis (e.g. "analysis of the normalized expression across all genes between samples..." (Bittner et al., 2000); "the purpose of normalization is to identify and remove systematic sources of variation..." (Dudoit et al., 2000b); "we also propose to truncate the log-ratios by a user-supplied constant M" (VanDerLaan and Bryan, 2000).) One of the difficulties of the technology is that the measurements at low expression levels are often misleading or inaccurate. A very low expression level means that the gene is absent for that sample, but there is no good way to distinguish between a low positive number or a large negative number when both essentially say that the gene is absent for that sample.

In our analyses, we set a threshold level and then standardized the data that were above the threshold level. Interest is only in genes which are in fact truly expressed. Given a sample, $i$, and a gene, $j$, the thresholding and

standardization algorithm is as follows. (This algorithm is due to Nguyen and Rocke, (Nguyen and Rocke, 2000).)

1. Let $A^{(o)}$ be the subset of genes with the lowest expression values. (Typically 10% of the genes.)
2. Let $m^{(o)} = \text{median } (A^{(o)})$
3. Let $\text{MAD}^{(o)} = \text{median } \{|x_j - m^{(o)}|, x_j \in A^{(o)}\}$
4. Let the threshold be $u^{(o)} = m^{(o)} + c \cdot s^{(o)}$ where $s^{(o)} = \text{MAD}^{(o)}/0.6745$ and $c = 2$. (Dividing MAD by 0.6745 makes it consistent for the standard deviation of a normal population.)
5. Let $A^{(1)} = \{x_j | x_j < u^{(o)}\}$
6. Repeat steps 2-5 until $A^{(k)} = A^{(k+1)}$
7. Keep genes such that 50% of the expression levels are above the threshold for that sample.
8. Rescale genes: $x_{ij} \leftarrow x_{ij} \cdot \overline{\overline{x_i}}/\overline{x_i}$ (where $\overline{\overline{x_i}}$ is the mean of all genes for sample $i$, and $\overline{x_i}$ is the mean of genes above the threshold for sample $i$.)

We then use the log of the new data, $\log x_{ij}$ as our transformed thresholded data. The new version of the data will be more elliptical and stable (because we are more confident in the values of the more highly expressed genes.)

# 3 Model Based Clustering

Many algorithms exist for clustering various types of data (Everitt, 1993). These algorithms use data, multivariate or univariate, as input, and as output the algorithm gives each datum a classification into a particular group. With microarray data, we can use such classifications to search for genetic pathways or groups of genes that might be regulated together. Some algorithms require that the number of clusters be pre-specified, and some algorithms allow for an unknown number of clusters. Those algorithms that do require as input a number of groups can be run multiple times with different values for the number of groups. The user can then choose the result that makes the most sense

according to the problem or according to some statistical criterion. Finding an appropriate criterion may prove to be a hard problem. For the method we use, finding the correct number of groups for a particular data set is beyond the scope of this work. Our methods are applicable to data with no a priori metric, so we restrict our work to partitioning clustering methods, and we disregard hierarchical clustering methods for the time being. These methods can be used to find a best fit to a problem with a given number of groups, or they can be applied to a problem by varying the parameter for number of groups.

Optimizational clustering methods use model based assumptions to derive different criteria which, when optimized, define a clustering of the dataset. Frequently used methods require that all points be assigned to a particular group. (See the clustering methods available in the software: mclust, kmeans, pam, clara, and fanny in S-Plus version 4.5 and proc cluster, proc fastclus, and proc varclus in SAS version 6.) Some methods, such as k-means, also use Euclidean distances or some other non-affine equivariant distance.

The clustering method we used, which will be described, assumes bivariate normal data, but the software can be set to any of a variety of distributions. (The outlier identification methods, however, treat the data as having some moments follow certain distributions.) Since the cluster shape is estimated from assigned points, it is required that $p+1$ points be assigned to each of the main clusters. However, this method allows for unassigned points, so there could easily be allowed a cluster of points which is smaller than $p+1$ included in the group of outlying points. The program is called EMMIX (McLachlan et al., 1998).

The optimization criterion used in this clustering algorithm is the maximization of the complete data log likelihood,

$$\log L_c = \sum_{l=1}^{g} \sum_{j=1}^{n} z_{lj} \log\{\pi_l \phi(x_j; \mu_l, \Sigma_l)\}$$

where

$$\pi_l = \text{probability } x_j \in \text{ cluster } l$$

$$z_{lj} = \begin{cases} 1 & x_j \text{ belongs to cluster } l \\ 0 & \text{else} \end{cases}$$

$\phi$ is the probability function associated with each cluster

and

$\mu_l$ and $\Sigma_l$ are the location and shape associated with cluster $l$.

The EMMIX procedure can be used with different functions for $\phi$, but in this work we take $\phi$ to be the normal distribution. The merits of this choice will be discussed further.

# 4 Outlier Detection and Overlapping Clusters

The EMMIX procedure gives a clustering configuration of the data, but it does not provide us with an idea of which points may be outlying or of which clusters may be overlapping. Using the EMMIX initialization to the clusterings, we can find the Minimum Covariance Determinant (MCD) of each cluster and apply distributional cutoffs to each MCD distance. Using distances provides us with an algebraic tool for measuring the distance of a point to the center of each of the clusters. These distances can be used to distinguish points that are in many clusters (small distances to many cluster centers) or are outlying (large distances to all cluster centers.) We use the MCD shape and location estimates because they are robust to a large proportion of outliers. The background for this procedure is given in Hardin and Rocke (2001a) and Hardin and Rocke (2001b).

## 4.1 Minimum Covariance Determinant

The Minimum Covariance Determinant (MCD) location and shape estimates are used as robust estimates of the location and shape of the clusters. Points that are outliers with respect to a particular cluster will not be involved in the location and shape calculations of that cluster, and points that are outliers with respect to all clusters will not be involved in the calculations of any clusters. The difference between the single population case and the multiple cluster case is that, in the latter, MCD samples need to be computed for each cluster. This

important difference leads to a need for a good robust starting point in the clustering situation.

## 4.2 Estimating the MCD

The exact MCD is impossible to find except in small samples or trivial cases. So, the algorithm used to estimate the MCD will be the estimator. The algorithm used in the multiple cluster case will be similar to the single population algorithm (Hawkins, 1999; Rousseeuw and VanDriessen, 1999) with the exception that the starting point of the algorithm will no longer be a random sub-sample of the data. The reason that it is important to have a non-random starting point for robust clustering is that random starts often give rise to shapes that are more representative of the entire data metric than the individual cluster metrics. Even with random samples of only $g \times (p+1)$ points (where $g$ is the number of clusters and $p$ is the dimension), it is highly unlikely that a random starting point would partition the points into their $g$ clusters respectively. From a starting point which reflects the entire data metric, it is difficult to separate the points into the correct $g$ clusters.

For a robust start, we used the program due to McLachlan et al. (1998). Their clustering algorithm uses three types of initializations (random partitioning, hierarchical clustering, and k-means clustering) and iteratively maximizes the log-likelihood associated with a normal mixture distribution (by using the EM algorithm.) From the different initializations, the output is the clustering structure that gives the maximum log-likelihood. The outlier detection methods described in this paper are not dependent on the particular robust clustering algorithm EMMIX. Any robust initialization would presumably give similar results. Random starts could be used if a condition was added to prevent the clusters from converging to the large dataset shape.

For each cluster $(j)$, a robust distance like $d^2_{S^*_j}(x_i, \overline{X}^*_j)$, where $S^*_j$ and $\overline{X}^*_j$ are the MCD shape and location estimates for cluster $j$, is likely to detect outliers because outlying points will not affect the MCD shape and location estimates. For points $x_i$ that are extreme, $d^2_{S^*_j}(x_i, \overline{X}^*_j)$ will be large for all $j$, and for points $x_i$ that are not extreme, $d^2_{S^*_j}(x_i, \overline{X}^*_j)$ will not be large for a particular $j$.

## 4.3 Single Population Distance Distributions

Mahalanobis squared distances give a one-dimensional measure of how far a point is from a location with respect to a shape. Using MSDs we can find points that are unusually far away from a location and call those points outlying. Unfortunately, using robust estimates gives MSDs with unknown distributional properties.

In Hardin and Rocke (2001), an approximate distributional result for MSDs based on location and shape derived from an MCD sample is given. Although the robust distances are asymptotically $\chi^2_p$, an F distribution fits the extreme points much more accurately across all sample sizes but especially in small samples. The distances based on the MCD metric can be expressed as,

$$\frac{c(m-p+1)}{pm} d^2_{S^*_X}(X_i, \overline{X}^*) \dot{\sim} F_{p,m-p+1}. \tag{1}$$

where $\overline{X}^*$ and $S^*_X$ are the location and shape estimates of the MCD sample, $p$ is the dimension of the sample, and $m$ and $c$ are both parameters based on the shape of the MCD sample. The unknown parameters, $m$ and $c$, can be estimated in three ways: using simulations, using an asymptotic result, or using an adjustment to the asymptotic result. The simulation results are the most accurate but also the most time consuming.

The parameter $c$ can be estimated by,

$$c = \frac{P(\chi^2_{p+2} < \chi^2_{(p,h/n)})}{\frac{h}{n}}$$

where $\chi^2_\nu$ is a Chi-Square random variable with $\nu$ degrees of freedom, and $\chi^2_{\nu,\epsilon}$ is the $\epsilon$ cutoff point for a $\chi^2_\nu$ random variable (Croux and Haesbroeck, 2000). We assume that the extreme points are asymptotically independent of the MCD estimates, and they do not enter into the calculations.

For $m$ there exists an asymptotic expression that is good in large samples and only moderately accurate in small samples (Croux and Haesbroeck, 2000). For small samples, an adjustment to the parameter is provided using a linear equation to estimate $m$ more accurately. The following interpolation formula is used to modify the theoretical parameter value of the degrees of freedom

(Hardin and Rocke, 2000a).

$$m_{pred} = m_{asy} \cdot e^{(0.725 - 0.00663p - 0.0780 \log(n))} \tag{2}$$

where $m_{pred}$ is the predicted degrees of freedom from adjusting the asymptotic degrees of freedom, $m_{asy}$, given by Croux and Haesbroeck (2000). Croux and Haesbroeck used influence functions to determine an asymptotic expression for the variance elements of the MCD sample. In determining our cutoffs in this work we used only the adjusted asymptotic degrees of freedom.

## 4.4 Multiple Population Distance Distributions

Using the same arguments from the single population setting in the cluster setting, an F distribution can be used to approximate distances which are large with respect to a cluster location and shape. However, in this setting there are new factors to consider such as how many points are in each cluster and whether extreme points simply belong to another cluster.

The number of points in a cluster is based on the initial clustering, and the MCD is found for each cluster. For cluster $j$, $\overline{X}_j^*$ and $S_j^*$ are the MCD mean and covariance. The distances $d_{ij} = d_{S_j^*}(X_i, \overline{X}_j^*)$ are the distances for each point, $i$, to each cluster, $j$. The same method as in the one cluster case is used to find the cluserwise degrees of freedom needed in the F distribution.

Consider $g$ groups of multivariate normal data in dimension $p$, and let $X_{ij} \sim N_p(\mu_j, \Sigma_j)$ where $i$ =gene and $j$ =cluster. Let $S_j$ be an estimate of $\Sigma_j$ such that, $m_j S_j \dot{\sim} \text{Wishart}_p(\Sigma_j, m_j)$. For the multiple cluster case,

$$c_j = \frac{P(\chi^2_{p+2} < \chi^2_{p,h_j/n_j})}{h_j/n_j},$$

and $m_j$ can be estimated from theoretical formulas or simulation. The distribution of the clusterwise MCD distances can be then approximated by:

$$\frac{c_j(m_j - p + 1)}{p m_j} d^2_{S_j^*}(X_i, \overline{X}_j^*) \dot{\sim} F_{p, m_j - p + 1}.$$

With these constructs in mind, the distances of interest are those associated with the cluster to which a point is closest. Let $\tilde{d}_i$ be the distance from point $x_i$ to the closest cluster. An outlying point, $x_i$, will be one with $\tilde{d}_i$ greater than some cutoff value.

### 4.5 Results of Clusterings

As discussed previously, the distances can be used to find the extent of overlapping clusters as well as genes which might be outlying. Figure 1 displays histograms of the distribution of the gene allocation the closest cluster center. We see that for a small number of clusters, most of the genes fall into one particular cluster with very few outliers (very few genes in cluster 0.) The clustering for 10 clusters spreads the genes out more evenly across clusters.

Though a gene might be closest to one particular cluster, it may also be within the bounds of many of the other clusters. Figure 2 gives an idea of the amount of overlap in the clustering structure. We see that for no clustering is there a complete distinction in the groups. For the grouping of the genes into 8 clusters, we see that most of the genes are in either 4 or 5 clusters. This leads us ask the following questions:

- Are the clusters of data really truly overlapping?
- Should we be searching for many more than 10 clusters because of the large number of genes?
- Are the assumptions being violated?

## 5 Possible Pitfalls in Model Based Clustering

From the last section we see that the distances do not give us the groups or discriminating tools for which we had hoped. There was significant overlapping of the clusters, and there was not a clear value for the correct number of clusters. This leads us to question what is happening with the data and the procedure in the context of this clustering problem. The insight below will help us to understand this type of data in the context of this and future problems.

## 5.1 Overlapping Clusters

It is possible that the clusters are indeed overlapping. It makes sense that the biological structure to the genes forces them to be correlated and regulated together. Often times a cancerous agent may cause a gene to be wildly variable with a similar mean to that in a non-cancerous sample. This might cause a cluster with large variability which overlaps a similar cluster with small variability. To test the validity of the overlapping clusters, we plotted the clusters onto projections that would most separate the clusters.

The dimension of our data is 38, and so it is impossible to visualize the clusterings in a 38-dimensional space. In order to visualize our clusterings, we can project them onto a space that should visually be the most separated groups of clusters. With two clusters, the projection is onto the line that connects the two cluster centers. (See figure 3.) With three clusters, the projection is onto the line that connects two of the cluster centers and also onto a line that is perpendicular to the first line and goes through the third cluster center. (See figure 4.) Again, we see that the clusters show quite a bit of overlap, though the variability is different and would be interesting to analyze.

## 5.2 Many small clusters

The clustering procedures were applied for 2 - 10 clusters. With over 7000 genes collected as data, it is possible that there are many small clusters. If the data do not have any cluster structure (i.e., the data are simply an ellipse or some other shape), we should see pairwise Mahalanobis squared distances plotted against $\chi^2_{38}$ quantiles as a smooth curve. If however, there are many distinct small clusters, the distances within a small cluster should be much smaller than the distances between clusters. This would manifest itself by having breaks in the curve of the pairwise distances vs. $\chi^2_{38}$ quantiles .

Figure 5 shows a plot of all the pairwise distances vs. $\chi^2_{38}$ quantiles. The plot appears to be quite smooth, but may be confounded by the large number of points (499,500 points.) Figure 6 shows a magnification of figure 5 for only the closest 10,000 pairs. Again, there are no discernible breaks indicating small

clusters. Though we will not at this point rule out the possibility of small clusters, these plots show that the genes are probably not separated into many distinct small clusters.

## 5.3 Normality

Our methods (and many other methods which have been applied to this type of data in the literature (Lee et al., 2000; Tanaka et al., 2000; VanDerLaan and Bryan, 2000; Dudoit et al., 2000a; Dudoit et al., 2000b)) make a normality assumption about the data. We know that Mahalanobis squared distances with the mean and covariance as the location and shape parameters have a chi-square distribution when the data are normal (Mardia et al., 1979). Figure 7 shows a plot of the Mahalanobis distances of those genes which are above the predetermined threshold versus the quantiles of a chi-square(38) distribution. It appears as though the data have very heavy tails and possibly a non-elliptical distribution. This implies that the user should be careful using formal statistical tests that are based on normality and on normal theory tests for number of clusters.

The departure from normality can also be seen in bivariate plots of two samples at a time. Figure 8 shows 12 samples (some from the ALL group, some from the AML group) with their gene expression values plotted against each other. Multivariate normality exists only when the individual univariate components are also distributed normally. Here we see a very skewed distribution which is obviously not normal in any of the univariate elements. We also see a strong positive correlation and a two cluster structure separating out groups of genes that are mostly high expressed and those that are mostly low expressed. Figure 9 shows the same plot adjusted for the positive correlation. By subtracting out the row means, we remove the cluster structure from before, and we see no other obvious cluster structure in these particular projections. We also see univariate structure that is closer to normality which leads us to suggest that analysis on row zero adjusted data might be a worthwhile endeavor.

# 6 Advantages to Model Based Clustering

With the visual clues from the previous section that there is some structure in the data that distinguishes between high expressed genes and low expressed genes, we investigated whether or not our clustering algorithm had done that. We choose eleven as a suitable number to differentiate between a "high" expression versus a "low" expression. Eleven is somewhat arbitrary because the units of the data are not well defined. However, eleven seems to do a good job at discriminating in our bivariate plots. Again, if we had chosen to use the average difference call data from Affymetrix this number would be changed.

Table 1(a)-(i) gives the number of genes in a particular cluster with a mean log expression above eleven and the number below eleven, based on the EMMIX clustering. We can see that there are very few groups which contain genes which are both strongly and weakly expressed. So, the clustering does accurately distinguish the two groups which are fairly certain to exist.

# 7 Conclusion

Software and algorithms will almost always output a cluster structure to the data it is given. The statistician and the data analyst need to be careful in interpreting what the software has output. How do we know if in fact there are any clusters? How will we know when we have found real clusters? In this paper we have addressed some concerns about clustering: are the assumptions being violated? are the clusters overlapping? are there points which should be labeled as outlying but instead are being forced into the cluster structure? Beyond the ideas discussed above, future concerns about this type of microarray data gene clustering are:

1. The dependence structure between the genes. (This might be avoided by removing one half of a pair of genes which are highly correlated, running the clustering algorithm, and then applying it to the whole data set.)

2. The standardization of the data. (This is an unresolved topic present in the literature which will hopefully be addressed successfully soon. There

needs to be a standard way to characterize and measure gene expressions.)

When the issues brought up in this paper are adequately addressed, the methods described above will be invaluable in measuring both the underlying overlap of clusters and points which do not seem to belong to clusters. Both of these structures will give insight into the genomic make up of the data.

# 8 Figures

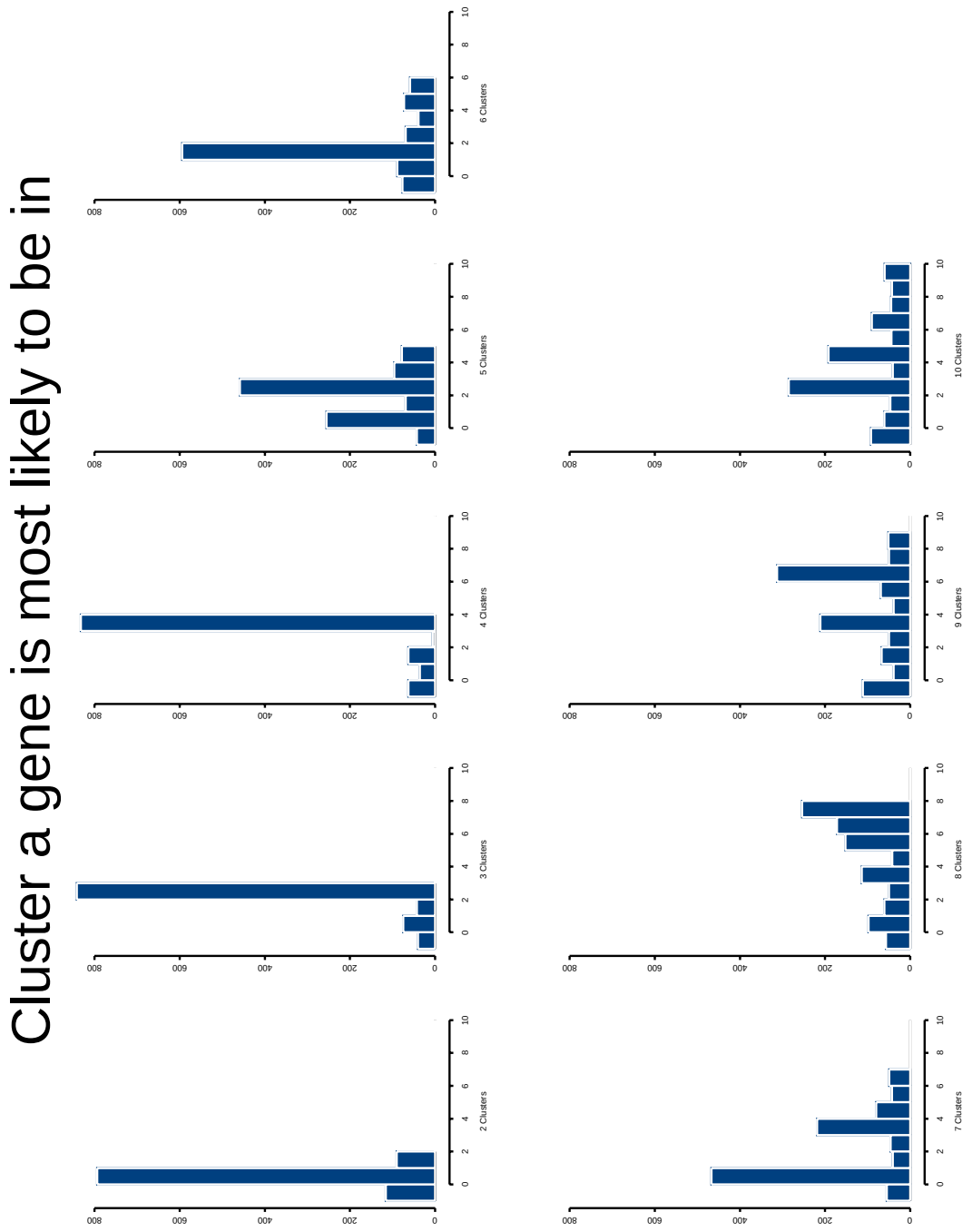


Figure 1: *Each bar represents the number of genes closest to one cluster center, cluster 1-k (where k is the assumed number of groups), or in cluster 0 which is the "group" of outlying points.*

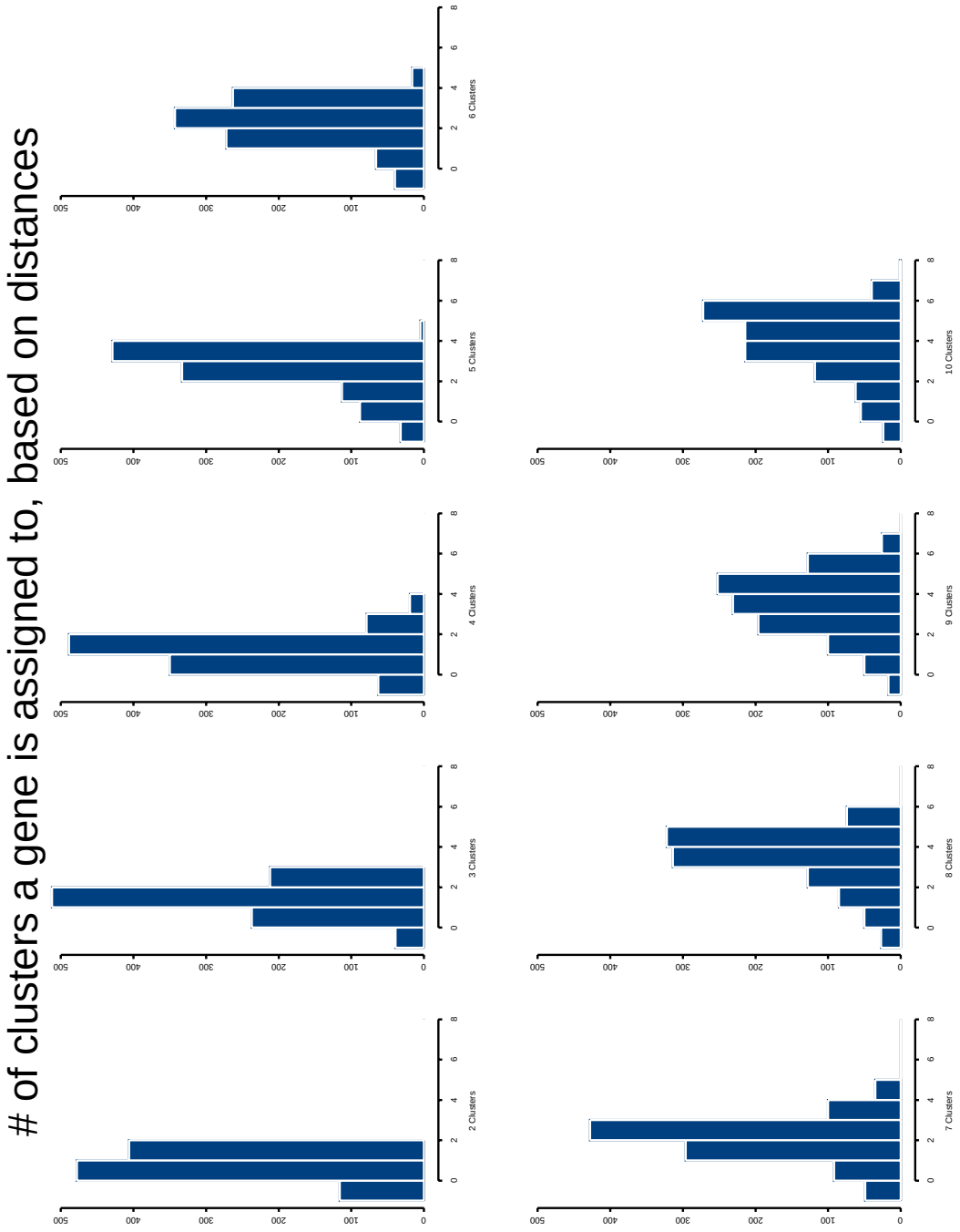


Figure 2: *Based on F-distances calculated at a 5% cutoff level, each bar represents the number of genes allocated to 0,1,2,...k clusters. A point that is allocated to 3 clusters gives an indication that there is some overlapping of the clusters.*

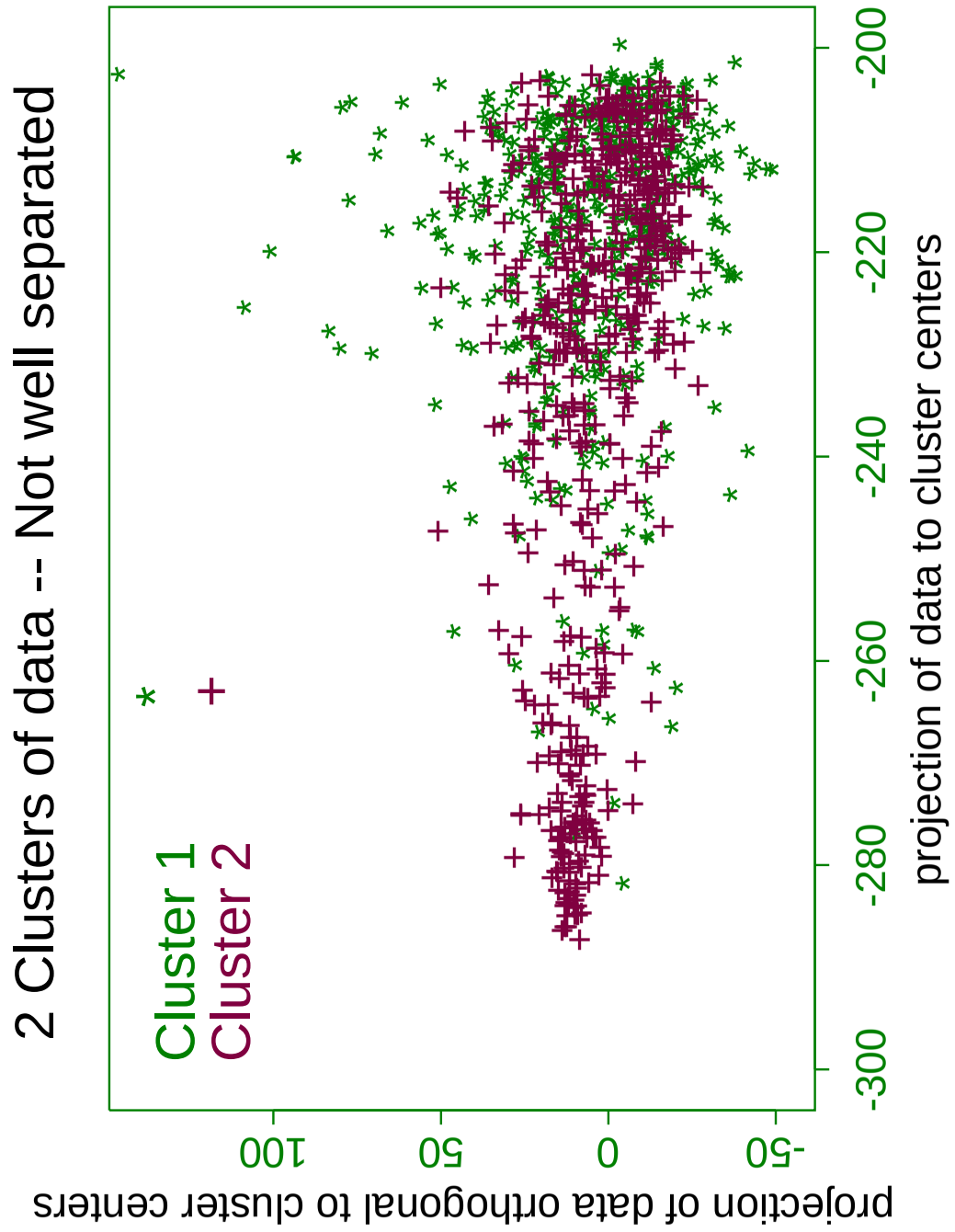


Figure 3: *2 clusters of data projected onto the line that connects the cluster centers (x-axis.) These two clusters were found from the* EMMIX *procedure, and are obviously not well separated, though they do seem to have different variance structures.*

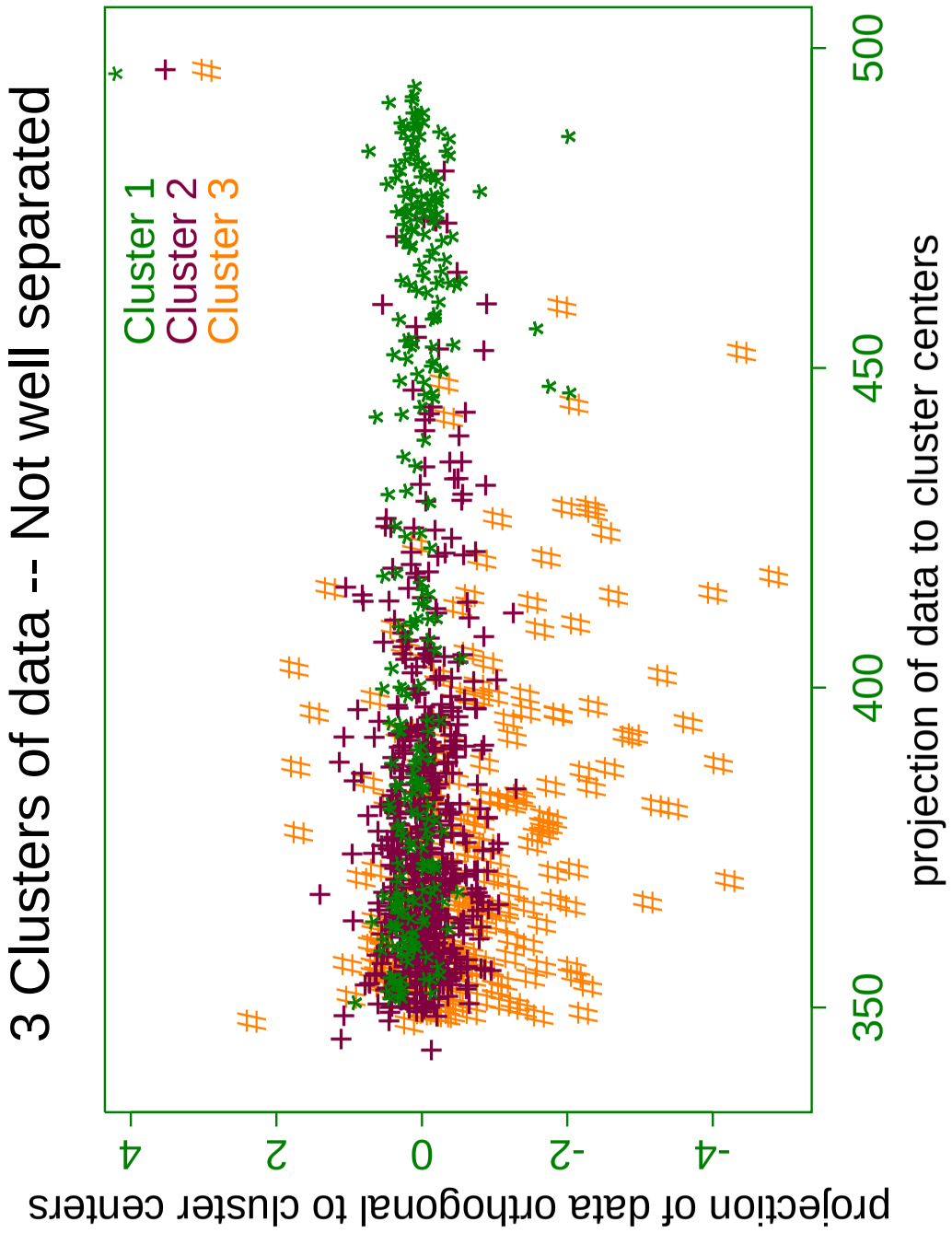


Figure 4: *Three clusters of data projected onto the line that connects two of the centers (x-axis) and also onto a line that is perpendicular to the first line and connects to the third cluster center (y-axis.) Again, the clusters were found from the* EMMIX *procedure and are not well separated. Cluster 3 seems to have much wider variability than the other two clusters.*

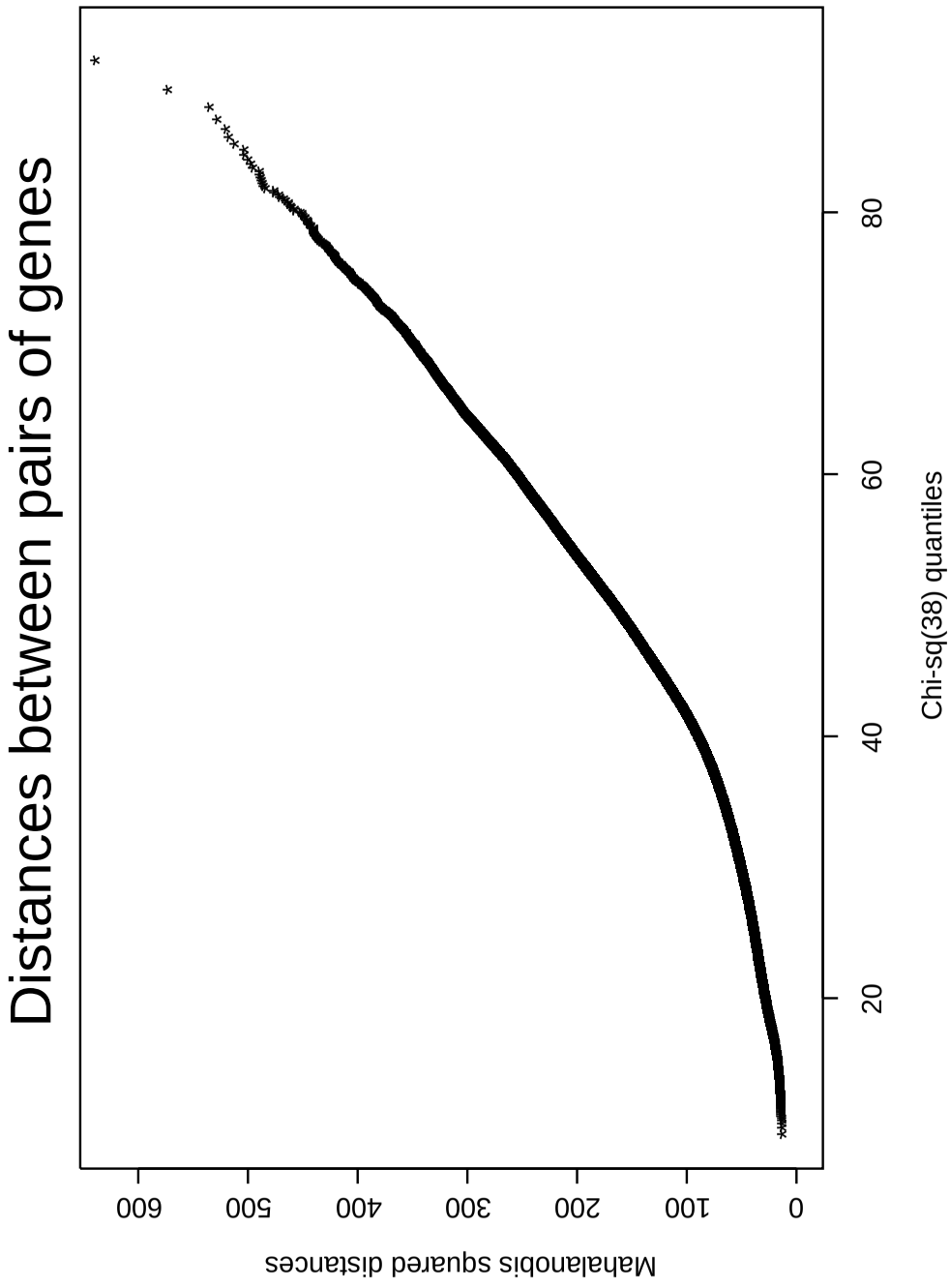


Figure 5: *All pairwise distances between pairs of genes vs. the quantiles of a* $\chi^2_{38}$ *distribution.*

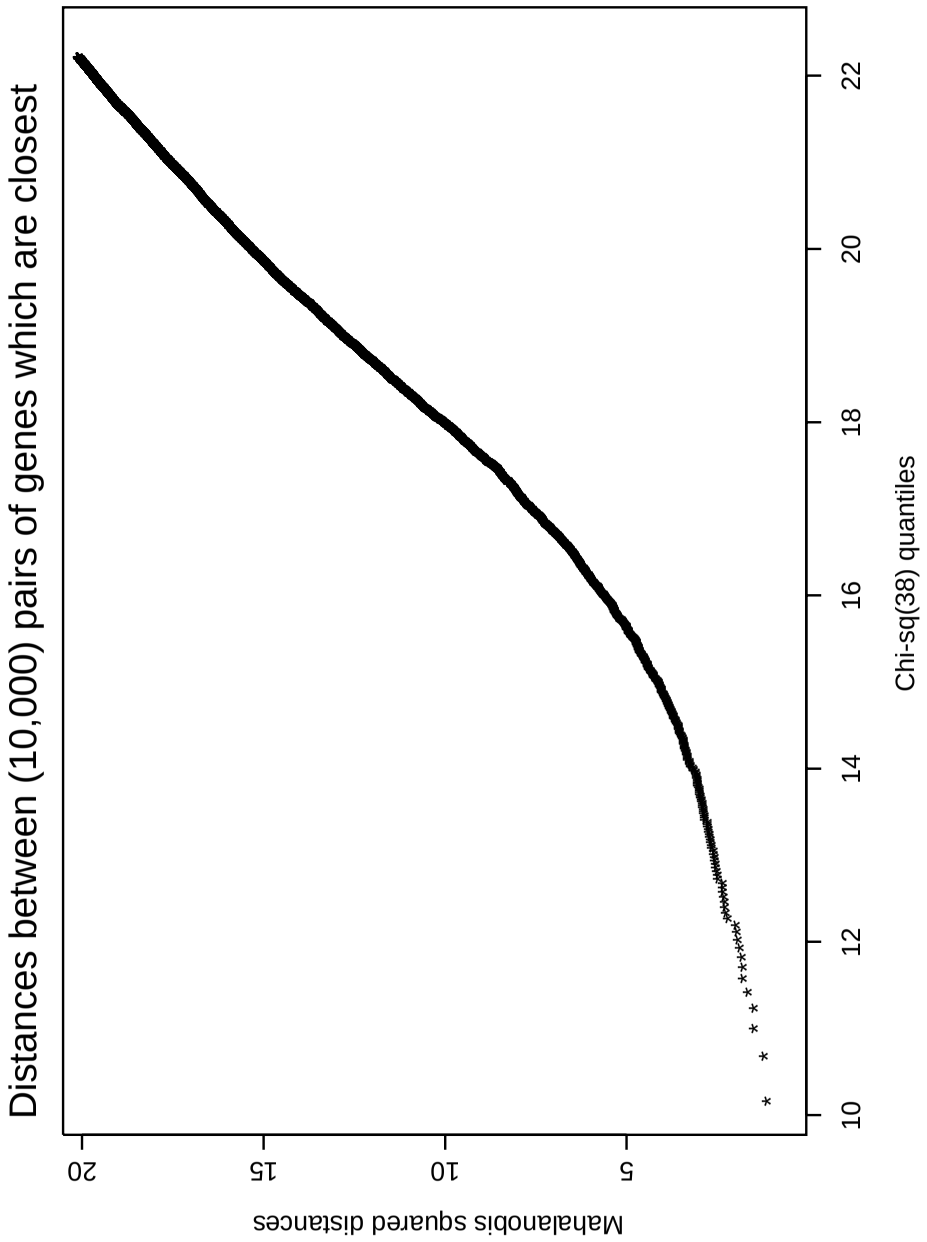


Figure 6: *The closest 10,000 pairwise distances between pairs of genes vs. the quantiles of a* $\chi^2_{38}$ *distribution.*

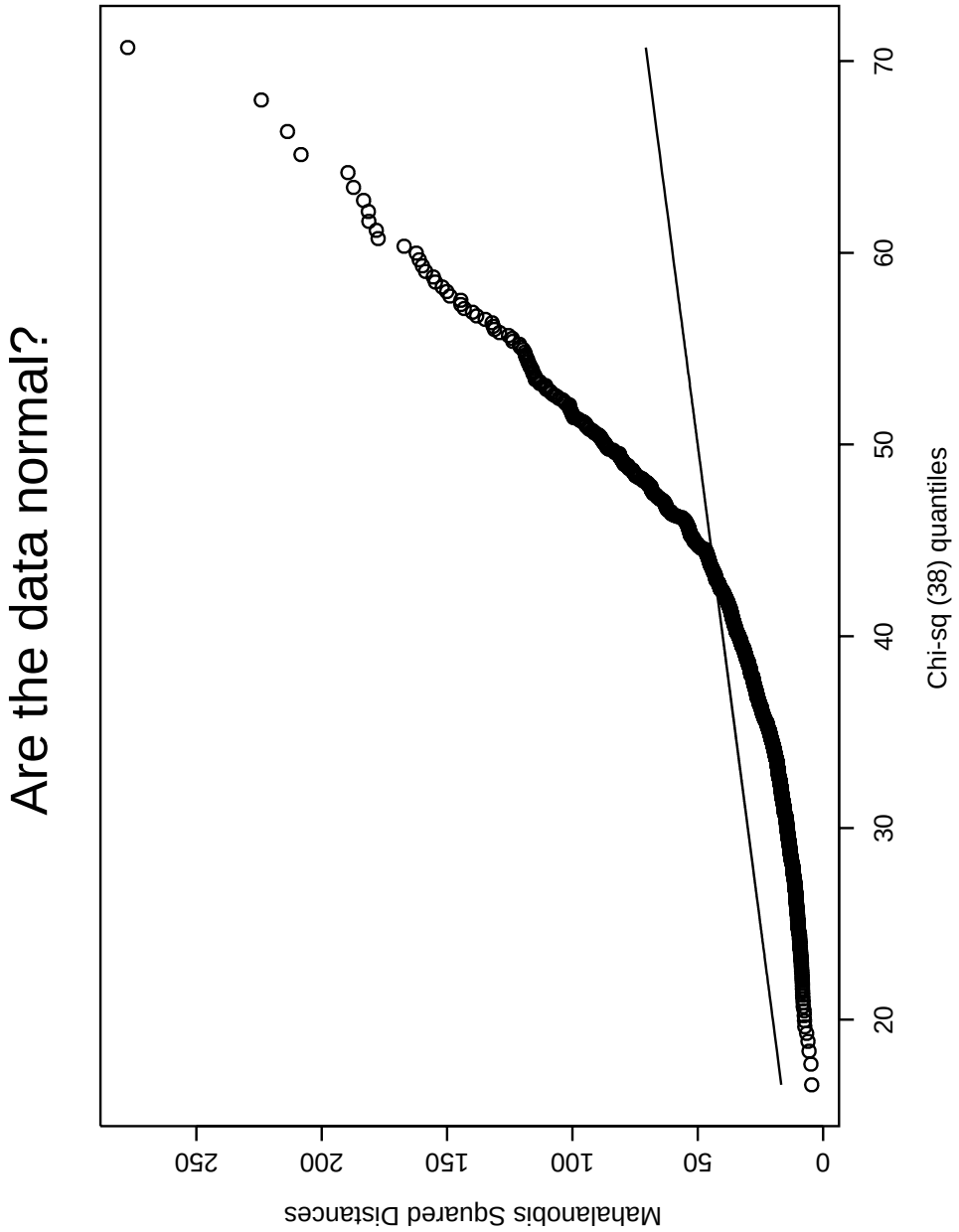


Figure 7: *Plot of the Mahalanobis squared distances of the genes versus* $\chi^2_{38}$ *quantiles. If the data had a normal distribution, they would follow the straight line.*

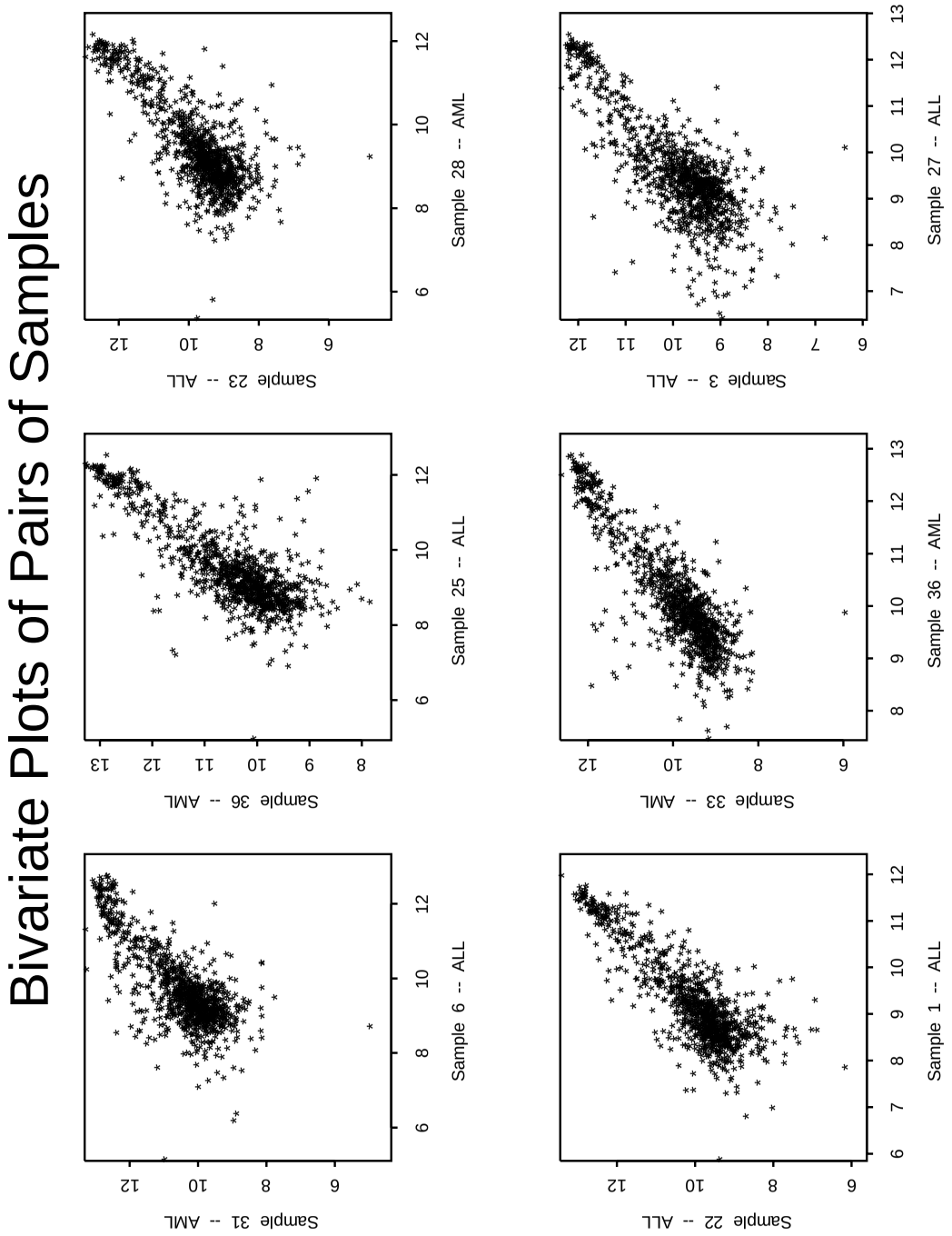


Figure 8: *Random selection of 12 samples plotted against each other. The samples appear to have a positive relationship and are possibly separated into two clusters (a low expression group and a high expression group.)*

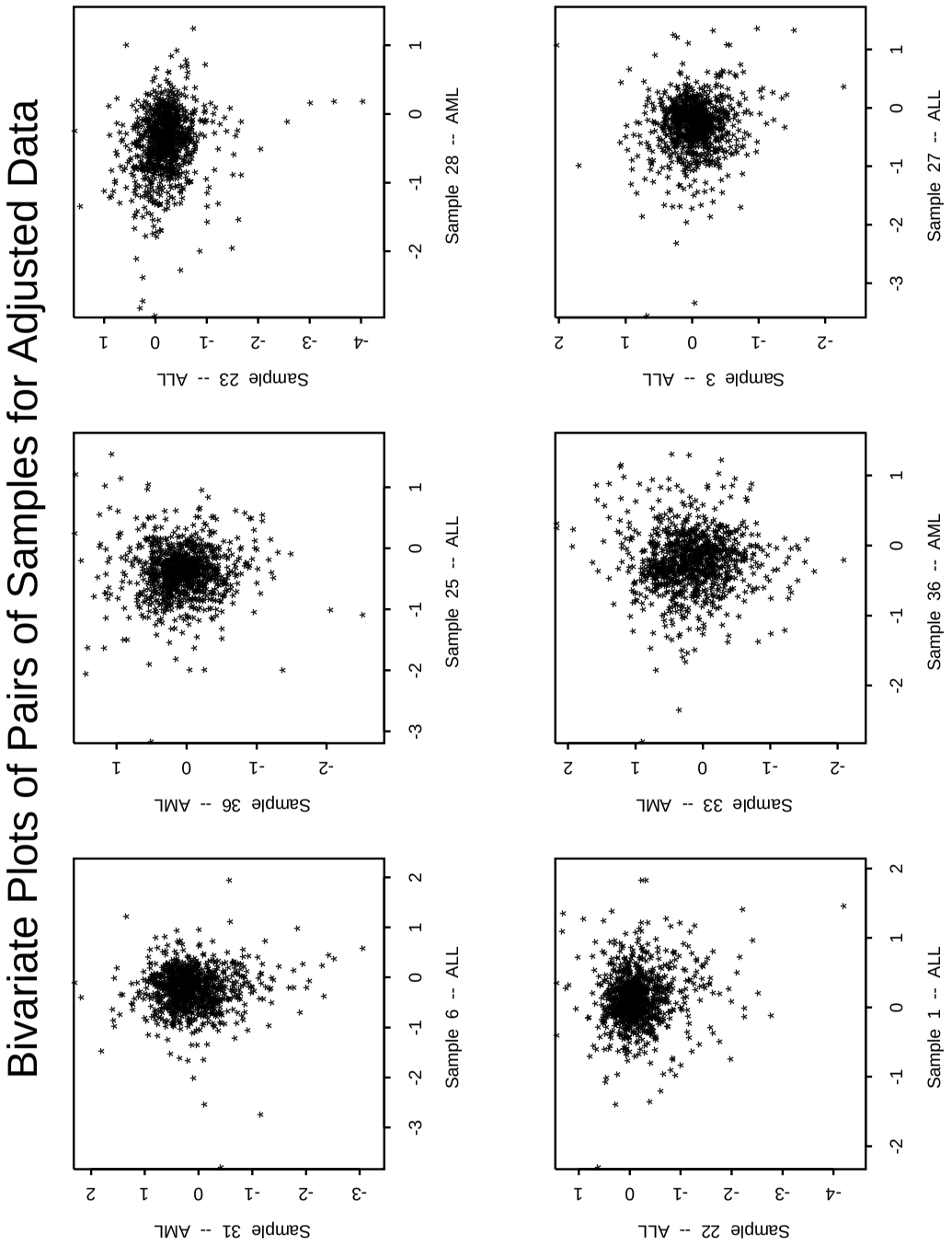


Figure 9: *Same random selection of 12 samples plotted against each other and adjusted for the positive correlation structure. (Accounted for the difference in two groups "all high expression" versus "all low expression.") Now we see no visible cluster structure from these particular projections.*

# 9 Tables

1. (a)

| Two Clusters | | |
|---|---|---|
| | 0 – 11 | 11+ |
| 0 | 19 | 2 |
| 1 | 441 | 17 |
| 2 | 396 | 125 |

(b)

| Three Clusters | | |
|---|---|---|
| | 0 – 11 | 11+ |
| 0 | 10 | 3 |
| 1 | 142 | 117 |
| 2 | 493 | 19 |
| 3 | 211 | 5 |

(c)

| Four Clusters | | |
|---|---|---|
| | 0 – 11 | 11+ |
| 0 | 10 | 0 |
| 1 | 1 | 39 |
| 2 | 9 | 105 |
| 3 | 530 | 0 |
| 4 | 306 | 0 |

(d)

| Five Clusters | | |
|---|---|---|
| | 0 – 11 | 11+ |
| 0 | 9 | 0 |
| 1 | 199 | 0 |
| 2 | 0 | 118 |
| 3 | 121 | 0 |
| 4 | 104 | 0 |
| 5 | 423 | 26 |

(e)

| Six Clusters | | |
|---|---|---|
| | 0 – 11 | 11+ |
| 0 | 8 | 0 |
| 1 | 471 | 2 |
| 2 | 229 | 0 |
| 3 | 0 | 104 |
| 4 | 39 | 0 |
| 5 | 75 | 0 |
| 6 | 34 | 38 |

(f)

| Seven Clusters | | |
|---|---|---|
| | 0 – 11 | 11+ |
| 0 | 13 | 0 |
| 1 | 158 | 0 |
| 2 | 44 | 0 |
| 3 | 15 | 38 |
| 4 | 157 | 0 |
| 5 | 469 | 0 |
| 6 | 0 | 47 |
| 7 | 0 | 59 |

(g)

| Eight Clusters | | |
|---|---|---|
| | 0 – 11 | 11+ |
| 0 | 5 | 0 |
| 1 | 77 | 0 |
| 2 | 0 | 85 |
| 3 | 0 | 59 |
| 4 | 70 | 0 |
| 5 | 45 | 0 |
| 6 | 105 | 0 |
| 7 | 106 | 0 |
| 8 | 448 | 0 |

(h)

| Nine Clusters | | |
|---|---|---|
| | 0 – 11 | 11+ |
| 0 | 6 | 0 |
| 1 | 39 | 0 |
| 2 | 0 | 98 |
| 3 | 54 | 0 |
| 4 | 96 | 0 |
| 5 | 39 | 0 |
| 6 | 62 | 0 |
| 7 | 487 | 0 |
| 8 | 18 | 46 |
| 9 | 55 | 0 |

(i)

| Ten Clusters | | |
|---|---|---|
| | 0 – 11 | 11+ |
| 0 | 1 | 0 |
| 1 | 61 | 0 |
| 2 | 49 | 0 |
| 3 | 462 | 0 |
| 4 | 44 | 0 |
| 5 | 97 | 0 |
| 6 | 6 | 44 |
| 7 | 76 | 0 |
| 8 | 0 | 52 |
| 9 | 0 | 48 |
| 10 | 60 | 0 |

Table 1: **Separation of High and Low Expression:** Based on the EMMIX clustering, the number of genes in each cluster which have a mean expression less than 11 or more than 11. I.e., for the clustering of four groups, in cluster two there were 114 genes. Nine of the genes had a mean log expression (across samples) less than eleven, 105 of the genes had a mean log expression (across samples) larger than eleven.

# References


Atkinson, A. (1994). Fast very robust methods for the detection of multiple outliers. *Journal of the American Statistical Association*, 89:1329–1339.

Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data*. John Wiley.

Bittner, M. et al. (2000). Molecular classification of cutaneous malignant melanoma by gene expression profiling. *nature*, 406:536–540.

Croux, C. and Haesbroeck, G. (2000). Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. *Journal of Multivariate Analysis*.

Dudoit, S., Fridlyand, J., and Speed, T. (2000a). Comparison of discrimination methods for the classification of tumors using gene expression data. Technical report, University of California, Berkeley. #576.

Dudoit, S., Yang, Y., Callow, M., and Speed, T. (2000b). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Technical report, University of California, Berkeley. #578.

Everitt, B. S. (1993). *Cluster Analysis*. John Wiley.

Gnanadesikan, R. and Kettenring, J. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, 28:81–124.

Golub, T. et al. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537.

Hadi, A. (1992). Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society B*, 54:761–771.

Hardin, J. and Rocke, D. (2000a). The distribution of robust distances. submitted.

Hardin, J. and Rocke, D. (2000b). Outlier detection in multiple cluster setting using the minimum covariance determinant. submitted.

Hastie, T., Tibshirani, R., Eisen, M., Brown, P., et al. (2000). Gene shaving: a new class of clustering methods for expression arrays. Technical report, Stanford University.

Hawkins, D. (1980). *Identification of Outliers*. Chapman and Hall.

Hawkins, D. (1999). Improved feasible solution algorithms for high breakdown estimation. *Computational Statistics and Data Analysis*, 30:1–11.

Lee, M.-L. T., Kuo, F., Whitmore, G., and Sklar, J. (2000). Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proceedings of the National Academy of Science*, pages 9834–9839.

Lopuhaä, H. and Rousseeuw, P. (1991). Breakdown of points of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics*, 19:229–248.

Mardia, K., Kent, J., and Bibby, J. (1979). *Multivariate Analysis*. Academic Press.

Maronna, R. and Yohai, V. (1995). The behavior of the Stahel-Donoho robust multivariate estimator. *Journal of the American Statistical Association*, 90:330–341.

McLachlan, G., Peel, D., and Basford, K. (1998). The emmix software for the fitting of mixtures of normal and t-components. Technical report, University of Queensland. http://www.maths.uq.oz.au/ gjm/emmix/emmix.html.

Nguyen, D. and Rocke, D. (2000). Classification of acute leukemia by partial least squares using gene expression data. Technical report, University of California, Davis.

Penny, K. (1995). Appropriate critical values when testing for a single multivariate outlier by using the mahalanobis distance. *Applied Statistics*, 45:73–81.

Rocke, D. and Woodruff, D. (1996). Identification of outliers in multivariate data. *Journal of the American Statistical Association*, 91:1047–1061.

Rousseeuw, P. (1984). Least median of squares regression. *Journal of the American Statistical Association*, 79:871–880.

Rousseeuw, P. and Leroy, A. (1987). *Robust Regression and Outlier Detection*. John Wiley.

Rousseeuw, P. and VanDriessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, pages 212–223.

Rousseeuw, P. and VanZomeren, B. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85:633–639.

Tanaka, T. et al. (2000). Genome-wide expression profiling of midgestation placenta and embryo using a 15,000 mouse developmental cDNA microarray. *Proceedings of the National Academy of Science*, pages 9127–9132.

VanDerLaan, M. and Bryan, J. (2000). Gene expression analysis with the parametirc bootstrap. *Biostatistics*, pages 1–24.