Thesis

Nicholas Conway

May 7, 2008

1 Background

1.1 DNA Microarrays

DNA microarrays are a technique and technology used to study the activation and expression of genes. Scientists are often interested in when certain genes are expressed, under what conditions and in which particular types of cells. Instead of looking at only a few genes at a time, microarrays allow scientists to look at thousands of genes simultaneously. Scientists can then see which genes are expressed under which circumstances. Scientists working with microarrays are often interested in the connections between different genes. Because genes typically do not work alone, but instead in tandem with many other genes, understanding the connections between genes helps scientists to build a picture of the function of a network of genes. The hope is to use this new understanding to develop applications, including therapies for diseases, notably cancer.

DNA microarrays consist of a large number, typically in the hundreds or thousands, of genes layed out in rows and columns on small supports, which are typically made of glass, but sometimes out of other materials. Messenger RNA, or mRNA, is gathered from two samples to be compared. mRNA codes for proteins in cells, and as such is often of intense interest to scientists. The mRNA to be examined is processed chemically, with different methods used depending on the particulars of the experiment. Either the mRNA itself or cDNA derived from it has chemical markers attached, which allows scientists to see the degree to which the mRNA has bound to the gene sequences, which in turn allows them to study which genes are actively coding for proteins under various circumstances. This knowledge can be extremely useful in seeking to gain an understanding of the mechanics of the cell, especailly with regards to the study of cancer.

The process for preparing DNA microarrays is costly and often difficult. The result is that microarray experiments often have a small number of trials. Furthermore, as the process is imperfect, data sets derived from micoarray experiments often have a substantial portion of bad data points. DNA microarrays are known to have a wide variety of problems in the consistency of their data. Expression is measured through the coloring of samples on the array. Quantified values are know to have biases due to problems with the dye used to show expression, pixel saturation, and computer image analysis, all of which can lead to inconsistencies in the data. Other problems are caused by hard to control laboratory conditions, as well as a low signal/noise ratio. All of these factors contribute to problems for data analysis. If meaningful conclusions are to be drawn from this type of data, it is necessary to use statistics that take into account the inherent noise in the data.

1.2 Distance

Genes which are often expressed together can be be related in function. Since scientists using microarrays are often interested in these connections, it is important to develop a good measure of the similarity of two genes on a microarray. One way of looking at this problem is to come up with a measure of distance between genes. Genes separated by a large distance would then be considered dissimilar and those separated by a small distance would be considered similar. This would then allow a scientist to conduct further studies on those genes marked as similar by our measure of distance. This is especially useful for clustering genes into groups.

Definition: For a sample a set X, a **distance function** is a function d which satisfies the following:

For $x, y, z \in X$,

$$d(x, y) \geq 0$$
$$d(x, y) = 0 \Longleftrightarrow x = y$$
$$d(x, y) = d(y, x)$$
$$d(x, y) \leq d(x, z) + d(z, y)$$

The last condition is known as the triangle inequality.

Sometimes we have a measure of similarity rather than distance. It is, however, simple enough to move between our ideas of similarity and distance. Our correlations will fall in the range $[-1, 1]$, so if we take the absolute value

of our correlation and subtract it from 1, we can arrive at a measure of distance. Likewise, we can subtract our measure of distance from 1 and end up with a measure of similarity. If we have a similarity s we can find a measure of distance d by

$$d = 1 - s$$

There are several familiar functions which can be used as measures of similarity or distance. Pearson's correlation is one of the most common. Others include the jackknife correlation[2], the Spearman rank correlation and Euclidean distance. The correlations are measures of similarity, unlike the Euclidean distance.

Although we treat the measures of distance derived from these correlations as rough measures of distance for the purpose of clustering, they are not (with the exception of Euclidean distance) actually distance metrics since they do not always satisfy the triangle inequality. Since the primary concern is whether there is a close correlation and not the direction of that correlation, we can deal with the absolute value of the correlation. Despite these problems, these correlations are useful for determining relatedness even if they cannot strictly be called a distance.

1.3 Robustness and Resistance

If we desire to cluster our data, we must first decided which correlation we wish to use. The trick is to pick a good measure of distance, one which reflects the problems often present in microarray data. Given the nature of microarray experiments, there is often only a small sample size and data sets often contain a significant proportion of outlying data. Particularly important is that our measure of distance not be overly influenced by outliers. A measure of distance which is not affected when a portion of the data it is drawn from is replaced with corrupt data is known as resistant. Statistics which are not unduly influenced by outliers are known as robust. There are a number of ways of measuring robustness, but one of particular concern to us is the breakdown bound. The breakdown bound is the proportion of data points which we can change without restriction while still seeing a bound on the change of the estimate.

For correlations, there is always a bound on the change of the estimate, since correlations fall in the range [-1,1]. Thus when talking about the robustness of correlations it makes more sense to think about breakdown in slightly different terms. When we talk about breakdown for a correlation, we refer to the proportion of data points which we can change without re-

striction while still seeing a bound on how close the estimate can come to 1 or negative 1. The most commonly know correlation, Pearson correlation has a breakdown bound of 0, because we can make the Pearson correlation arbitrarily close to 1 or -1 simply by changing one data value, and so is not a robust measure. We will look into this in more detail in section 2.5. In light of the Pearson correlation's lack of robustness, we set out to find a more robust distance/similarity metric for our analysis of DNA microarray data.

2 M-estimators

2.1 Definition of an M-estimator of location

An M-estimator is a statistic which is produced through the minimization of some objective fuction. M-estimates can be defined as follows [3]:

Definition: For a sample $(x_1, ..., x_n)$ and a function ρ, the M-estimate T_n is the value of t which minimizes

$$\sum_{i=1}^{n} \rho(x_i, t).$$

Since we are looking at the minimization of a function, if we know the derivative of that function it often easier to deal with the derivative. In this case we can find T_n by solving the following equation for t:

$$\sum_{i=1}^{n} \psi(x_i, t) = 0,$$

where $c\psi$ is the derivative of ρ with respect to t where c is a constant. We are not particularly concerned with c, as it will divide through and be absorbed in the 0 on the right side of the equation and thus not affect the solution of the equation.

2.2 A Basic Example - The Mean

The mean is an M-estimator with objective function

$$\rho(x, t) = (x - t)^2.$$

To find the M-estimator of this objective function, we minimize with respect to t

$$\sum_{i=1}^{n} \rho(x_i, t) = \sum_{i=1}^{n} (x_i - t)^2.$$

This is best achieved finding the ψ-function for our ρ-function.

$$\frac{d}{dt}(x-t)^2 = 2(x-t).$$

If we drop the constant 2, we get

$$\psi(x, t) = (x - t).$$

Thus if we want to find the M-estimator associated with this ψ-function we would solve

$$\sum_{i=1}^{n} (x_i - t) = 0$$

We can now see that our M-estimator is indeed the mean. Solving the above equation we see that

$$\sum_{i=1}^{n} (x_i - t) = 0 => \sum_{i=1}^{n} x_i = tn => (1/n) \sum_{i=1}^{n} x_i = t.$$

Thus $T_n = (1/n) \sum_{i=1}^{n} x_i$, the sample mean. The mean is an M-estimator, but it is not resistant to outliers. If a very large outlier were added to the sample space the mean would change drastically. If we are worried about this effect we should look for a more resistant M-estimator. But in order to undertake that search we first need a better idea of what we are looking for. What we need is to define some criteria by which we evaluate M-estimators.

2.3 Criteria for Evaluating an M-estimator

One such criteria is the breakdown bound. The breakdown bound as defined in Hoaglin, Mosteller and Tukey [3] is:

> The largest possible fraction of the observations for which there is a bound on the change in the estimate when that fraction of the sample is altered without restriction.

Under this definition, the breakdown bound of the mean is 0. If we alter even one observation without restriction, there is no bound on how large or small we can make the mean. Say we have a sample $(y, x_2, ..., x_n)$. Then let a number C be given. Let us replace y with

$$x_1 = n(C - \sum_{i=2}^{n} x_i/n)$$

Then we can see that

$$\sum_{i=1}^{n} x_i/n = x_1/n + \sum_{i=2}^{n} x_i/n = n(C - \sum_{i=2}^{n} x_i/n)/n + \sum_{i=2}^{n} x_i/n =$$

$$C - \sum_{i=2}^{n} x_i/n + \sum_{i=2}^{n} x_i/n = C \geq C$$

Thus we can make the mean as large as we like by changing only one observation. Clearly if we are concerned about the effects of outliers on our estimate, the mean is not a good choice, and we can see this from our analysis of the breakdown bound.

Another M-estimator is the median. The ρ-function for the median is

$$\rho(x, t) = |x - t|$$

Although the absolute value function does not have a derivative at 0, if we treat it as if the derivative was 0 at 0, we can find a workable ψ-function. With this in mind, we can say that

$$\psi(x, t) = sgn(x - t)$$

where

$$sgn(u) = \begin{cases} 1 & u > 0 \\ -1 & u < 0 \\ 0 & u = 0 \end{cases}$$

With the median we can see that there is a very large breakdown bound, since only the middle element, or two middle elements in the case where the size of the sample is even, has an effect on the median. In order to make the median as large or as small as we want, we would have to replace nearly half of the elements of the sample.

2.4 Tukey's Biweight

A much better choice for robust estimation than the median or the mean is the biweight. Tukey's biweight refers to a class of estimators with the objective function,

$$\rho(u) = \begin{cases} \frac{c}{6}(1-(1-(\frac{u}{c})^2)^3) & |u| \leq c \\ \frac{1}{6} & |u| \geq c \end{cases}$$

where c is a constant chosen depending on our requirement for robustness [3]. The value of c controls how far from the center of the data a data point need be before it is treated as an outlier. The greater we make c, the farther away from the center of the data the point greater than which data points will have no effect, known as the rejection point, is set. Data points for which $|u| \geq c$ are weighted to zero, which means they have no effect on the calculation of the estimate. The setting of c gives the data analyst a way to control the robustness of the estimate.

2.5 Robustness of the Pearson correlation

In order to talk about the breakdown bound of a Pearson correlation we need to be careful in our definition of breakdown bound. Pearson correlation is by definition limited to the range $[-1, 1]$, so no matter how we changed the data set we could never make its value arbitrarily large. However, by changing only one value in our data set, we can make the Pearson correlation arbitrarily close to 1 or -1. Pearson correlation is the sum of the product of the z-scores of point in the data set divided by the size of the set. That is,

$$r = \frac{\sum_{i=1}^{n} z_{x_i} z_{y_i}}{n-1}$$

where z is

$$z_{x_i} = \frac{x_i - \overline{x}}{s}$$

where $\overline{x}$ is the sample mean and s is the sample standard deviation.

Let us say that we have an arbitrary set of data of size n. Then let us remove the n^{th} data point from the set. We then replace the n^{th} data point with a new data point whose x and y coordinates are equal. Call this point (L, L). Then

$$lim_{L\to\infty}\bar{x} = \frac{L}{n}$$

and

$$lim_{L\to\infty}\bar{y} = \frac{L}{n}$$

since as L becomes very large the other values of the data set become negligible in calculating the averages. Thus, for $i \in (1, 2, ..., n-1)$, as $L \to \infty$

$$x_i - \bar{x} \to 0 - \frac{L}{n} = -\frac{L}{n}$$

and

$$y_i - \bar{y} \to 0 - \frac{L}{n} = -\frac{L}{n}$$

Thus

$$(x_i - \bar{x})^2 \to \left(\frac{L}{n}\right)^2 \forall i \neq n$$

and

$$(y_i - \bar{y})^2 \to \left(\frac{L}{n}\right)^2 \forall i \neq n$$

Thus

$$\sum_{i=1}^{n-1}(x_i - \bar{x})^2 \to (n-1)\left(\frac{L}{n}\right)^2$$

and

$$\sum_{i=1}^{n-1}(y_i - \bar{y})^2 \to (n-1)\left(\frac{L}{n}\right)^2$$

Also, as $L \to \infty$

$$x_n - \bar{x} \to L - \frac{L}{n} = \frac{L(n-1)}{n}$$

and

$$y_n - \bar{y} \to L - \frac{L}{n} = \frac{L(n-1)}{n}$$

which implies that

$$(x_i - \bar{x})^2 \to \left(\frac{L(n-1)}{n}\right)^2$$

and

$$(y_i - \bar{y})^2 \to \left(\frac{L(n-1)}{n}\right)^2$$

Thus, as $L \to \infty$

$$\sum_{i=1}^{n}(x_i-\bar{x})^2 \to (n-1)\left(\frac{L}{n}\right)^2+\left(\frac{L(n-1)}{n}\right)^2$$

and

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 \to (n-1)\left(\frac{L}{n}\right)^2+\left(\frac{L(n-1)}{n}\right)^2$$

Then, as $L \to \infty$, the standard deviations of x and y are simply

$$s_x^2 = \sqrt{\frac{(\sum_{i=1}^{n}(x_i-\bar{x})^2)}{n-1}} \to \sqrt{\frac{(n-1)\left(\frac{L}{n}\right)^2+\left(\frac{L(n-1)}{n}\right)^2}{n-1}}$$

$$= \sqrt{\frac{\frac{L^2}{n^2}\left((n-1)+(n-1)^2\right)}{n-1}}$$

$$= \sqrt{\frac{\frac{L^2}{n^2}\left((n-1)+(n^2-2n+1)\right)}{n-1}}$$

$$= \sqrt{\frac{\frac{L^2}{n^2}(n^2-n)}{n-1}} = \sqrt{\frac{L^2 n(n-1)}{n^2(n-1)}} =$$

$$= \sqrt{\frac{L^2}{n}}$$

and similarly

$$s_y^2 \to \sqrt{\frac{L^2}{n}}$$

Then for $x_i \in (1,2,...,n-1)$,

$$z_{x_i} \to \frac{-\frac{L}{n}}{\sqrt{\frac{L^2}{n}}}$$

and

$$z_{y_i} \to \frac{-\frac{L}{n}}{\sqrt{\frac{L^2}{n}}}$$

Thus for $x_i \in (1,2,...,n-1)$,

$$z_{x_i} z_{y_i} \to \frac{\left(-\frac{L}{n}\right)^2}{\frac{L^2}{n}} = \frac{1}{n}$$

which implies that

$$\sum_{i=1}^{n-1} z_{x_i} z_{y_i} = (n-1)\frac{1}{n} = \frac{n-1}{n}$$

For x_n we have

$$z_{x_n} \to \frac{\left(\frac{L(n-1)}{n}\right)}{\sqrt{\frac{L^2}{n}}}$$

and

$$z_{y_n} \to \frac{\left(\frac{L(n-1)}{n}\right)}{\sqrt{\frac{L^2}{n}}}$$

so

$$z_{x_n} z_{y_n} \to \frac{\left(\frac{L(n-1)}{n}\right)^2}{\frac{L^2}{n}} = \frac{(n-1)^2}{n}$$

Thus

$$\sum_{i=1}^{n} z_{x_i} z_{y_i} = z_{x_n} z_{y_n} + \sum_{i=1}^{n-1} z_{x_i} z_{y_i} \to \frac{n-1}{n} + \frac{(n-1)^2}{n} = n-1$$

which implies that

$$r = \frac{\sum z_x z_y}{n-1} \to \frac{n-1}{n-1} = 1$$

Thus, by making L arbitrarily large, we can force the Pearson correlation to approach 1. This is not a desirable feature if we are concerned with the robustness of our correlation.

Although the above proof is for two dimensions, the same principle holds even in higher dimensions. Essentially, if we make one point in our data set extremely large, all of the other data points will appear to be very close together, almost becoming a single point. The Pearson correlation will report a very high correlation, either close to 1 or -1, depending on the direction in which we choose to make our point large. However, if this outlier is due to experimental error or other flaws in the collection of data, rather than

reflecting a true value, we could mistake a data set with no linear association between variables for one with a very high correlation.

2.6 W-Estimation of Location

Calculating the value of an M-estimator is not usually as easy as it is with the mean and median. W-estimation, a form of M-estimation, gives us a good way to actually find an estimate [3]. For a given M-estimator, the ψ-function defines a weight function:

$$uw(u) = \psi(u).$$

Then, since

$$\sum_{i=1}^{n} \psi(x_i - t) = 0,$$

we have

$$\sum_{i=1}^{n}(x_i - T_n)w(x_i - T_n) = 0 \Longrightarrow$$

$$\sum_{i=1}^{n}[x_i w(x_i - T_n) - T_n w(x_i - T_n)] = 0 \Longrightarrow$$

$$\sum_{i=1}^{n} x_i w(x_i - T_n) - \sum_{i=1}^{n} T_n w(x_i - T_n) = 0 \Longrightarrow$$

$$\sum_{i=1}^{n} x_i w(x_i - T_n) - T_n \sum_{i=1}^{n} w(x_i - T_n) = 0 \Longrightarrow$$

$$T_n \sum_{i=1}^{n} w(x_i - T_n) = \sum_{i=1}^{n} x_i w(x_i - T_n) \Longrightarrow$$

$$T_n = \frac{\sum_{i=1}^{n} x_i w(x_i - T_n)}{\sum_{i=1}^{n} w(x_i - T_n)}.$$

It is not usually possible to calculate T_n in closed form. However, it does lead us to an iterative approach. If we let $T_n^{(k)}$ be the estimate at the kth iterative step then

$$T_n^{(k+1)} = \frac{\sum_{i=1}^{n} x_i w(x_i - T_n^{(k)})}{\sum_{i=1}^{n} w(x_i - T_n^{(k)})}.$$

Then we have a method for finding the estimate. All we have to do is choose a $T_n^{(0)}$ as starting point and iterate until we are satisfied that the estimate has converged. $T_n^{(0)}$ is typically a robust measure of the center of the data set.

2.7 Existence and Uniqueness

There is no guarantee that an M-Estimate exists. It is, however, possible to prove existence and uniqueness under particular conditions. Before we prove this, first we need to prove a lemma [5].

Lemma: Let f and g be strictly monotone functions. Let Y_1 be a random variable not concentrated at one point (i.e. $\nexists\, x$ such that $P(Y_1 = x) = 1$). Then

$$cov[f(Y), g(Y)] > 0$$

Proof of Lemma: Let Y_2 be a random variable independent of and distributed identically to Y_1. Then

$$E[f(Y_1) - f(Y_2)][g(Y_1) - g(Y_2)] =$$

$$E[f(Y_1)g(Y_1) - f(Y_1)g(Y_2) - f(Y_2)g(Y_1) + f(Y_2)g(Y_2)] =$$

$$E[f(Y_1)g(Y_1)] - E[f(Y_1)g(Y_2)] - E[f(Y_2)g(Y_1)] + E[f(Y_2)g(Y_2)] =$$

$$E[f(Y_1)g(Y_1)] - E[f(Y_1)]E[g(Y_2)] + E[f(Y_2)g(Y_2)] - E[f(Y_2)]E[g(Y_1)]$$

since Y_1 and Y_2 are independent. This is then equal to

$$E[f(Y_1)g(Y_1)] - E[f(Y_1)]E[g(Y_1)] + E[f(Y_2)g(Y_2)] - E[f(Y_2)]E[g(Y_2)]$$

since Y_1 and Y_2 are identically distributed. By definition of covariance we then have that this is equal to

$$cov[(f(Y_1), g(Y_1)] + cov[(f(Y_2), g(Y_2)] = 2cov[(f(Y_1), g(Y_1)]$$

since Y_1 and Y_2 are identically distributed. Thus, we can say that

$$E[f(Y_1) - f(Y_2)][g(Y_1) - g(Y_2)] = 2cov[f(Y_1), g(Y_1)] \Rightarrow$$

$$\frac{1}{2}E[f(Y_1) - f(Y_2)][g(Y_1) - g(Y_2)] = cov[f(Y_1), g(Y_1)]$$

Then since f and g are monotone and $P(Y_1 = Y_2) \neq 1$, we have

$$[f(Y_1) - f(Y_2)][g(Y_1) - g(Y_2)] > 0 \Rightarrow$$
$$E[f(Y_1) - f(Y_2)][g(Y_1) - g(Y_2)] > 0 \Rightarrow cov[(f(Y_1), g(Y_1)] > 0$$

Let us say that we have a pair of equations for M-estimates of location and scale, known as a simultaneous M-estimate of location and scale. Let us call the M-estimate of location T_n and the M-estimate of scale S_n. Then we have

$$\sum \psi\left(\frac{x_i - T_n}{S_n}\right) = 0$$

and

$$\sum \chi\left(\frac{x_i - T_n}{S_n}\right) = 0$$

where χ is the objective function for the scale estimate. We can then modify these equations for use with the distribution rather than the data set. Thus we have

$$\int \psi\left(\frac{x - T(F)}{S(F)}\right) F(dx) = 0$$

and

$$\int \chi\left(\frac{x - T(F)}{S(F)}\right) F(dx) = 0$$

where F is the underlying distribution.

Then under certain conditions we can prove the existence and uniqueness of the solution of these equations.

Theorem: Let us assume that ψ and χ are differentiable, that ψ' is positive over its domain, that $\psi(x) = 0$ at $x = 0$ and χ has a minimum at $x = 0$ and χ'/ψ' is either strictly increasing or strictly decreasing. Then there is a unique solution to the above system of equations [5].

Proof: Let us consider the map

$$(t,s) \to \left(\int \psi\left(\frac{x-t}{s}\right) dF, \int \chi\left(\frac{x-t}{s}\right) dF \right)$$

Let $y = (x-t)/s$. Then the Jacobian of our map is

$$-\frac{1}{s}\begin{bmatrix} \int \psi'(y)dF & \int y\psi'(y)dF \\ \int \chi'(y)dF & \int y\chi'(y)dF \end{bmatrix}$$

Then let us define

$$dF^* = \frac{\psi'(y)}{E_F(\psi'(y))}dF.$$

Then we can rewrite the Jacobian like this:

$$-\frac{1}{s}E_F\left[\psi'(y)\right]\begin{bmatrix} 1 & E_{F^*}(y) \\ E_{F^*}\left(\frac{\chi'}{\psi'}\right) & E_{F^*}y\left(\frac{\chi'}{\psi'}\right) \end{bmatrix}$$

The determinant of this matrix is

$$\left(\frac{E_F\psi'(y)}{s}\right)^2 E_{F^*}y\left(\frac{\chi'}{\psi'}\right) - \left(\frac{E_F\psi'(y)}{s}\right)^2 E_{F^*}(y)E_{F^*}\left(\frac{\chi'}{\psi'}\right) =$$

$$\left(\frac{E_F\psi'(y)}{s}\right)^2 \left(E_{F^*}y\left(\frac{\chi'}{\psi'}\right) - E_{F^*}(y)E_{F^*}\left(\frac{\chi'}{\psi'}\right)\right) =$$

$$\left(\frac{E_F\psi'(y)}{s}\right)^2 cov_{F^*}\left(y, \frac{\chi'}{\psi'}\right),$$

so by our lemma we know that the determinant of this matrix is strictly positive. Since the determinant of the Jacobian is strictly positive and the diagonal elements are strictly negative we can conclude that our map is one-to-one.

Thus by the intermediate value theorem, there is a unique solution to our equations.

This proof does not, however, cover the case of Tukey's biwieght. We have a ψ-function which has a zero at $u = 0$, but $\psi' = (1-u^2)(1-5u^2)$ is not strictly greater than zero.

3 Resistant Method for Microarrays

An iterative estimate of Tukey's biweight can be used as a robust measure of location and scale of the expression levels between two genes on a microarray [1]. The scale estimate can then further be turned into a correlation between two genes. Tukey's biweight is especially advantageous because there is a parameter which can be used to control the breakdown, allowing the analyst to choose the level of robustness in the method.

The iterative method requires an initial estimate of location and scale. The coordinate-wise median and Median Absolute Deviation (MAD) are good choices for the initial estimates. There are better choices, such as the Minimum Covariance Determinant, for a starting point for the iterative method, but they take longer to compute. After running the iteration for a few steps the difference from the inital estimates disappears, so instead of using more computation-intensive initial estimates we can use our more easily computed inital estimates and longer iterations.

Before we explain the iterative method, we should define the objective function we are going to be using. In our case, we will use the objective function for Tukey's Biweight.

$$\rho(d) = \begin{cases} \frac{c^2}{6}\left[1 - (1 - \left(\frac{d}{c}\right)^2)^3\right] & |d| \leq c \\ \frac{1}{6} & |d| > c \end{cases}$$

After deciding on our objective funtion and finding our initial estimates of location (T) and scale (S), we calculate the distance from each point in the data set to the center of the data set. This is found by

$$d_j = \sqrt{(X_j - T)' S^{-1} (X_j - T)}$$

where j is the gene we are considering, and X_j is the data from gene j. We then find our parameter of constraint k by solving

$$n^{-1} \sum_{j=1}^{n} \rho\left(\frac{d_j}{k}\right) = b_0$$

where

$$b_0 = E\left[\rho\left(\frac{d}{k}\right)\right]$$

for k, where n is the number of samples. b_0 is the expected value of $\rho\left(\frac{d}{k}\right)$ when the data are normally distributed. The purpose of the constraint, k is to prevent excessive downweighting as the iterative process moves forward.

Since with the biweight there is a rejection point beyond which observations are weighted to zero, there is a danger that with each successive iteration more and more observations will be weighted to zero. The weighting of outlying data points to zero means that they are not considered in the calculation of the scale estimate. The new scale estimate may lead to new points being weighted to zero because the old outliers are no longer considered. The constraint counteracts this effect.

Next we calculate the weight functions to be used in finding the next iteration of our location and scale estimates.

$$\psi(d) = \frac{\partial \rho(d)}{\partial d}$$

$$w(d) = \frac{\psi(d)}{d}$$

$$v(d) = d\psi(d)$$

where ρ is our objective function. We then calculate the next iteration of our location and scale estimates.

$$T^{(i+1)} = \frac{\sum_j w(d_j^{(i)}/k^{(i)}) X_j}{\sum_j w(d_j^{(i)}/k^{(i)})}$$

$$S^{(i+1)} = \frac{\sum_j w(d_j^{(i)}/k^{(i)})(X_j - T^{i+1})(X_j - T^{i+1})'}{\sum_j v(d_j^{(i)}/k^{(i)})}$$

These are weighted estimates. The biweight function is used to calculate a weight for every point, and these weights are in turn used to calculate the estimates. The estimate of location is simply a weighted mean with weights derived from the biweight. The estimate of scale is similarly weighted.

We then recalculate the distance of each point in the set from the center of the set with respect to the new estimates of location and scale.

$$d_j^{(i+1)} = \sqrt{(X_j - T^{(i+1)})' \left(S^{(i+1)}\right)^{-1} (X_j - T^{(i+1)})}$$

These new distances are then used to recalculate k and in turn the next iteration of our estimates for location and scale. When we have iterated a satisfactory number of times, we can then use our estimates of location and scale to find a correlation between genes.

If we are interested in genes p and q we simply take the element of S in the p^{th} row and the q^{th} column and call it s_{pq}. This can be thought of as a

robust version of the covariance of X_p and X_q. Thus, the biwieght correlation between these two genes is

$$r_{pq} = \frac{s_{pq}}{\sqrt{s_{pp}s_{qq}}}$$

With this new correlation in hand, we can then compare the biweight correlation of two genes with their Pearson correlation. This will allow us to find pairs of genes for which the Pearson correlation may be affected by the presence of an outlier. These pairs can then be examined to see if this is the case.

4 Conclusion

Robust measures of distance are needed for the analysis of DNA microarrays due to problems inherent to the process of conducting microarray experiments. In evaluating the robustness of estimates, we look at the concept of breakdown bound. The traditional definition of breakdown bound for estimates of location provides a method for that evaluation, as does a similar concept for measuring the robustness of a correlation. Using this method, we can see that the most commonly used correlation, the Pearson correlation, has a breakdown bound of 0, and so is completely non-resistant to outliers. The iterative biweight estimate provides a more robust choice than Pearson correlation. Since we can set the breakdown bound for the biweight as we choose, it is a flexible choice as a distant metric, with the data analyst deciding on the level of robustness needed. The biweight correlations can be compared with the Pearson correlation, allowing the data analyst to flag genes for which the Pearson correlation may not be giving an accurate picture.

References

[1] Hardin J, Mitani A, Hicks L, VanKoten B: **A robust measure of correlation between two genes on a microarray.** *BMC Bioinformatics* 2007, **8**:220.

[2] Heyer L, Kruglyak S, Yooseph S: **Exploring Expression Data: Identification and Analysis of Coexpressed Genes.** *Genome Research* 1999, **9**:1106-1115.

[3] Hoaglin DC, Mosteller F, Tukey JW, Eds: *Understanding Robust and Exploratory Data Analysis* Wiley Classics Library, Wiley-Interscience, New York; 2000.

[4] Hubbel E, Liu W, Mei R: **Robust estimators for expression analysis.** *Bioinformatics* 2002, **18**:1585-1592.

[5] Huber PJ: *Robust Statistics* John Wiley & Sons, New York; 1981.

[6] Maronna RA, Martin RD, Yohai VJ: *Robust Statistics Theory and Methods* John Wiley & Sons, Chichester; 2006.

[7] Rousseeuw PJ, Leroy AM: *Robust Regression and Outlier Detection.* John Wiley & Sons, New York; 1987.

[8] Wilcox R: *Introduction to Robust Estimation and Hypothesis Testing* Elsevier Academic Press; 2005.