Technical Report: Asymptotic Tests For Location and Variance Using the Distance Matrix

Douglas Hayden
Massachusetts General Hospital Biostatistics Center

Mark Kon
Boston University Department of Mathematics and Statistics

May 22, 2008

Contents

List of Tables

List of Figures

1 Introduction

This technical report summarizes some of the work done to try to develop asymptotic tests for location and variance differences between two groups of microarrays or more generally any high dimensional vectors. This work was undertaken to find an alternative to the permutation tests for location and variance developed with Dr. David Schoenfeld at the Massachusetts General Hospital Biostatistics Center.

The permutation tests are based on the distance matrix of pairwise distances between all microarrays of both groups. It was hoped that the test statistics used for the permutation test would turn out to be asymptotically normal under very weak assumptions about the underlying probablity distribution of the microarrays which is essentially unknowable due to its very high dimension.

As it turned out, under the assumption of additive errors and squared Euclidean distance, the location test is not asymptotically normal without further assumptions whereas the variance test and the equivalence test are asymptotically normal.

2 One Independent Sample

2.1 The Model

This section proposes a one sample model of the distance matrix which may also be used as a global null model for the comparison of two samples assumed, under the null, to be drawn from the same distribution.

Let $\mathrm{X_1, \ldots, X_n}$ be a random sample of independent and identically distributed random variables. Let $\mathrm{d_{ij} = d[X_i, X_j]}$ be a function (loosely called a distance measure) mapping pairs of random variables into the real line such that:

$$\mathrm{d_{ij} \equiv d_0 \Leftrightarrow i = j} \tag{1}$$

and

$$\mathrm{d_{ij} = d_{ji}} \tag{2}$$

where $\mathrm{d_0}$ is a fixed constant. For example if if $\mathrm{d[\cdot]}$ is Pearson correlation then $\mathrm{d_0} = 1$ while if $\mathrm{d[\cdot]}$ is Euclidean distance then $\mathrm{d_0} = 0$.

It follows from Equations 1 and 2 that all the information is contained in the set of $\mathrm{n(n-1)/2}$ pairwise distances $\mathrm{\{d_{ij}\}}$, $\mathrm{1 \le i < j \le n}$. The $\mathrm{\{d_{ij}\}}$ are identically distributed with common mean $\mathrm{E[d_{ij}] = d}$, but $\mathrm{d_{ij}}$ and $\mathrm{d_{kl}}$ are independent if and only if they do not share a common index. That is, if $\mathrm{\{i, j\} \cap \{k, l\}} = \phi$. If $\mathrm{d_{ij}}$ and $\mathrm{d_{kl}}$ do share a common index, then $\mathrm{d_{ij}}$ and $\mathrm{d_{kl}}$ are dependent.

It follows that we can assume a two parameter covariance matrix for $\mathrm{\{d_{ij}\}}$ with parameters:

$$\mathrm{Variance[d_{ij}]} = \sigma^2 \tag{3}$$

for all $\mathrm{d_{ij}}$ and

$$\mathrm{Covariance[d_{ij}, d_{kl}]} = \rho\sigma^2 \tag{4}$$

if $\mathrm{d_{ij}}$ and $\mathrm{d_{kl}}$ are dependent.

2.2 Sample Average Distance

Let I be the set of all $\mathrm{n_I}$ independent pairs and D the set of all $\mathrm{n_D}$ dependent pairs of $\{\mathrm{d_{ij}},\mathrm{d_{kl}}\}$. A simple counting argument gives;

$$\mathrm{n_I} = \frac{\mathrm{n(n-1)(n-2)(n-3)}}{8} \tag{5}$$

and

$$\mathrm{n_D} = \frac{\mathrm{n(n-1)(n-2)}}{2} \tag{6}$$

It follows that the sample average distance:

$$\mathrm{\bar{d}_{xx}} = \frac{2}{\mathrm{n(n-1)}} \sum_{\mathrm{i<j}} \mathrm{d_{ij}} \tag{7}$$

has variance

$$\frac{2\sigma^2}{\mathrm{n(n-1)}} + \frac{4(\mathrm{n}-2)\rho\sigma^2}{\mathrm{n(n-1)}} \tag{8}$$

which vanishes as $\mathrm{n} \to \infty$ so that $\mathrm{\bar{d}}$ is an unbiased and consistent estimator of d.

2.3 Covariance Parameter Estimators

Define the two sample statistics:

$$\mathrm{S_I^2} = \frac{1}{2\mathrm{n_I}} \sum_{\{\mathrm{d_{ij}},\mathrm{d_{kl}}\}\epsilon \mathrm{I}} (\mathrm{d_{ij}} - \mathrm{d_{kl}})^2 \tag{9}$$

$$\mathrm{S_D^2} = \frac{1}{2\mathrm{n_D}} \sum_{\{\mathrm{d_{ij}},\mathrm{d_{kl}}\}\epsilon \mathrm{D}} (\mathrm{d_{ij}} - \mathrm{d_{kl}})^2 \tag{10}$$

It follows that

$$\mathrm{E[S_I^2]} = \sigma^2 \tag{11}$$

$$\mathrm{E[S_I^2 - S_D^2]} = \rho\sigma^2 \tag{12}$$

and the statistic

$$\mathrm{S}^2 = \frac{2\mathrm{S}_\mathrm{I}^2}{\mathrm{n}(\mathrm{n}-1)} + \frac{4(\mathrm{n}-2)(\mathrm{S}_\mathrm{I}^2 - \mathrm{S}_\mathrm{D}^2)}{\mathrm{n}(\mathrm{n}-1)} \tag{13}$$

is an unbiased estimate of the variance of the sample average distance given by Equation 7. Note that for $\mathrm{n} < 4$, $\mathrm{n}_\mathrm{I} = 0$ and the proposed covariance parameter estimators do not exist.

2.4 Consistency of Covariance Parameter Estimators

In this section we show that S_I^2 and S_D^2 are consistent estimators. Since they are unbiased it suffices to show that their variance vanishes as $\mathrm{n} \to \infty$.

Since each estimator is the average of identically distributed squared differences of the form, $(\mathrm{d}_\mathrm{ij} - \mathrm{d}_\mathrm{kl})^2$, it is only necessary to count the number of non-zero covariance terms in the covariance matrix of the $(\mathrm{d}_\mathrm{ij} - \mathrm{d}_\mathrm{kl})^2$ of each estimator.

Let's consider the general case first to clarify the idea. Suppose we have a sample of k identically distributed random variables, $\mathrm{x}_\mathrm{i}, \ldots, \mathrm{x}_\mathrm{k}$, with common variance σ^2 and covariance $\{\rho_\mathrm{ij}\sigma^2\}_{\mathrm{i}\neq\mathrm{j}}$. Suppose there are k_0 covariance terms that are identically zero, then since the covariance matrix of the variables x_i is k by k and since there are $(\mathrm{k}^2 - \mathrm{k})$ off diagonal covariance terms, the variance of the sample mean is given by:

$$\frac{\sigma^2}{\mathrm{k}} + \frac{(\mathrm{k}^2 - \mathrm{k} - \mathrm{k}_0)\bar{\rho}\sigma^2}{\mathrm{k}^2} \tag{14}$$

where $\bar{\rho}$ is the mean of the non-zero ρ_ij. Thus it can be seen that the variance vanishes as $\mathrm{k} \to \infty$ provide $\mathrm{k}_0 \sim \circ(\mathrm{k}^2)$.

Consider first S_I^2, which is the average of the terms $(\mathrm{d}_\mathrm{ij} - \mathrm{d}_\mathrm{kl})^2$ where the elements of the pair $\{\mathrm{d}_\mathrm{ij}, \mathrm{d}_\mathrm{kl}\}$ are independent. Each of these pairs is independent of any other pair drawn from the $\mathrm{n} - 4$ by $\mathrm{n} - 4$ sub-matrix remaining after excluding the rows and columns common to $\mathrm{i}, \mathrm{j}, \mathrm{k}, \mathrm{l}$. Thus, by Equation 5, each pair is independent of

$$\frac{(\mathrm{n}-4)(\mathrm{n}-5)(\mathrm{n}-6)(\mathrm{n}-7)}{8} \tag{15}$$

other pairs and since there are

$$\frac{\mathrm{n}(\mathrm{n}-1)(\mathrm{n}-2)(\mathrm{n}-3)}{8} \tag{16}$$

pairs the number of identically zero covariance terms is at least

$$\frac{\mathrm{n}(\mathrm{n}-1)(\mathrm{n}-2)(\mathrm{n}-3)(\mathrm{n}-4)(\mathrm{n}-5)(\mathrm{n}-6)(\mathrm{n}-7)}{64} \tag{17}$$

which is of order

$$\left(\frac{\mathrm{n}(\mathrm{n}-1)(\mathrm{n}-2)(\mathrm{n}-3)}{8}\right)^2 \tag{18}$$

as desired.

Consider next S_D^2, which is the average of the terms $(\mathrm{d}_\mathrm{ij} - \mathrm{d}_\mathrm{kl})^2$, where the elements of the pair $\{\mathrm{d}_\mathrm{ij}, \mathrm{d}_\mathrm{kl}\}$ are dependent. Each of these pairs is independent of any other pair drawn from the $\mathrm{n}-3$ by $\mathrm{n}-3$ sub-matrix remaining after excluding rows and columns common to $\mathrm{i}, \mathrm{j}, \mathrm{k}, \mathrm{l}$. Thus, by Equation 6, each pair is independent of

$$\frac{(\mathrm{n}-3)(\mathrm{n}-4)(\mathrm{n}-5)}{2} \tag{19}$$

other pairs and since there are

$$\frac{\mathrm{n}(\mathrm{n}-1)(\mathrm{n}-2)}{2} \tag{20}$$

pairs the number of identically zero covariance terms is at least

$$\frac{\mathrm{n}(\mathrm{n}-1)(\mathrm{n}-2)(\mathrm{n}-3)(\mathrm{n}-4)(\mathrm{n}-5)}{4} \tag{21}$$

which is of order

$$\left(\frac{\mathrm{n}(\mathrm{n}-1)(\mathrm{n}-2)}{2}\right)^2 \tag{22}$$

as desired.

2.5 Test Statistics

Under this global null model we will consider test statistics of the form:

$$\mathrm{T} = \sum_{\mathrm{i}<\mathrm{j}} \mathrm{c}_\mathrm{ij} \mathrm{d}_\mathrm{ij} \tag{23}$$

where the c_ij are specified constants such that:

$$\sum_{\mathrm{i}<\mathrm{j}} \mathrm{c}_\mathrm{ij} = 0 \tag{24}$$

so that

$$\mathrm{E}[\mathrm{T}] = 0 \tag{25}$$

and

$$\mathrm{Variance[T]} = \sum_{\mathrm{i<j}} \sum_{\mathrm{k<l}} \mathrm{c_{ij} c_{kl} Cov[d_{ij}, d_{kl}]} \tag{26}$$

where

$$\mathrm{Cov[d_{ij}, d_{kl}]} = 0 \quad \text{if} \quad \#[\{\mathrm{i, j}\} \cap \{\mathrm{k, l}\}] = 0 \tag{27}$$

$$\mathrm{Cov[d_{ij}, d_{kl}]} = \rho\sigma^2 \quad \text{if} \quad \#[\{\mathrm{i, j}\} \cap \{\mathrm{k, l}\}] = 1 \tag{28}$$

$$\mathrm{Cov[d_{ij}, d_{kl}]} = \sigma^2 \quad \text{if} \quad \#[\{\mathrm{i, j}\} \cap \{\mathrm{k, l}\}] = 2 \tag{29}$$

2.6 Simplest Example

To make the ideas concrete consider just four independent and identically distributed random variables $\mathrm{X_1, X_2, X_3, X_4}$. The six pairwise distances, $\{\mathrm{d_{12}}$, $\mathrm{d_{13}}$, $\mathrm{d_{14}}$, $\mathrm{d_{23}}$, $\mathrm{d_{24}}$, $\mathrm{d_{34}}\}$, can be laid out as the upper triangular region of the distance matrix shown in Table 1. There are three independent pairs $\{\mathrm{d_{12}, d_{34}}\}$, $\{\mathrm{d_{14}, d_{23}}\}$, $\{\mathrm{d_{13}, d_{24}}\}$ so the covariance matrix has six zero and twenty four non-zero off diagonal elements as shown in Table 2.

Table 1: Triangular Upper Half of Distance Matrix

	$\mathrm{X_1}$	$\mathrm{X_2}$	$\mathrm{X_3}$	$\mathrm{X_4}$
$\mathrm{X_1}$	$\mathrm{d_0}$	$\mathrm{d_{12}}$	$\mathrm{d_{13}}$	$\mathrm{d_{14}}$
$\mathrm{X_2}$	.	$\mathrm{d_0}$	$\mathrm{d_{23}}$	$\mathrm{d_{24}}$
$\mathrm{X_3}$	.	.	$\mathrm{d_0}$	$\mathrm{d_{34}}$
$\mathrm{X_4}$	.	.	.	$\mathrm{d_0}$

Table 2: Covariance Matrix of $\{\mathrm{d}_{12}, \mathrm{d}_{13}, \mathrm{d}_{14}, \mathrm{d}_{23}, \mathrm{d}_{24}, \mathrm{d}_{34}\}$

	d_{12}	d_{13}	d_{14}	d_{23}	d_{24}	d_{34}
d_{12}	σ^2	$\rho\sigma^2$	$\rho\sigma^2$	$\rho\sigma^2$	$\rho\sigma^2$	0
d_{13}	$\rho\sigma^2$	σ^2	$\rho\sigma^2$	$\rho\sigma^2$	0	$\rho\sigma^2$
d_{14}	$\rho\sigma^2$	$\rho\sigma^2$	σ^2	0	$\rho\sigma^2$	$\rho\sigma^2$
d_{23}	$\rho\sigma^2$	$\rho\sigma^2$	0	σ^2	$\rho\sigma^2$	$\rho\sigma^2$
d_{24}	$\rho\sigma^2$	0	$\rho\sigma^2$	$\rho\sigma^2$	σ^2	$\rho\sigma^2$
d_{34}	0	$\rho\sigma^2$	$\rho\sigma^2$	$\rho\sigma^2$	$\rho\sigma^2$	σ^2

3 Two Independent Samples

In this section we discuss the pairwise distances between two independent samples of microarrays. Let $X_1, \ldots, X_n$ be an IID random sample from distribution F and $Y_1, \ldots, Y_m$ be an IID random sample from distribution G and let $d_{ij} = d[X_i, Y_j]$.

As a simple example let $n = 3$ and $m = 4$ giving the distance matrix shown below.

	Y_1	Y_2	Y_3	Y_4
X_1	d_{11}	d_{12}	d_{13}	d_{14}
X_2	d_{21}	d_{22}	d_{23}	d_{24}
X_3	d_{31}	d_{32}	d_{33}	d_{34}

The $\{d_{ij}\}$ are identically distributed with common mean $E[d_{ij}] = d$, but d_{ij} and d_{kl} are independent if and only if $i \neq k$ and $j \neq l$.

We can assume a three parameter covariance matrix for $\{d_{ij}\}$ with parameters:

$$\text{Variance}[d_{ij}] = \sigma^2 \tag{30}$$

for all d_{ij} and

$$\text{Covariance}[d_{ij}, d_{kl}] = \rho_R \sigma^2 \tag{31}$$

if they share a commom row so that $i = k$ and

$$\text{Covariance}[d_{ij}, d_{kl}] = \rho_C \sigma^2 \tag{32}$$

if they share a common column so that $j = l$.

3.1 Sample Average Distance

Let I be the set of all n_I independent pairs, R the set of all n_R row dependent pairs, and C the set of all n_C column dependent pairs of the $\{d_{ij}, d_{kl}\}$. A simple counting argument gives:

$$n_I = \frac{nm(n-1)(m-1)}{2} \tag{33}$$

and

$$n_R = \frac{nm(m-1)}{2} \tag{34}$$

and

$$\mathrm{n_C} = \frac{\mathrm{nm(n-1)}}{2} \tag{35}$$

It follows that the sample average distance

$$\bar{\mathrm{d}}_{\mathrm{xy}} = \frac{1}{\mathrm{nm}} \sum_{\mathrm{i}} \sum_{\mathrm{j}} \mathrm{d_{ij}} \tag{36}$$

has variance

$$\frac{\sigma^2}{\mathrm{nm}} + \frac{(\mathrm{m}-1)\rho_{\mathrm{R}}\sigma^2}{\mathrm{nm}} + \frac{(\mathrm{n}-1)\rho_{\mathrm{C}}\sigma^2}{\mathrm{nm}} \tag{37}$$

which vanishes as $\mathrm{n}, \mathrm{m} \to \infty$ so that $\bar{\mathrm{d}}_{\mathrm{xy}}$ is an unbiased and consistent estimator of d.

3.2 Covariance Parameter Estimators

Define the three sample statistics:

$$\mathrm{S_I^2} = \frac{1}{2\mathrm{n_I}} \sum_{\{\mathrm{d_{ij}},\mathrm{d_{kl}}\}\epsilon \mathrm{I}} (\mathrm{d_{ij}} - \mathrm{d_{kl}})^2 \tag{38}$$

$$\mathrm{S_R^2} = \frac{1}{2\mathrm{n_R}} \sum_{\{\mathrm{d_{ij}},\mathrm{d_{ik}}\}\epsilon \mathrm{R}} (\mathrm{d_{ij}} - \mathrm{d_{ik}})^2 \tag{39}$$

$$\mathrm{S_C^2} = \frac{1}{2\mathrm{n_C}} \sum_{\{\mathrm{d_{ij}},\mathrm{d_{kj}}\}\epsilon \mathrm{C}} (\mathrm{d_{ij}} - \mathrm{d_{kj}})^2 \tag{40}$$

It follows that

$$\mathrm{E}[\mathrm{S_I^2}] = \sigma^2 \tag{41}$$

$$\mathrm{E}[\mathrm{S_I^2} - \mathrm{S_R^2}] = \rho_{\mathrm{R}}\sigma^2 \tag{42}$$

$$\mathrm{E}[\mathrm{S_I^2} - \mathrm{S_C^2}] = \rho_{\mathrm{C}}\sigma^2 \tag{43}$$

and the statistic

$$\mathrm{S}^2 = \frac{\mathrm{S_I^2}}{\mathrm{nm}} + \frac{(\mathrm{m}-1)\rho_{\mathrm{R}}(\mathrm{S_I^2} - \mathrm{S_R^2})}{\mathrm{nm}} + \frac{(\mathrm{n}-1)\rho_{\mathrm{C}}(\mathrm{S_I^2} - \mathrm{S_C^2})}{\mathrm{nm}} \tag{44}$$

is an unbiased estimate of the variance of the sample average distance given by Equation 37.

3.3 Consistency of Covariance Parameter Estimators

In this section we show that S_I^2, S_R^2, and S_C^2 are consistent estimators. Since they are unbiased it suffices to show that their variance vanishes as $\mathrm{n}, \mathrm{m} \to \infty$.

Since each estimator is the average of identically distributed squared differences of the form, $(\mathrm{d}_\mathrm{ij} - \mathrm{d}_\mathrm{kl})^2$, it is only necessary to count the number of non-zero covariance terms in the covariance matrix of the $(\mathrm{d}_\mathrm{ij} - \mathrm{d}_\mathrm{kl})^2$ of each estimator.

Let's consider the general case first to clarify the idea. Suppose we have a sample of k identically distributed random variables, $\mathrm{x}_\mathrm{i}, \ldots, \mathrm{x}_\mathrm{k}$, with common variance σ^2 and covariance $\{\rho_\mathrm{ij}\sigma^2\}_{\mathrm{i}\neq\mathrm{j}}$. Suppose there are k_0 covariance terms that are identically zero. Then since the covariance matrix of the x_i is k by k and since there are $(\mathrm{k}^2 - \mathrm{k})$ off diagonal covariance terms the variance of the sample mean is given by:

$$\frac{\sigma^2}{\mathrm{k}} + \frac{(\mathrm{k}^2 - \mathrm{k} - \mathrm{k}_0)\bar{\rho}\sigma^2}{\mathrm{k}^2} \tag{45}$$

where $\bar{\rho}$ is the mean of the non-zero ρ_ij. Thus it can be seen that the variance vanishes as $\mathrm{k} \to \infty$ provide $\mathrm{k}_0 \sim \circ(\mathrm{k}^2)$.

Consider first S_I^2, which is the average of the terms $(\mathrm{d}_\mathrm{ij} - \mathrm{d}_\mathrm{kl})^2$, where the elements of the pair $\{\mathrm{d}_\mathrm{ij}, \mathrm{d}_\mathrm{kl}\}$ are independent. Each of these pairs is independent of any other pair drawn from the $\mathrm{n} - 2$ rows and $\mathrm{m} - 2$ columns remaining after excluding rows i, k and columns j, l. Thus, by Equation 33, each pair is independent of

$$\frac{(\mathrm{n}-2)(\mathrm{m}-2)(\mathrm{n}-3)(\mathrm{m}-3)}{2} \tag{46}$$

other pairs and since there are

$$\frac{\mathrm{nm}(\mathrm{n}-1)(\mathrm{m}-1)}{2} \tag{47}$$

pairs the number of identically zero covariance terms is at least

$$\frac{\mathrm{nm}(\mathrm{n}-1)(\mathrm{m}-1)(\mathrm{n}-2)(\mathrm{m}-2)(\mathrm{n}-3)(\mathrm{m}-3)}{4} \tag{48}$$

which is of order

$$\left(\frac{\mathrm{nm}(\mathrm{n}-1)(\mathrm{m}-1)}{2}\right)^2 \tag{49}$$

as desired.

Consider next $\mathrm{S}_{\mathrm{R}}^2$, which is the average of the terms $(\mathrm{d}_{\mathrm{ij}} - \mathrm{d}_{\mathrm{ik}})^2$, where the elements of the pair $\{\mathrm{d}_{\mathrm{ij}}, \mathrm{d}_{\mathrm{ik}}\}$ share a common row. Each of these pairs is independent of any other pair drawn from the $\mathrm{n} - 1$ rows and $\mathrm{m} - 2$ columns remaining after excluding row i and columns j, k. Thus, by Equation 34, each pair is independent of

$$\frac{(\mathrm{n}-1)(\mathrm{m}-2)(\mathrm{m}-3)}{2} \tag{50}$$

other pairs and since there are

$$\frac{\mathrm{nm}(\mathrm{m}-1)}{2} \tag{51}$$

pairs the number of identically zero covariance terms is at least

$$\frac{\mathrm{nm}(\mathrm{m}-1)(\mathrm{n}-1)(\mathrm{m}-2)(\mathrm{m}-3)}{4} \tag{52}$$

which is of order

$$\left(\frac{\mathrm{nm}(\mathrm{m}-1)}{2}\right)^2 \tag{53}$$

as desired.

A similar argument applies to $\mathrm{S}_{\mathrm{C}}^2$ which is the average of the column dependent pairs.

4 Additive Errors and Euclidean Distance

4.1 The Model

In this section we specialize our general model to the case of additive errors and Euclidean distance.

Consider IID samples $\mathrm{X}_1, \ldots, \mathrm{X}_{\mathrm{n}}$ and $\mathrm{Y}_1, \ldots, \mathrm{Y}_{\mathrm{m}}$ from two different distributions. Under an additive error model we have:

$$\begin{aligned} \mathrm{X}_{\mathrm{i}} &= \mu_{\mathrm{x}} + \epsilon_{\mathrm{i}} && (54) \\ \mathrm{Y}_{\mathrm{i}} &= \mu_{\mathrm{y}} + \delta_{\mathrm{i}} && (55) \end{aligned}$$

where μ_{x} and μ_{y} are the respective mean vectors of the X_{i} and Y_{i} and ϵ_{i} and δ_{i} are random error vectors with expected value 0 and variance covariance matrices Σ_{x}

and Σ_y respectively. Since the samples are IID the random error vectors are also independent.

If we use squared Euclidean distance as our distance measure then the within X mean distance is:

$$\bar{d}_{xx} = \frac{2}{n(n-1)} \sum_{i<j} |X_i - X_j|^2 \tag{56}$$

and the within Y mean distance is:

$$\bar{d}_{yy} = \frac{2}{m(m-1)} \sum_{i<j} |Y_i - Y_j|^2 \tag{57}$$

and the between X and Y mean distance is:

$$\bar{d}_{xy} = \frac{1}{nm} \sum_{i,j} |X_i - Y_j|^2 \tag{58}$$

These sample statistics have expected values:

$$\begin{aligned} E[\bar{d}_{xx}] &= 2\mathrm{Tr}[\Sigma_x] && (59)\\ E[\bar{d}_{yy}] &= 2\mathrm{Tr}[\Sigma_y] && (60)\\ E[\bar{d}_{xy}] &= |\mu_x - \mu_y|^2 + \mathrm{Tr}[\Sigma_x] + \mathrm{Tr}[\Sigma_y] && (61) \end{aligned}$$

where $\mathrm{Tr}[]$ is the trace operator which gives the sum of the diagonal elements of a matrix.

4.2 A Lemma

We need a simple lemma to use in the proofs of asymptotic normality. Let e_{ij} be defined as any one of the following three inner products:

$$e_{ij} = (X_i - \mu_x)^T (X_j - \mu_x) \tag{62}$$

or

$$e_{ij} = (Y_i - \mu_y)^T (Y_j - \mu_y) \tag{63}$$

or

$$e_{ij} = (X_i - \mu_x)^T (Y_j - \mu_y) \tag{64}$$

It follows that for $i \neq j \neq k \neq i$ we have:

$$E[e_{ij}] = 0 \tag{65}$$

and

$$\mathrm{cov}[\mathrm{e_{ij}}, \mathrm{e_{ik}}] = 0 \tag{66}$$

and

$$\mathrm{cov}[\mathrm{e_{ij}}, \mathrm{e_{ii}}] = 0 \tag{67}$$

The proof of this follows immediately from the one dimensional case. Let $\mathrm{e_i}, \mathrm{e_j}, \mathrm{e_k}$ be mean zero pairwise independent scalars. Then $\mathrm{e_{ii}} = \mathrm{e_i e_i}, \mathrm{e_{ij}} = \mathrm{e_i e_j}, \mathrm{e_{ik}} = \mathrm{e_i e_k}$ and

$$\mathrm{E[e_{ij}] = E[e_i]E[e_j] = 0}$$

and

$$\begin{aligned}
\mathrm{cov}[\mathrm{e_{ij}}, \mathrm{e_{ik}}] &= \mathrm{E[e_i e_j e_i e_k] - E[e_i e_j]E[e_i e_k]} \\
&= \mathrm{E[e_i e_i]E[e_j]E[e_k] - E[e_i]E[e_j]E[e_i]E[e_k]} \\
&= 0
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{cov}[\mathrm{e_{ij}}, \mathrm{e_{ii}}] &= \mathrm{E[e_i e_i e_i e_j] - E[e_i e_j]E[e_i e_i]} \\
&= \mathrm{E[e_i e_i e_i]E[e_j] - E[e_i]E[e_j]E[e_i e_i]} \\
&= 0
\end{aligned}$$

4.3 Independent Asymptotically Normal Random Variables

In this section we prove that a linear combination of two independent asymptotically normal random variables is also asymptotically normal. We will use this result in the proofs of asymptotic normality.

Let $\mathrm{W_n}$ and $\mathrm{Z_m}$ be two sequences of random variables which are independent for all values of n and m and which converge in distribution to the normal random variables W and Z respectively as n and m go to infinity. Let a and b be constants and let $\phi_{\mathrm{X}}(\mathrm{t})$ denote the characteristic function of the random variable X. Then by continuity [1] $\mathrm{aW_n}$ and $\mathrm{bZ_m}$ converge in distribution to the normal random variables aW and bZ respectively. Since $\mathrm{W_n}$ and $\mathrm{Z_m}$ are independent the characteristic function of $\mathrm{aW_n + bZ_m}$ factors into $\phi_{\mathrm{aW_n}}(\mathrm{t})\phi_{\mathrm{bZ_m}}(\mathrm{t})$, which, by Levy's continuity theorem [2] converges to $\phi_{\mathrm{aW}}(\mathrm{t})\phi_{\mathrm{bZ}}(\mathrm{t})$ as n and m go to infinity. Since aW and bZ are both normal random variables, the product of their characteristic functions is also that of a normal random variable. Thus the linear combination $\mathrm{aW_n + bZ_m}$ converges in distribution to a normal random variable by Levy's continuity theorem.

4.4 One Sample Mean Distance: Asymptotic Normality

We have $\mathrm{X_1, \ldots, X_n}$ an IID random sample with common mean μ_x. The sample mean squared Euclidean distance is:

$$\mathrm{\bar{d}_{xx}} = \frac{2}{\mathrm{n(n-1)}} \sum_{\mathrm{i<j}} |\mathrm{X_i - X_j}|^2 \tag{68}$$

or by adding and subtracting μ_x:

$$\frac{2}{\mathrm{n(n-1)}} \sum_{\mathrm{i<j}} |\mathrm{X_i} - \mu_\mathrm{x} + \mu_\mathrm{x} - \mathrm{X_j}|^2 \tag{69}$$

which expands to:

$$\frac{2}{\mathrm{n(n-1)}} \sum_{\mathrm{i<j}} |\mathrm{X_i} - \mu_\mathrm{x}|^2 + |\mathrm{X_j} - \mu_\mathrm{x}|^2 - 2(\mathrm{X_i} - \mu_\mathrm{x})^\mathrm{T}(\mathrm{X_j} - \mu_\mathrm{x}) \tag{70}$$

which simplifies to a sum of the two terms:

$$\frac{2}{\mathrm{n}} \sum_{\mathrm{i}} |\mathrm{X_i} - \mu_\mathrm{x}|^2 \tag{71}$$

$$-\frac{4}{\mathrm{n(n-1)}} \sum_{\mathrm{i<j}} (\mathrm{X_i} - \mu_\mathrm{x})^\mathrm{T}(\mathrm{X_j} - \mu_\mathrm{x}) \tag{72}$$

Now term 71 is twice the mean of n IID terms. Suppose each has mean μ and variance $\mathrm{V_1}$. Term 72 is minus twice the mean of $\mathrm{n(n-1)/2}$ identically distributed and uncorrelated terms of mean zero (by the lemma). Suppose each has variance $\mathrm{V_2}$. We can center and rescale $\mathrm{\bar{d}_{xx}}$ to get:

$$\frac{\sqrt{\mathrm{n}}}{\sqrt{\mathrm{V_1}}}(\frac{\mathrm{\bar{d}_{xx}}}{2} - \mu) = \frac{\sqrt{\mathrm{n}}}{\sqrt{\mathrm{V_1}}} \left(\frac{1}{\mathrm{n}} \sum_{\mathrm{i}} |\mathrm{X_i} - \mu_\mathrm{x}|^2 - \mu \right) - \frac{2\sqrt{\mathrm{n}}}{\mathrm{n(n-1)}\sqrt{\mathrm{V_1}}} \sum_{\mathrm{i<j}} (\mathrm{X_i} - \mu_\mathrm{x})^\mathrm{T}(\mathrm{X_j} - \mu_\mathrm{x}) \tag{73}$$

Now as $\mathrm{n} \to \infty$, the first term on the right hand side of Equation 73 converges in distribution to $\mathrm{N}[0,1]$ by the Central Limit Theorem [3]. Also since the second term on the right in Equation 73, has mean zero and variance

$$\frac{2 * \mathrm{V_2}}{\mathrm{(n-1)V_1}} \tag{74}$$

which vanishes as $\mathrm{n} \to \infty$, the term converges in probability to zero by Chebeychev's inequality [3] . Thus by Slutsky's Theorem [3] the entire right hand side of Equation 73 converges in distribution to $\mathrm{N}[0,1]$ as $\mathrm{n} \to \infty$ and $\mathrm{\bar{d}_{xx}}$ is asymptotically normal.

4.5 Two Sample Mean Distance: Asymptotic Normality

We have $\mathrm{X}_1, \dots, \mathrm{X}_\mathrm{n}$ an IID random sample with common mean μ_x and $\mathrm{Y}_1, \dots, \mathrm{Y}_\mathrm{m}$ an IID random sample with common mean μ_y. The sample mean squared Euclidean distance is:

$$\bar{\mathrm{d}}_{\mathrm{xy}} = \frac{1}{\mathrm{nm}} \sum_{\mathrm{i,j}} |\mathrm{X}_\mathrm{i} - \mathrm{Y}_\mathrm{j}|^2 \tag{75}$$

or by adding and subtracting $\mu_\mathrm{x} - \mu_\mathrm{y}$:

$$\frac{1}{\mathrm{nm}} \sum_{\mathrm{i,j}} |\mathrm{X}_\mathrm{i} - \mu_\mathrm{x} + \mu_\mathrm{x} - \mu_\mathrm{y} + \mu_\mathrm{y} - \mathrm{Y}_\mathrm{j}|^2 \tag{76}$$

which expands to a sum of the following terms:

$$\begin{aligned}
\mathrm{t}_1 &= \frac{1}{\mathrm{n}} \sum_{\mathrm{i}} \left(|\mathrm{X}_\mathrm{i} - \mu_\mathrm{x}|^2 + 2(\mu_\mathrm{x} - \mu_\mathrm{y})^\mathrm{T}(\mathrm{X}_\mathrm{i} - \mu_\mathrm{x}) \right) && (77)\\
\mathrm{t}_2 &= \frac{1}{\mathrm{m}} \sum_{\mathrm{j}} \left(|\mathrm{Y}_\mathrm{j} - \mu_\mathrm{y}|^2 - 2(\mu_\mathrm{x} - \mu_\mathrm{y})^\mathrm{T}(\mathrm{Y}_\mathrm{j} - \mu_\mathrm{y}) \right) && (78)\\
\mathrm{t}_3 &= -\frac{1}{\mathrm{nm}} \sum_{\mathrm{i,j}} 2(\mathrm{X}_\mathrm{i} - \mu_\mathrm{x})^\mathrm{T}(\mathrm{Y}_\mathrm{j} - \mu_\mathrm{y}) && (79)\\
\mathrm{t}_4 &= |\mu_\mathrm{x} - \mu_\mathrm{y}|^2 && (80)
\end{aligned}$$

Now t_1 is the mean of n IID terms. Suppose each has mean μ_1 and variance V_1. Similarly t_2 is the mean of m IID terms. Suppose each has mean μ_2 and variance V_2. Also t_3 is a mean of nm uncorrelated terms of mean zero (by the lemma). Suppose each has variance V_3. Noting that t_4 is a constant we can center and rescale $\bar{\mathrm{d}}_{\mathrm{xy}}$ to get:

$$\begin{aligned}
\frac{\sqrt{\mathrm{nm}}\left(\bar{\mathrm{d}}_{\mathrm{xy}} - (\mu_1 + \mu_2 + \mathrm{t}_4)\right)}{\sqrt{\mathrm{mV}_1 + \mathrm{nV}_2}} &= \frac{\sqrt{\mathrm{nm}}}{\sqrt{\mathrm{mV}_1 + \mathrm{nV}_2}}(\mathrm{t}_1 - \mu_1) && (81)\\
&+ \frac{\sqrt{\mathrm{nm}}}{\sqrt{\mathrm{mV}_1 + \mathrm{nV}_2}}(\mathrm{t}_2 - \mu_2) && (82)\\
&- \frac{\sqrt{\mathrm{nm}}}{\sqrt{\mathrm{mV}_1 + \mathrm{nV}_2}}\mathrm{t}_3 && (83)
\end{aligned}$$

The right hand side of Equation 81 can be re-written as:

$$\frac{\sqrt{\mathrm{nm}}}{\sqrt{\mathrm{mV_1+nV_2}}}(\mathrm{t_1}-\mu_1) = \frac{\sqrt{\mathrm{nm}}\sqrt{\mathrm{V_1}}}{\sqrt{\mathrm{mV_1+nV_2}}\sqrt{\mathrm{n}}}\frac{\sqrt{\mathrm{n}}(\mathrm{t_1}-\mu_1)}{\sqrt{\mathrm{V_1}}} \tag{84}$$

$$= \left(1+\frac{\mathrm{nV_2}}{\mathrm{mV_1}}\right)^{-\frac{1}{2}}\frac{\sqrt{\mathrm{n}}(\mathrm{t_1}-\mu_1)}{\sqrt{\mathrm{V_1}}} \tag{85}$$

Assuming that $\mathrm{n/m}$ is either constant or converges to a constant as $\mathrm{n}\to\infty$ and $\mathrm{m}\to\infty$ we have that the right hand side of Equation 85 converges to a constant times a $\mathrm{N}[0,1]$ random variable in distribution by the Central Limit and Slutsky's Theorems [3]. A similar result holds for Term 82.

Since Term 83 has mean value zero and variance

$$\frac{\mathrm{V_3}}{\mathrm{mV_1+nV_2}} \tag{86}$$

which vanishes as $\mathrm{n,m}\to\infty$ the term converges in probability to zero by Chebeychev's Inequality [3] and is ignorable in the limit. Thus, since $\mathrm{t_1}$ and $\mathrm{t_2}$ are asymptotically normal and independent for all values of n and m it follows that $\bar{\mathrm{d}}_{\mathrm{xy}}$ is asymptotically normal.

4.6 Two Sample Location Test

As above we assume $\mathrm{X_1,\dots,X_n}$ are an IID random sample with common mean μ_{x} and $\mathrm{Y_1,\dots,Y_m}$ are an IID random sample with common mean μ_{y}. We also have the one and two sample mean squared Euclidean distances:

$$\bar{\mathrm{d}}_{\mathrm{xx}} = \frac{2}{\mathrm{n(n-1)}}\sum_{\mathrm{i<j}}|\mathrm{X_i-X_j}|^2 \tag{87}$$

$$\bar{\mathrm{d}}_{\mathrm{yy}} = \frac{2}{\mathrm{m(m-1)}}\sum_{\mathrm{i<j}}|\mathrm{Y_i-Y_j}|^2 \tag{88}$$

$$\bar{\mathrm{d}}_{\mathrm{xy}} = \frac{1}{\mathrm{nm}}\sum_{\mathrm{i,j}}|\mathrm{X_i-Y_j}|^2 \tag{89}$$

A reasonable location test statistic is given by:

$$\Delta_{\mathrm{l}} = \bar{\mathrm{d}}_{\mathrm{xy}} - \frac{\bar{\mathrm{d}}_{\mathrm{xx}}+\bar{\mathrm{d}}_{\mathrm{yy}}}{2} \tag{90}$$

which has expected value $|\mu_x - \mu_y|^2$.

It follows from the identities derived above in 71 and 72 and 77 to 80 that Δ_l is the sum of the six following terms:

$$2(\mu_x - \mu_y)^T((\bar{X} - \mu_x) - (\bar{Y} - \mu_y)) \tag{91}$$

$$-2(\bar{X} - \mu_x)^T(\bar{Y} - \mu_y) \tag{92}$$

$$|\mu_x - \mu_y|^2 \tag{93}$$

$$\frac{2}{n(n-1)} \sum_{i<j} (X_i - \mu_x)^T (X_j - \mu_x) \tag{94}$$

$$\frac{2}{m(m-1)} \sum_{i<j} (Y_i - \mu_y)^T (Y_j - \mu_y) \tag{95}$$

A little algebra shows that term 94 is equivalent to:

$$\frac{1}{n(n-1)} \left(\sum_{i,j} (X_i - \mu_x)^T (X_i - \mu_x) - \sum_i |X_i - \mu_x|^2 \right) \tag{96}$$

or:

$$\frac{n}{(n-1)} |\bar{X} - \mu_x|^2 - \frac{1}{n(n-1)} \sum_i |X_i - \mu_x|^2 \tag{97}$$

Since a similar result holds for term 95 and since μ_x and μ_y are just place holders and can be replaced by $\bar{X}$ and $\bar{Y}$ we have:

$$\Delta_l = |\bar{X} - \bar{Y}|^2 - \frac{1}{n(n-1)} \sum_i |X_i - \bar{X}|^2 - \frac{1}{m(m-1)} \sum_i |Y_i - \bar{Y}|^2 \tag{98}$$

The second two terms in Equation 98 converge to zero in probability so the asymptotic distribution of Δ_l is that of $|\bar{X} - \bar{Y}|^2$. For example, for univariate data under the location null hypothesis, $H_0 : \mu_x = \mu_y$, $|\bar{X} - \bar{Y}|^2$ correctly scaled is asymptotically distributed as χ^2 with one degree of freedom. On the other hand if X and Y are high dimensional with IID components, $|\bar{X} - \bar{Y}|^2$ is asymptotically normal. Thus, the asymptotic distribution of Δ_l is not independent of the structure of the underlying data.

4.7 Two Sample Variance Test

We have $\mathrm{X}_1, \ldots, \mathrm{X}_\mathrm{n}$ and $\mathrm{Y}_1, \ldots, \mathrm{Y}_\mathrm{m}$ IID random samples from different distributions. We also have the one sample mean squared Euclidean distances:

$$\bar{\mathrm{d}}_{\mathrm{xx}} = \frac{2}{\mathrm{n}(\mathrm{n}-1)} \sum_{\mathrm{i}<\mathrm{j}} |\mathrm{X}_\mathrm{i} - \mathrm{X}_\mathrm{j}|^2 \tag{99}$$

$$\bar{\mathrm{d}}_{\mathrm{yy}} = \frac{2}{\mathrm{m}(\mathrm{m}-1)} \sum_{\mathrm{i}<\mathrm{j}} |\mathrm{Y}_\mathrm{i} - \mathrm{Y}_\mathrm{j}|^2 \tag{100}$$

$$\tag{101}$$

A reasonable variance test statistic is given by:

$$\Delta_\mathrm{v} = \bar{\mathrm{d}}_{\mathrm{xx}} - \bar{\mathrm{d}}_{\mathrm{yy}} \tag{102}$$

which has expected value $2(\mathrm{Tr}[\Sigma_\mathrm{x}] - \mathrm{Tr}[\Sigma_\mathrm{y}])$. Since $\bar{\mathrm{d}}_{\mathrm{xx}}$ and $\bar{\mathrm{d}}_{\mathrm{yy}}$ are independent and asymptotically normal, Δ_v is also asymptotically normal. Its variance can be computed from the one sample variance estimators derived above.

4.8 Two Sample Equivalance Test

As above we have two groups of vectors, $\mathrm{X}_\mathrm{i}, \ldots, \mathrm{X}_\mathrm{n}$ and $\mathrm{Y}_\mathrm{i}, \ldots, \mathrm{Y}_\mathrm{m}$. Suppose that the $\mathrm{X}_\mathrm{i}, \ldots, \mathrm{X}_\mathrm{n}$ are drawn from a reference group and we wish to test whether the $\mathrm{Y}_\mathrm{i}, \ldots, \mathrm{Y}_\mathrm{m}$ are equivalent to them in mean and variability. We have the one and two sample mean squared Euclidean distances:

$$\bar{\mathrm{d}}_{\mathrm{xx}} = \frac{2}{\mathrm{n}(\mathrm{n}-1)} \sum_{\mathrm{i}<\mathrm{j}} |\mathrm{X}_\mathrm{i} - \mathrm{X}_\mathrm{j}|^2 \tag{103}$$

$$\bar{\mathrm{d}}_{\mathrm{xy}} = \frac{1}{\mathrm{nm}} \sum_{\mathrm{i},\mathrm{j}} |\mathrm{X}_\mathrm{i} - \mathrm{Y}_\mathrm{j}|^2 \tag{104}$$

and a reasonable equivalence test statistic is given by:

$$\Delta_\mathrm{e} = \bar{\mathrm{d}}_{\mathrm{xy}} - \bar{\mathrm{d}}_{\mathrm{xx}} \tag{105}$$

which has expected value $|\mu_\mathrm{x} - \mu_\mathrm{y}|^2 + (\mathrm{Tr}[\Sigma_\mathrm{y}] - \mathrm{Tr}[\Sigma_\mathrm{x}])$.

Note that Δ_e should be small if each Y_i is drawn from the same distribution as each X_i but should be large if the two distributions differ in mean or the Y_i have greater variability.

Adding and subtracting means and doing a little algebra reduces Δ_e to a sum of the five terms:

$$\mathrm{t}_1 = \frac{1}{\mathrm{m}}\sum_{\mathrm{j}}\left(|\mathrm{Y_j}-\mu_\mathrm{y}|^2 - 2(\mathrm{Y_j}-\mu_\mathrm{y})^\mathrm{T}(\mu_\mathrm{x}-\mu_\mathrm{y})\right) \tag{106}$$

$$\mathrm{t}_2 = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}}\left(-|\mathrm{X_i}-\mu_\mathrm{x}|^2 + 2(\mathrm{X_i}-\mu_\mathrm{x})^\mathrm{T}(\mu_\mathrm{x}-\mu_\mathrm{y})\right) \tag{107}$$

$$\mathrm{t}_3 = -\frac{1}{\mathrm{nm}}\sum_{\mathrm{i,j}}2(\mathrm{X_i}-\mu_\mathrm{x})^\mathrm{T}(\mathrm{Y_j}-\mu_\mathrm{y}) \tag{108}$$

$$\mathrm{t}_4 = \frac{1}{\mathrm{n(n-1)}}\sum_{\mathrm{i<j}}4(\mathrm{X_i}-\mu_\mathrm{x})^\mathrm{T}(\mathrm{X_j}-\mu_\mathrm{x}) \tag{109}$$

$$\mathrm{t}_5 = \frac{1}{\mathrm{nm}}\sum_{\mathrm{i,j}}|\mu_\mathrm{x}-\mu_\mathrm{y}|^2 \tag{110}$$

Since terms 106 and 107 are independent means of IID terms and are thus asymptotically normal and independent, and since terms 108 and 109 converge to zero in probability at a rate proportional to 1/nm, we expect Δ_e to be asymptotically normal. In fact, a formal proof, similar to the proof given in Section 4.5 for the asymptotic normality of the two sample mean distance, shows this to be the case.

Note that under the hypothesis of equal variability we have

$$\mathrm{E}[\Delta_\mathrm{e}] = |\mu_\mathrm{x}-\mu_\mathrm{y}|^2 \tag{111}$$

so that under equal variability Δ_e is also a reasonable choice of test statistic for testing a location difference.

4.9 Simulation

A simulation was run in R to see if the asymptotic location, variability, and equivalence tests have normal distributions under the global null hypothesis. In addition the p-values of the asymptotic tests were compared to those of the permutation test.

Figures 1, 2, and 3 are based on two groups of twenty simulated “microarrays” each having only two “genes”. Thus each microarray is simply a pair of independent $\mathrm{N}[0,1]$ random variables. The distance measure used is mean squared Euclidean distance and 200 repetitions were run.

The upper panel is a normal QQ-plot of the asymptotic test z-statistic which should follow a 45 degree line if the sampling distribution is approximately $\mathrm{N}[0,1]$ as hypothesized.

The bottom panel plots the permutation test p-value versus the asymptotic test p-value together with a 45 degree reference line.

As can be seen the location test does not appear asymptotically normal, as expected, while the variability and equivalence tests perform very well and give p-values nearly equivalent to the permutation test.

Figure 1: Location Test

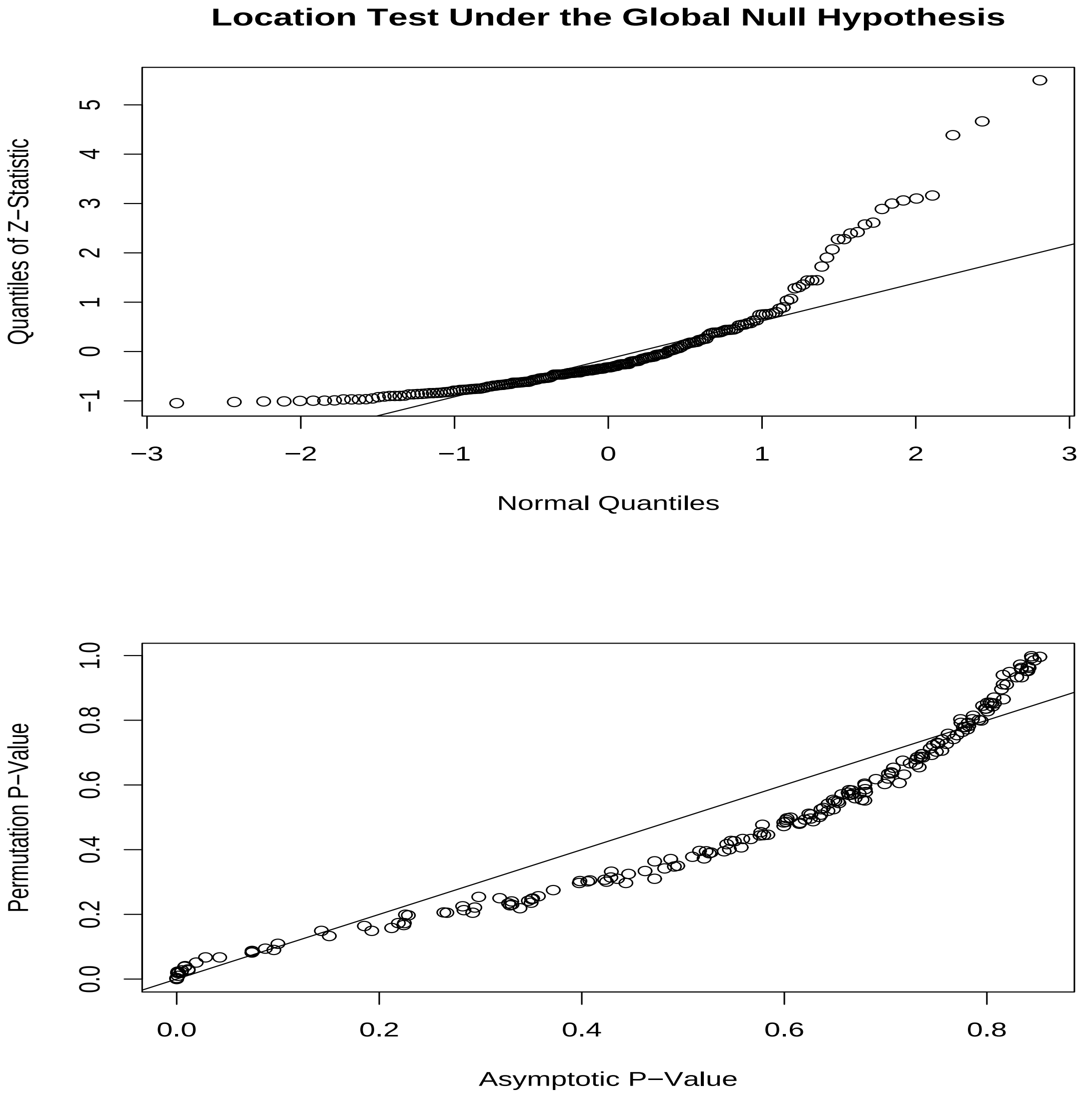


Figure 2: Variability Test

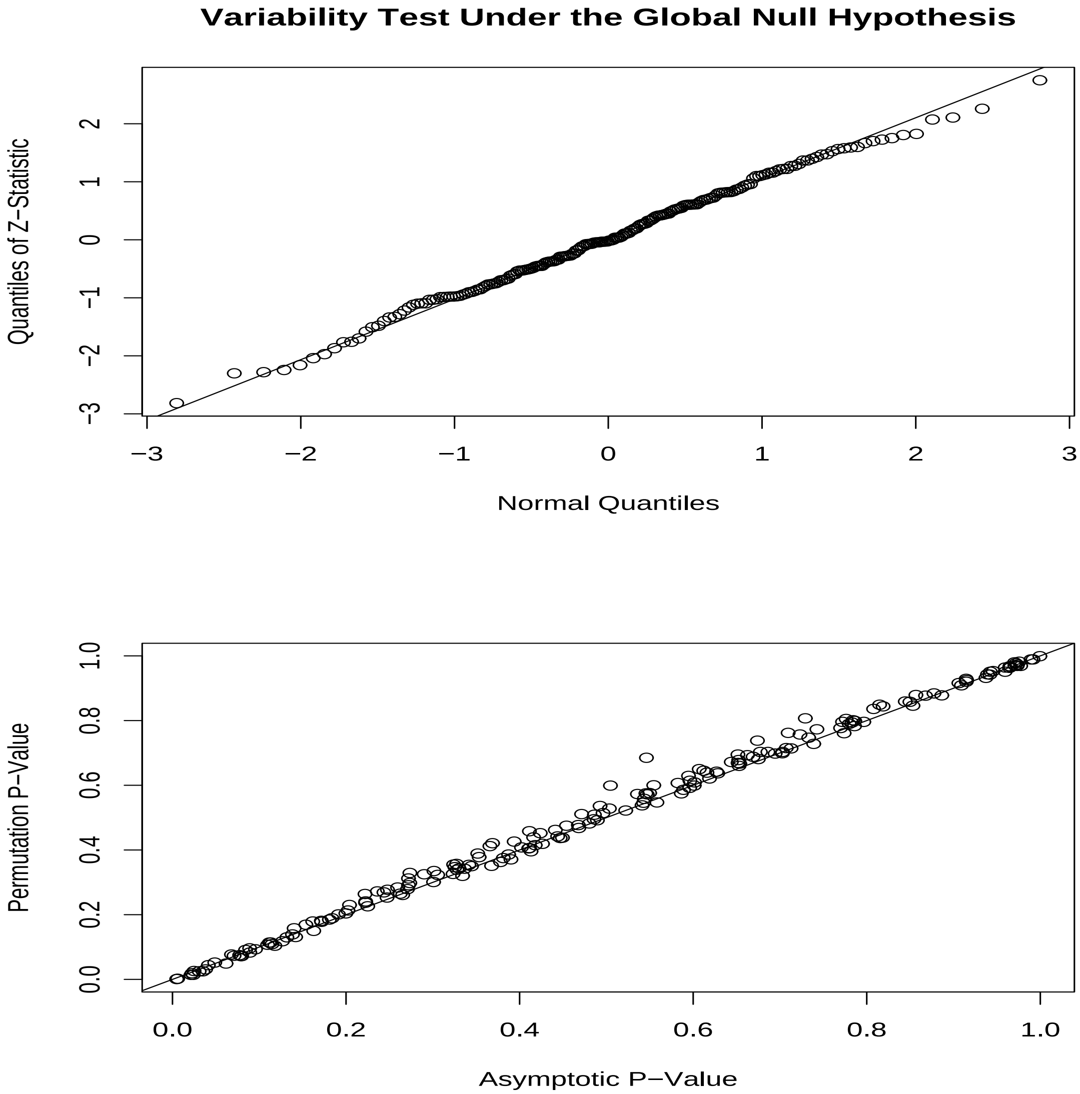
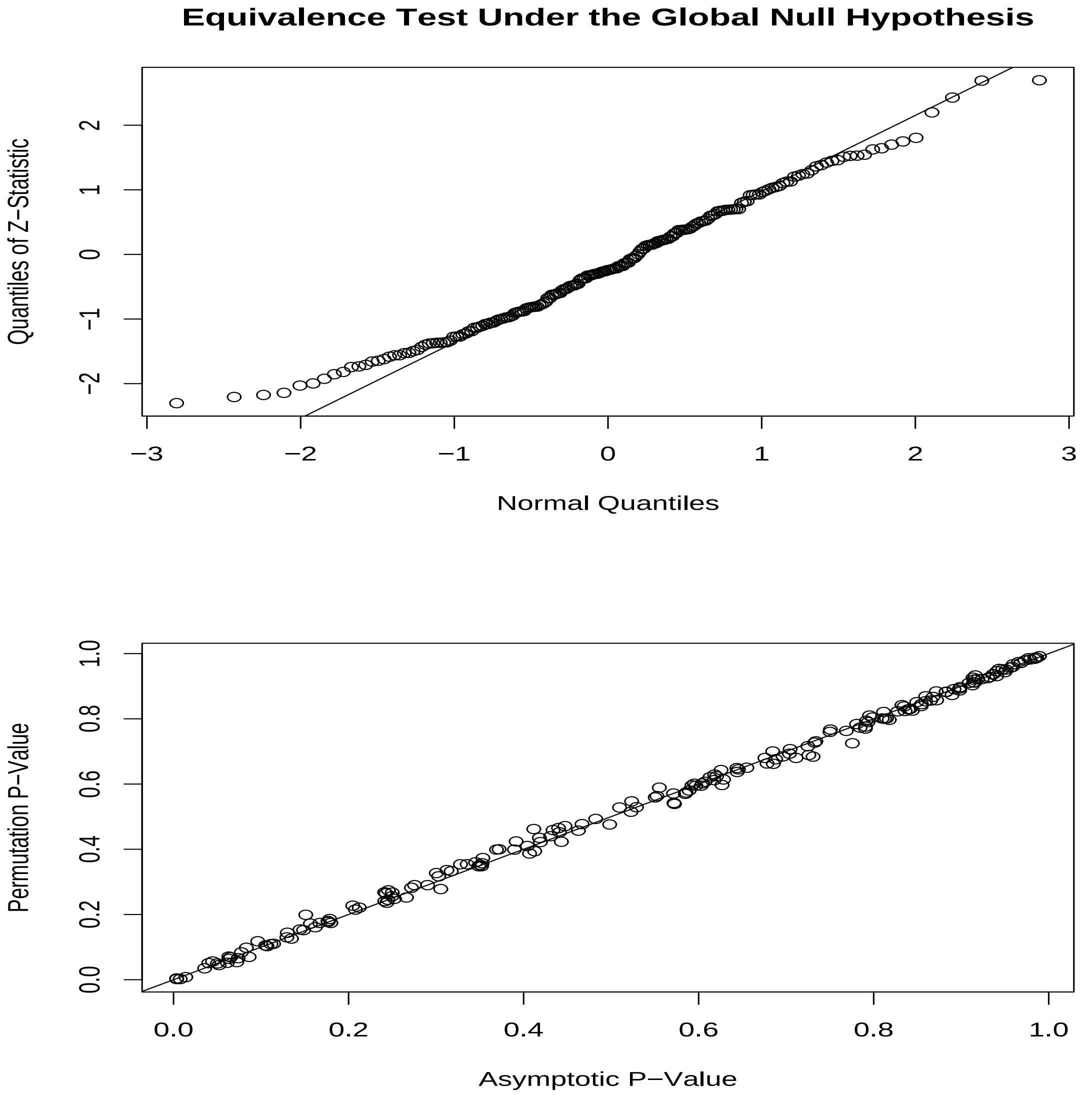


Figure 3: Equivalence Test



5 Speculative Alternatives to the Location Test

We note that the above proposed location test is not guaranteed to be asymptotically normal under squared Euclidean distance. Therefore we wish to consider several possible alternative ways of testing for a location difference.

5.1 "Magic Coordinate" Test

The idea of the magic coordinate test is to project all the data onto a direction of maximal separation and then simply do a univariate test on this line. The direction will need to be found by some algorithm such as support vector machines. We examine here how this might work out.

We have as usual two groups of vectors, $\mathrm{X_i, \ldots, X_n}$ and $\mathrm{Y_i, \ldots, Y_m}$, and a specified direction given by a unit vector C. Choose any point, say B, as a base point and project all the data onto the line through B in the C direction. For the vector $\mathrm{X_i}$ the signed length from B in the C direction is given by:

$$\mathrm{L(X_i) = C^T(X_i - B)} \tag{112}$$

thus the average of $\mathrm{L(X_i)}$ over $\mathrm{X_i, \ldots, X_n}$ is:

$$\begin{aligned} \mathrm{\frac{1}{n}\sum_i L(X_i)} &= \mathrm{\frac{1}{n}\sum_i C^T(X_i - B)} && (113) \\ &= \mathrm{C^T(\bar{X} - B)} && (114) \end{aligned}$$

Note that 113 is an average of IID terms so that 114 is asymptotically normal. Further it has sample variance:

$$\mathrm{V_x = \frac{1}{n(n-1)}\sum_i \left(C^T(X_i - \bar{X})\right)^2} \tag{115}$$

and since similar results hold for $\mathrm{Y_i, \ldots, Y_m}$ we have under the null hypothesis, $\mathrm{H_0} : \mu_\mathrm{x} = \mu_\mathrm{y}$, that:

$$\Delta_\mathrm{L} = \mathrm{\frac{C^T(\bar{X} - \bar{Y})}{\sqrt{V_x + V_y}}} \tag{116}$$

is an asymptotically $\mathrm{N}[0, 1]$ test for a location difference in the C direction.

Lets look more closely at the variance estimator. Letting the subscripts j and k represent the jth and kth components of each vector we have:

$$
\begin{aligned}
\mathrm{V_x} &= \frac{1}{\mathrm{n(n-1)}}\sum_{\mathrm{i}}\left(\sum_{\mathrm{j}}\mathrm{C_j(X_{ij}-\bar{X}_j)}\right)^2 && (117)\\
&= \frac{1}{\mathrm{n(n-1)}}\sum_{\mathrm{i}}\sum_{\mathrm{j}}\sum_{\mathrm{k}}\mathrm{C_j(X_{ij}-\bar{X}_j)C_k(X_{ik}-\bar{X}_k)} && (118)\\
&= \sum_{\mathrm{j}}\sum_{\mathrm{k}}\mathrm{C_jC_k}\frac{\sum_{\mathrm{i}}\mathrm{(X_{ij}-\bar{X}_j)(X_{ik}-\bar{X}_k)}}{\mathrm{n(n-1)}} && (119)\\
&= \sum_{\mathrm{j}}\sum_{\mathrm{k}}\mathrm{C_jC_k\widehat{cov}(\bar{X}_j,\bar{X}_k)} && (120)\\
&= \mathrm{C^T\widehat{\Sigma}_{\bar{X}}C} && (121)
\end{aligned}
$$

where $\widehat{\Sigma}_{\bar{X}}$ is the sample covariance matrix of $\bar{X}$. Thus V_x is the well known sample estimator of the variance of $C^T\bar{X}$.

Since the direction C has been chosen to maximally separate the two groups the size of the test based Δ_L is smaller than the nominal size of the normal Z-test so that the test is anti-conservative. Second the maximal value of $C^T(\bar{X}-\bar{Y})$ occurs when C is parallel to $(\bar{X}-\bar{Y})$ but the maximal value of Δ_L depends on $(\bar{X}-\bar{Y})$, $\widehat{\Sigma}_{\bar{X}}$, and $\widehat{\Sigma}_{\bar{Y}}$.

5.2 Split Sample Location Test

If, as was shown above, the average within group and between group mean squared Euclidean distance is indeed asymptotically normal, then an asymptotically normal location test can be constructed by splitting each group of microarrays in half and using one half to estimate the between group mean distance and the other to estimate the within group mean distance. Since these estimates are independent and asymptotically normal any linear combination will also be asymptotically normal.

So suppose we have two groups each with - for notational convenience - an even number of microarrays $X_i, \ldots, X_{2n}$ and $Y_i, \ldots, Y_{2m}$. We can randomly split them into two groups each of the form $X_{ai}, \ldots, X_{an}$ and $X_{bi}, \ldots, X_{bn}$ and

$\mathrm{Y_{ai}}, \ldots, \mathrm{Y_{am}}$ and $\mathrm{Y_{bi}}, \ldots, \mathrm{Y_{bm}}$. Define the location test statistic:

$$\Delta_{\mathrm{l}} = \frac{1}{\mathrm{nm}} \sum_{\mathrm{i,j}} \mathrm{d}(\mathrm{X_{ai}}, \mathrm{Y_{aj}}) - \frac{1}{2} \left(\frac{2}{\mathrm{n(n-1)}} \sum_{\mathrm{i<j}} \mathrm{d}(\mathrm{X_{bi}}, \mathrm{X_{bj}}) + \frac{2}{\mathrm{m(m-1)}} \sum_{\mathrm{i<j}} \mathrm{d}(\mathrm{Y_{bi}}, \mathrm{Y_{bj}}) \right) \tag{122}$$

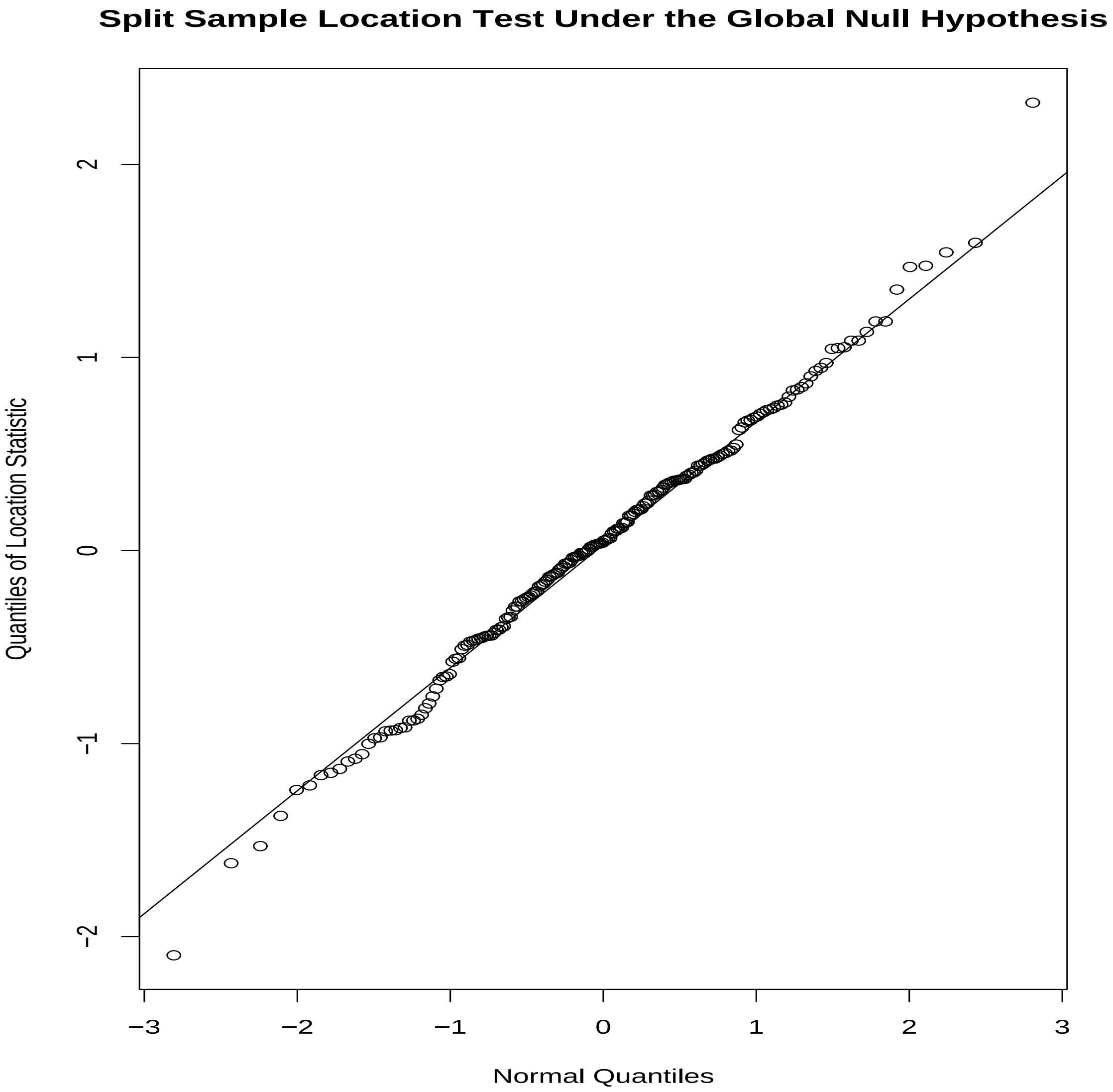
The statistic Δ_{l} is asymptotically normal, has expected value $|\mu_{\mathrm{x}} - \mu_{\mathrm{y}}|^2$, and we know how to estimate its variance from work done above. It is necessary, of course, to prove that the random partition into half groups minimizes the variance of Δ_{l} over all possible partitions of each group into two sub-groups. Also Δ_{l} has the obvious problem that a different random partition into half groups will yield a slightly different p-value or confidence interval.

We performed a simulation in R to see if the split sample location test has a normal distribution under the global null hypothesis. To keep the coding simple we did not scale Δ_{l} by its estimated standard error.

Figure 4 is a normal QQ-plot of Δ_{l} based on two groups of twenty simulated "microarrays" each having only two "genes". Thus each microarray is simply a pair of independent $\mathrm{N}[0, 1]$ random variables. The distance measure used is mean squared Euclidean distance and 200 repetitions were run.

The figure should follow a straight line if the sampling distribution is approximately normal as hypothesized so that, as can be seen, the distribution of Δ_{l} is nearly normal.

Figure 4: Split Sample Location Test



As a way to finesse the arbitrary splitting we considered the possibility of averaging the test statistic over several splits and then comparing the average (suitably scaled) to the percentiles of the standard normal despite the fact that we do not know the sampling distribution of the average. The argument for such a test follows from the assumption that the distribution of the sample average of identically distributed (but not necessarily independent) random variables must narrow so that (with some thought) such a test will in fact be conservative without sacrificing power under alternatives that would have "reasonable" power under the split sample test.

In fact this premise turns out to be false as there is a way to construct a simple counter example for a pair of identically distributed discrete random variables. The basic idea is to construct correlated random variables with most of the mass in the tails.

We present an example for the identically distributed random variables X_1 and X_2 which take the values $\{1, 2, 3, 4, 5, 6, 7, 8\}$. As will be evident this construction can easily be extended to discrete random variables taking an arbitrarily large number of values. Table 3 shows the joint distribution of X_1 and X_2 although it has *not* been normalized to have total mass equal to one so that the pattern in how the probability mass is distributed is evident.

Letting $\bar{\mathrm{X}} = (\mathrm{X}_1 + \mathrm{X}_2)/2$ it can be seen from the joint distribution that:

$$\mathrm{P}[\bar{\mathrm{X}} \geq 8] < \mathrm{P}[\mathrm{X}_1 \geq 8] \tag{123}$$

but:

$$\mathrm{P}[\bar{\mathrm{X}} \geq 7] > \mathrm{P}[\mathrm{X}_1 \geq 7] \tag{124}$$

and:

$$\mathrm{P}[\bar{\mathrm{X}} \geq 6] > \mathrm{P}[\mathrm{X}_1 \geq 6] \tag{125}$$

and:

$$\mathrm{P}[\bar{\mathrm{X}} \geq 5] = \mathrm{P}[\mathrm{X}_1 \geq 5] \tag{126}$$

and a symmetrical result holds for $\mathrm{P}[\bar{\mathrm{X}} \leq 1], \mathrm{P}[\bar{\mathrm{X}} \leq 2], \mathrm{P}[\bar{\mathrm{X}} \leq 3]$, and $\mathrm{P}[\bar{\mathrm{X}} \leq 4]$ so that except for the extreme values 1 and 8 the distribution of $\bar{\mathrm{X}}$ is essentially broader than the distribution of X_1.

Table 3: Joint Probability Distribution of X_1 and X_2

8	0	0	0	0	$\frac{1}{64}$	$\frac{1}{8}$	$\frac{1}{2}$	1
7	0	0	0	0	$\frac{1}{128}$	$\frac{1}{16}$	$\frac{1}{4}$	$\frac{1}{2}$
6	0	0	0	0	$\frac{1}{256}$	$\frac{1}{32}$	$\frac{1}{16}$	$\frac{1}{8}$
5	0	0	0	0	$\frac{1}{512}$	$\frac{1}{256}$	$\frac{1}{128}$	$\frac{1}{64}$
4	$\frac{1}{64}$	$\frac{1}{128}$	$\frac{1}{256}$	$\frac{1}{512}$	0	0	0	0
3	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{256}$	0	0	0	0
2	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{16}$	$\frac{1}{128}$	0	0	0	0
1	1	$\frac{1}{2}$	$\frac{1}{8}$	$\frac{1}{64}$	0	0	0	0
	1	2	3	4	5	6	7	8

6 The Permutation Test

6.1 Introduction

This section describes the permutation test in the simplest case of two independent groups and squared Euclidean distance.

Consider an experiment comparing two levels of some experimental factor that might influence apparent gene expression. The gene expression values, or some function of the gene expression values which measures the biological signal for each gene, can be represented by two groups of column vectors of signal values. We will represent these column vectors by $\mathrm{X}_{1,1}, \ldots, \mathrm{X}_{1,\mathrm{N}_1}$ and $\mathrm{X}_{2,1}, \ldots, \mathrm{X}_{2,\mathrm{N}_2}$, for factor levels 1 and 2 respectively where N_1 and N_2 are the number of arrays in groups 1 and 2 respectively. For simplicity we will refer to these vectors of signal values as microarrays. Let $\mathrm{D}[\mathrm{X}_{\mathrm{i,j}}, \mathrm{X}_{\mathrm{k,l}}]$ be the dissimilarity or distance between two microarrays.

Inference concerning the variability and location of groups 1 and 2 can be based on the three means:

$$\bar{\mathrm{D}}_{11} = \frac{2}{\mathrm{N}_1(\mathrm{N}_1 - 1)} \sum_{\mathrm{i}<\mathrm{j}\leq \mathrm{N}_1} \mathrm{D}[\mathrm{X}_{1,\mathrm{i}}, \mathrm{X}_{1,\mathrm{j}}] \tag{127}$$

$$\bar{\mathrm{D}}_{22} = \frac{2}{\mathrm{N}_2(\mathrm{N}_2 - 1)} \sum_{\mathrm{i}<\mathrm{j}\leq \mathrm{N}_2} \mathrm{D}[\mathrm{X}_{2,\mathrm{i}}, \mathrm{X}_{2,\mathrm{j}}] \tag{128}$$

$$\bar{\mathrm{D}}_{12} = \frac{1}{\mathrm{N}_1\mathrm{N}_2} \sum_{\mathrm{i}\leq \mathrm{N}_1, \mathrm{j}\leq \mathrm{N}_2} \mathrm{D}[\mathrm{X}_{1,\mathrm{i}}, \mathrm{X}_{2,\mathrm{j}}] \tag{129}$$

Where

1. $\bar{\mathrm{D}}_{11}$ is the mean distance between microarrays within group 1.
2. $\bar{\mathrm{D}}_{22}$ is the mean distance between microarrays within group 2.
3. $\bar{\mathrm{D}}_{12}$ is the mean distance between microarrays between groups 1 and 2.

6.2 Additive Errors and Squared Euclidean Distance

In this section we derive the expected value of $\bar{\mathrm{D}}_{11}$, $\bar{\mathrm{D}}_{22}$, and $\bar{\mathrm{D}}_{12}$ assuming a completely randomized design, an additive error model, and squared Euclidean distance. Under these assumptions we have:

$$\mathrm{X}_{1,\mathrm{i}} = \mu_1 + \epsilon_{1,\mathrm{i}} \tag{130}$$

$$\mathrm{X}_{2,\mathrm{i}} = \mu_2 + \epsilon_{2,\mathrm{i}} \tag{131}$$

where μ_1 and μ_2 are the respective mean vectors of the microarrays in groups 1 and 2 and $\epsilon_{1,\mathrm{i}}$ and $\epsilon_{2,\mathrm{i}}$ are random error vectors with expected value 0 and variance covariance matrices Σ_1 and Σ_2 respectively. The errors are assumed to be independent across microarrays.

The squared Euclidean distance between any two microarrays, $\mathrm{X}_{\mathrm{i,j}}$ and $\mathrm{X}_{\mathrm{k,l}}$is

$$|\mathrm{X}_{\mathrm{i,j}} - \mathrm{X}_{\mathrm{k,l}}|^2 = (\mathrm{X}_{\mathrm{i,j}} - \mathrm{X}_{\mathrm{k,l}})^{\mathrm{T}}(\mathrm{X}_{\mathrm{i,j}} - \mathrm{X}_{\mathrm{k,l}}) \tag{132}$$

and

$$\mathrm{E}[\bar{\mathrm{D}}_{11}] = 2\mathrm{Tr}[\Sigma_1] \tag{133}$$

$$\mathrm{E}[\bar{\mathrm{D}}_{22}] = 2\mathrm{Tr}[\Sigma_2] \tag{134}$$

$$\mathrm{E}[\bar{\mathrm{D}}_{12}] = |\mu_1 - \mu_2|^2 + \mathrm{Tr}[\Sigma_1] + \mathrm{Tr}[\Sigma_2] \tag{135}$$

where $\mathrm{Tr}[]$ is the trace operator which gives the sum of the diagonal elements of a matrix.

We can now define test statistics to compare the variability and location of groups 1 and 2. To compare variability let:

$$\Delta_{\mathrm{v}} = \bar{\mathrm{D}}_{11} - \bar{\mathrm{D}}_{22} \tag{136}$$

and to compare location let:

$$\Delta_{\mathrm{l}} = \bar{\mathrm{D}}_{12} - \frac{\bar{\mathrm{D}}_{11} + \bar{\mathrm{D}}_{22}}{2} \tag{137}$$

These test statistics have expected values:

$$\mathrm{E}[\Delta_{\mathrm{v}}] = 2(\mathrm{Tr}[\Sigma_1] - \mathrm{Tr}[\Sigma_2]) \tag{138}$$

and:

$$\mathrm{E}[\Delta_{\mathrm{l}}] = |\mu_1 - \mu_2|^2 \tag{139}$$

Inference concerning the magnitude of Δ_{v} and Δ_{l} can be made using a permutation test. Each permutation consists of assigning N_1 microarrays to group 1 and the remaining N_2 to group 2. Note that for each permutation the pairwise distances are simply re-indexed, they do not have to be recalculated. Only the values of $\bar{\mathrm{D}}_{11}$, $\bar{\mathrm{D}}_{22}$, and $\bar{\mathrm{D}}_{12}$ and Δ_{v} and Δ_{l} have to be recalculated based on the re-indexing.

Let $\Delta_{\mathrm{v}}^{\mathrm{obs}}$ and $\Delta_{\mathrm{l}}^{\mathrm{obs}}$ be the observed values of Δ_{v} and Δ_{l} and let Δ_{v}^{*} and Δ_{l}^{*} be the values from a permutation. If there are a total of B permutations, and assuming $\Delta_{\mathrm{v}}^{\mathrm{obs}} > 0$, then

$$\mathrm{p_v} = \frac{\mathrm{Number}[\Delta_{\mathrm{v}}^{*} \geq \Delta_{\mathrm{v}}^{\mathrm{obs}}]}{\mathrm{B}} \tag{140}$$

is a one-sided p-value [4] for rejecting the null hypothesis that $\mathrm{Tr}[\Sigma_1] = \mathrm{Tr}[\Sigma_2]$. If $\Delta_{\mathrm{v}}^{\mathrm{obs}} < 0$ then the inequality in Equation 140 is simply reversed. Similarly

$$\mathrm{p_l} = \frac{\mathrm{Number}[\Delta_{\mathrm{l}}^{*} \geq \Delta_{\mathrm{l}}^{\mathrm{obs}}]}{\mathrm{B}} \tag{141}$$

is a one-sided p-value for rejecting the null hypothesis that $\mu_1 = \mu_2$.

Sometimes investigators design an experiment to compare a new method to a proven “gold standard”. In such a case interest centers on showing that the new method is equivalent to the gold standard. To be equivalent it should not differ in mean and not exhibit greater variability. Assuming that the microarrays in group 1 were prepared using the gold standard, a summary statistic which can be used to reject the null hypothesis of equivalence is given by:

$$\Delta_{\mathrm{e}} = \bar{\mathrm{D}}_{12} - \bar{\mathrm{D}}_{11} \tag{142}$$

If this statistic is large, then group 2 either has a different mean or more var iability than group 1. This can easily be seen from its expected value under mean squared Euclidean distance:

$$\mathrm{E}[\Delta_{\mathrm{e}}] = |\mu_1 - \mu_2|^2 + (\mathrm{Tr}[\Sigma_2] - \mathrm{Tr}[\Sigma_1]) \tag{143}$$

This also makes intuitive sense. If the two methods are equivalent then the distance between the microarrays in groups 1 and 2 should not be any greater than the distance between the microarrays within group 1. Inference concerning the magnitude of Δ_{e} can be made using a permutation test exactly as for Δ_{l} and Δ_{v}.

It should be noted that the statistics Δ_{v}, Δ_{l}, and Δ_{e} are all special cases of Mantel’s U statistic [5] and Δ_{l} is similar to a special case of the MRPP statistic [6].

References

[1] van der Vaart, AW. *Asymptotic Statistics* Cambridge University Press 1998 p 7

[2] van der Vaart, AW. *Asymptotic Statistics* Cambridge University Press 1998 p 14

[3] Casella G. and Berger RL. *Statistical Inference* Wadsworth and Brookes/Cole 1990

[4] Efron B., Tibshirani RJ. *An Introduction to the Bootstrap* Chapman & Hall/CRC, 1993; pp202-19

[5] Mantel N. *The Detection of Disease Clustering and a Generalized Regression Approach* Cancer Research, February 1967, 27 Part 1, pp 209-220

[6] Berry KJ., Kvamme KL., Mielke Jr. PW. *Improvements in the Permutation Test for the Spatial Analysis of the Distribution of Artifacts into Classes* American Antiquity, 1983, Vol. 48, No. 3 pp 547-553