Non-linear tests for identifying differentially expressed genes or genetic networks

Hao Xiong
Department of Computer Science, Texas A&M University, 301 Harvey R. Bright Bldg, College Station, TX 77843-3112, USA

ABSTRACT
Motivation: One of the recently developed statistics for identifying differentially expressed genetic networks is Hotelling $T^2$ statistic, which is a quadratic form of difference in linear functions of means of gene expressions between two types of tissue samples, and so their power is limited.

Results: To improve the power of test statistics, a general statistical framework for construction of non-linear tests is presented, and two specific non-linear test statistics that use non-linear transformations of means are developed. Asymptotical distributions of the non-linear test statistics under the null and alternative hypothesis are derived. It has been proved that under some conditions the power of the non-linear test statistics is higher than that of the $T^2$ statistic. Besides theory, to evaluate in practice the performance of the non-linear test statistics, they are applied to two real datasets. The preliminary results demonstrate that the $P$-values of the non-linear statistics for testing differential expressions of the genetic networks are much smaller than those of the $T^2$ statistic. And furthermore simulations show the Type I errors of the non-linear statistics agree with the threshold used and the statistics fit the $\chi^2$ distribution.

Contact: hxiong@cs.tamu.edu

Supplementary information: Supplementary data are available on Bioinformatics online.

1 INTRODUCTION

Microarray technology can simultaneously measure expression levels of thousands or even ten thousands of genes and produce an avalanche of data. It did not take long before scientists avail themselves of this valuable tool in studying variation of genomewide gene expression over different tissue samples, different experimental conditions or different time points of biological process (Brown and Botstein, 1999).

It is widely recognized that the conditions of the cell and cellular processes are influenced by a large number of genes interwoven into networks, rather than a few genes (Strohman, 2002). To study individual biological components alone is not sufficient to discover the rules underlying complex biological systems. Therefore, a systems level study of genetic networks and identification of differentially expressed genetic networks holds the key to unraveling the relationship between genotype and phenotype (Xiong et al., 2004; Khalil and Hill, 2005; Lu et al., 2005).

The ideal statistics for testing differentially expressed genetic networks should have high power while keeping false positive rates at a specified level. A multivariate statistic for testing differentially expressed genetic networks is Hotelling $T^2$ statistic (Anderson, 1984; Xiong et al., 2004; Lu et al., 2005). However, it is a quadratic function on the difference of means of expression levels between two types of tissue samples (e.g. tumor and normal tissue samples), and the difference is a linear function of expression levels. One strategy to improve the power of test statistics is to amplify difference in the means of gene expression. A natural way to amplify such difference is to transform gene expression levels. It is not difficult to show that any linear transformation of gene expression in $T^2$ statistics will not change their pre-transformation values. To overcome this problem, I propose to use non-linear transformations of means of gene expression in normal tissue ($X$) and abnormal tissue ($Y$), i.e. $f(X)$ and $f(Y)$, expecting statistics based on difference $|f(X) - f(Y)|$ will be more powerful than those based on $|X - Y|$

The main purpose of this report is to develop a statistic framework for constructing non-linear statistics for testing differentially expressed genes or genetic networks and propose several non-linear statistics for gene expression data analysis. To do so, I first investigate the properties of non-linear transformation, then study how to construct test statistics based on them and derive asymptotic distributions of non-linear test statistics under the null and alternative hypothesis. Since different non-linear tests may have different power, selection of non-linear statistics is critical to the successful application of non-linear tests to gene expression data analysis. I compare the power of several non-linear test statistics and Hotelling $T^2$ statistic. Finally, to evaluate their performance non-linear test statistics are applied to two real gene expression datasets and simulations are run to obtain Type I error and distribution of statistics.

2 METHODS

2.1 Non-linear functions of sample means of gene expressions and their distributions

For convenience of theoretical analysis, I assume that the number of tissue samples is large enough to allow application of large sample theory to gene expression data analysis although in practice the assumption often does not hold. Under this assumption, the sample mean of gene expression will be asymptotically distributed as normal distributions. Let $X$ be a vector of means of gene expressions. Then, $X$ will asymptotically have a normal distribution $N(\mu, \frac{1}{n}\Sigma)$, where

$$\mu = E[X], \quad \text{Cov}(X, X) = \frac{1}{n}\Sigma.$$
functions, a Hessian matrix of non-linear function and Jacobian and Hessian matrices of the non-linear functions. If the norm of the mean expressions between abnormal and normal tissues sample, \( f(Y) - f(X) \) is asymptotically distributed as the following multivariate normal distribution

\[
    f(Y) - f(X) \sim N(f(\mu_X) - f(\mu_Y), \Lambda)
\]

where \( \mu_X = E[X] \), \( B = (\partial^2 f/\partial \mu^2) \) a Jacobian matrix of the vector of functions, \( \Sigma = nCov(X, X) \). Difference in non-linear transformation of means of gene expressions between abnormal and normal tissues sample:

\[
    \mu_Y = E[Y], \Lambda = \frac{1}{n} \Sigma + \frac{1}{n} B \Sigma B^T, C = \left( \frac{\partial f}{\partial \mu_Y} \right)^T, \Sigma_Y = n \text{Cov}(Y, Y),
\]

\( B \) and \( \Sigma \) are defined as before, \( n \) and \( n \) are the number of abnormal and normal tissue samples respectively.

The principle behind \( T^2 \) test statistics in gene expression data analysis is to compare difference in the means of gene expression in abnormal and normal samples, and that difference, if amplified, may improve power to identify differentially expressed genes or genetic networks. One strategy to amplify the difference is to transform non-linearly the means of gene expressions so that the non-linear difference should be larger. Therefore, the goal is to search for such non-linear transformations. With this goal in mind, I have investigated how difference of non-linear transformation of means can amplify difference of the means in gene expressions.

Define

\[
    f(\mu) = [f_1(\mu), \ldots, f_s(\mu)]^T, H_i = \left( \frac{\partial f_i}{\partial \mu_i, \partial \mu_j} \right)^T,
\]

a Hessian matrix of non-linear function \( f(\mu) \) and \( H = \begin{bmatrix} H_1 & \vdots & H_s \end{bmatrix} \), an array.

Then by Taylor expansion, we have (Bates and Watts, 1980)

\[
    f(\mu_Y) - f(\mu_X) \approx B(\mu_Y - \mu_X) + \frac{1}{2}(\mu_Y - \mu_X)^T H(\mu_Y - \mu_X)
\]

\[
    = [B + \frac{1}{2} H(\mu_Y - \mu_X)](\mu_Y - \mu_X)
\]

(1)

From Equation (1), the difference in the non-linear functions of the means of the gene expressions between abnormal and normal tissues depends on the Jacobian and Hessian matrices of the non-linear functions. If the norm of the coefficient matrix of the vector \( \mu_Y - \mu_X \) satisfies

\[
    \|B + \frac{1}{2} H(\mu_Y - \mu_X)\| > 1,
\]

then

\[
    \|f(\mu_Y) - f(\mu_X)\| > \|\mu_Y - \mu_X\|
\]

which implies that the norm of differences in non-linear functions of the means of the gene expressions between the abnormal and normal tissues is larger than that of the original difference in the means of the gene expressions under this condition. The matrix \( B + (1/2) H(\mu_Y - \mu_X) \) characterizes the strength of non-linearity of the non-linear functions (Bates and Watts, 1980) and hence provides information for searching non-linear functions which can be used to construct non-linear test statistics with high power.

### 2.2 Test statistics

The results of non-linear functions of asymptotically norm random vectors can be used to construct non-linear test statistics for testing differential expressions of genes or genetic networks. The quadratic form \( X^T \Sigma_X X \) of asymptotically norm random vectors provides a statistic framework for construction of test statistics.

Suppose that there are \( k \) genes in the pathway being tested. Let \( x_i \) be the expression of the \( j \)-th gene in the \( i \)-th normal tissue sample and \( y_j \) be the expression of the \( j \)-th gene in the \( i \)-th abnormal tissue sample. Define

\[
    X_i = (x_{i1}, \ldots, x_{ik})^T, Y_i = (y_{i1}, \ldots, y_{ik})^T
\]

\[
    X = \frac{1}{n} \sum_{i=1}^n X_i, Y = \frac{1}{n} \sum_{i=1}^n Y_i
\]

\[
    f(X) = [f_1(X), f_2(X), \ldots, f_s(X)]^T, f(Y) = [f_1(Y), f_2(Y), \ldots, f_s(Y)]^T
\]

The pooled-sample variance-covariance matrix of the indicator variables for the marker genotypes is defined as

\[
    S = \frac{1}{n + m - 2} \left( \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^T + \sum_{i=1}^n (Y_i - \bar{Y})(Y_i - \bar{Y})^T \right)
\]

Let \( A \) be an estimator of matrix \( \Lambda \), where \( \Lambda = (1/n) \Sigma_X \Sigma_Y + (1/n) \Sigma_X \Sigma_Y^T \) as defined before. Under the null hypothesis, we have \( B = C \) and \( \Sigma = \Sigma_Y^T \). The estimator \( A \) can be obtained by substituting pooled-sample estimation of the covariance matrices \( \Sigma_X, \Sigma_Y \) and pooled estimation of the Jacobian matrices \( B, C \) into the equation defining the matrix \( \Lambda \). For convenience, estimator \( \tilde{\Lambda} \) under the null hypothesis, will be denoted by \( A_0 \). The non-linear statistics can be defined as

\[
    T_n = [f(Y) - f(X)](\tilde{\Lambda} - f(X) - f(X))^T
\]

(2)

where \( (\tilde{\Lambda})^{-1} \) is the generalized inverse of the matrix \( \tilde{\Lambda} \). Let \( r = \text{rank}(\tilde{\Lambda}) \). It can be shown (Greenwood and Nikulin, 1996) that under the null hypothesis of no differential expressions of the gene or genetic network, i.e., \( H_0: \mu_1 = \mu_2, \) the statistic \( T_n \) is asymptotically distributed as a central \( \chi^2_n \) distribution, and under the alternative hypothesis \( H_1: \mu_1 \neq \mu_2 \), the statistic \( T_n \) is asymptotically distributed as a non-central \( \chi^2_n \) distribution with the following non-centrality parameter:

\[
    \lambda_n = [f(\mu_2) - f(\mu_1)]^T \tilde{\Lambda}^{-1} [f(\mu_2) - f(\mu_1)]
\]

(3)

Now we consider a special vector-valued non-linear function. Let \( g(x) \) be a real valued non-linear function and has a non-zero derivative at its mean \( E[x] = \mu \). Define

\[
    f(\mu) = [g(\mu_1), \ldots, g(\mu_k)]^T, f(x) = [g(x_1), \ldots, g(x_k)]^T
\]

where \( \mu_1, \ldots, \mu_k \) are the parameters and \( \mu = [\mu_1, \ldots, \mu_k]^T \). Thus, we have

\[
    f(X) = [g(x_1), \ldots, g(x_k)]^T, f(Y) = [g(y_1), \ldots, g(y_k)]^T
\]

(4)

Under this definition, Jacobian matrices \( B, C \) have the following simple forms:

\[
    b_{ij} = \frac{\partial g(\mu_j)}{\partial \mu_i}, c_{ij} = 0 (i \neq j) = \frac{\partial g(\mu_j)}{\partial \mu_j}
\]

(5)

Test statistic \( T_n \) in Equation (2) defines a class of non-linear tests. Various non-linear functions satisfying some regularity conditions can be used to construct the test statistics. Table 1 lists some of non-linear functions used in this study and their corresponding derivatives.

### 2.3 Comparisons of power of non-linear test statistics and the Hotelling’s \( T^2 \) statistic by approximate formula

To evaluate the performance of non-linear statistics for testing differential expressions of genes or genetic networks, we need to compare the power of the non-linear test statistics and the Hotelling’s \( T^2 \) statistic. Calculation of the power of non-linear test statistic is based on computation of the non-centrality parameter in Equation (3). The non-centrality parameters can be
approximated by Taylor expansion. In the Supplementary material, I show that under some conditions the non-centrality parameters of the non-linear statistics are larger than that of the Hotelling’s $T^2$ statistic. That means that under above assumed conditions the power of the non-linear test statistics is higher than that of Hotelling’s $T^2$ test statistic.

3 RESULTS

3.1 Null distribution of the non-linear test statistics

In the previous sections I have shown that when the sample size is large enough to apply large sample theory, the distribution of the non-linear statistics under the null hypothesis of no differential expressions is asymptotically a central $\chi^2$ distribution. To examine the validity of this statement, I performed a series of simulation studies. Two datasets: the expression profiles of seven genes in invasive lobular and ductal carcinomas of breast in which there are 38 invasive ductal carcinoma (IDC) and 21 invasive lobular carcinoma (ILC) patients (Zhao et al., 2004), and the expression profiles of four genes in 72 lung neuroendocrine tumor samples and 19 normal samples from GEO database at http://www.ncbi.nlm.nih.gov/geo/gds/gds_browse.cgi?gds=619 were used for simulations. The samples in two types of breast cancers, and lung tumor and normal samples were randomly permuted, 100 000 simulations were repeated. In each simulation, the non-linear statistics were calculated. Figure 1A and B plot the histograms of the statistics based on the quadratic and Gaussian transformations applied to breast cancer samples with the theoretical central $\chi^2$ distributions superimposed, respectively. Figure 2A and B plot the histograms of the statistics based on the quadratic and Gaussian transformations applied to lung samples with the theoretical central $\chi^2$ distributions superimposed, respectively. It can be seen that the distributions of

<table>
<thead>
<tr>
<th>Function</th>
<th>Derivative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polynomial $x^2 + x + 1$</td>
<td>$2x + 1$</td>
</tr>
<tr>
<td>Gaussian $e^{-\frac{x}{\sigma^2}}$</td>
<td>$\frac{e^{-x}}{\sigma^2} \times e^{-\frac{x}{\sigma^2}}$</td>
</tr>
</tbody>
</table>
the non-linear statistics are similar to the theoretical central \( \chi^2 \) distributions even under the scenario of modest tissue sample size. Table 2 summarizes the type I error rates of the non-linear test statistics and the Hotelling \( T^2 \) statistic applied to the simulated breast cancer samples and lung samples. It shows that the estimated type I error rates of the non-linear test statistics were not appreciably different from the nominal levels: \( \alpha = 0.05 \) and \( \alpha = 0.01 \). These results demonstrate that the tests based on quadratic and Gaussian transformation of the expression levels are still valid even for the reasonably small sample sizes.

### 3.2 Power of the non-linear test statistics and Hotelling’s test statistic

In Section 2.3 I showed that under some conditions, the power of some non-linear statistics is higher than that of the Hotelling \( T^2 \) statistic using approximation approach. Now the exact analytic methods are used to calculate their power. The power of two non-linear statistics based on quadratic and Gaussian functions and the Hotelling \( T^2 \) statistic with the significance level \( \alpha = 0.001 \) is shown in Figure 3. I assume that the variances \( \sigma_x^2 \) and \( \sigma_y^2 \) of the gene expression in the normal and abnormal tissue samples are equal to 5 and 1 respectively, and the numbers of both normal and abnormal tissue samples are equal to 100. In Figure 3, for ease of presentation, I consider the expression of only one gene. Figure 3 plots the power of the test statistics as a function of difference in gene expressions between normal and abnormal tissues, and demonstrates that the non-linear test statistics in general have higher power than the Hotelling’s \( T^2 \) statistic. The difference in power between the non-linear statistics and the Hotelling’s \( T^2 \) statistic increases as the difference in gene expression levels between the abnormal and normal tissue samples increases.

### 3.3 Real data examples

To further evaluate the performance of the non-linear test statistics two real datasets were used. One dataset was the gene expression profiles of invasive lobular and ductal carcinomas of breast (Zhao et al., 2004). There were 38 IDC and 21 ILC patients with over 42,000 genes profiled using cDNA. Specimens were separated by the modified Scarff-Bloom-Richardson method. The non-linear statistics and the Hotelling’s \( T^2 \) statistic were applied to this dataset for testing differential expressions of two pathways: cell cycle regulation pathway and MKKK pathway between IDC samples and ILC samples. I have taken care that, for better comparison, the number of genes here match that of genes used in calculating type I error rates. Cell cycle regulation pathway includes seven genes: CDC2, CDK7, HK2, KRAS2, PLK1, PRKCA and STAT1, and MKKK pathway includes seven genes: DLK1, MAP3K14, MAP3K3, MAP3K4, MAP3K7, MAP4K3 and MST1. Table 3 lists \( P \)-values of \( T^2 \); quadratic and Gaussian non-linear statistics for testing differential expressions of cell cycle regulation and MKKK pathways. In general non-linear statistics show stronger significance.

Another dataset was for lung neuroendocrine tumor and normal gene profiles. There were 72 various tumor samples and 19 normal samples from GEO database. Table 4 lists \( P \)-values of the Hotelling’s \( T^2 \), quadratic and Gaussian non-linear statistics for testing differential expressions of P53 pathway and Androgen pathway between the lung cancer samples and normal samples.

<table>
<thead>
<tr>
<th>Pathway</th>
<th>( T^2 )</th>
<th>Quadratic</th>
<th>Gaussian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell cycle regulation</td>
<td>0.0175943</td>
<td>5.28406e−005</td>
<td>4.5096e−005</td>
</tr>
<tr>
<td>MKKK</td>
<td>0.000769</td>
<td>4.01396e−006</td>
<td>4.02049−e006</td>
</tr>
<tr>
<td>P53</td>
<td>2.0055e−03</td>
<td>1.2055e−006</td>
<td>6.2354e−007</td>
</tr>
<tr>
<td>Androgen</td>
<td>1.13046e−4</td>
<td>3.9074e−012</td>
<td>3.8891e−013</td>
</tr>
</tbody>
</table>

The \( T^2 \) statistic was applied to this dataset for testing differential expressions of two pathways: cell cycle regulation pathway includes seven genes: CDC2, CDK7, HK2, KRAS2, PLK1, PRKCA and STAT1, and MKKK pathway includes seven genes: DLK1, MAP3K14, MAP3K3, MAP3K4, MAP3K7, MAP4K3 and MST1. Table 3 lists \( P \)-values of \( T^2 \); quadratic and Gaussian non-linear statistics for testing differential expressions of cell cycle regulation and MKKK pathways. In general non-linear statistics show stronger significance.

Another dataset was for lung neuroendocrine tumor and normal gene profiles. There were 72 various tumor samples and 19 normal samples from GEO database. Table 4 lists \( P \)-values of the Hotelling’s \( T^2 \), quadratic and Gaussian non-linear statistics for testing differential expressions of P53 pathway and Androgen pathway.

<table>
<thead>
<tr>
<th>Pathway</th>
<th>( T^2 )</th>
<th>Quadratic</th>
<th>Gaussian</th>
</tr>
</thead>
<tbody>
<tr>
<td>P53</td>
<td>2.0055e−03</td>
<td>1.2055e−006</td>
<td>6.2354e−007</td>
</tr>
<tr>
<td>Androgen</td>
<td>1.13046e−4</td>
<td>3.9074e−012</td>
<td>3.8891e−013</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pathway</th>
<th>( T^2 )</th>
<th>Quadratic</th>
<th>Gaussian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell cycle regulation</td>
<td>0.0175943</td>
<td>5.28406e−005</td>
<td>4.5096e−005</td>
</tr>
<tr>
<td>MKKK</td>
<td>0.000769</td>
<td>4.01396e−006</td>
<td>4.02049−e006</td>
</tr>
<tr>
<td>P53</td>
<td>2.0055e−03</td>
<td>1.2055e−006</td>
<td>6.2354e−007</td>
</tr>
<tr>
<td>Androgen</td>
<td>1.13046e−4</td>
<td>3.9074e−012</td>
<td>3.8891e−013</td>
</tr>
</tbody>
</table>

The \( T^2 \) statistic was applied to this dataset for testing differential expressions of two pathways: cell cycle regulation pathway includes seven genes: CDC2, CDK7, HK2, KRAS2, PLK1, PRKCA and STAT1, and MKKK pathway includes seven genes: DLK1, MAP3K14, MAP3K3, MAP3K4, MAP3K7, MAP4K3 and MST1. Table 3 lists \( P \)-values of \( T^2 \); quadratic and Gaussian non-linear statistics for testing differential expressions of cell cycle regulation and MKKK pathways. In general non-linear statistics show stronger significance.

Another dataset was for lung neuroendocrine tumor and normal gene profiles. There were 72 various tumor samples and 19 normal samples from GEO database. Table 4 lists \( P \)-values of the Hotelling’s \( T^2 \), quadratic and Gaussian non-linear statistics for testing differential expressions of P53 pathway and Androgen pathway.

<table>
<thead>
<tr>
<th>Pathway</th>
<th>( T^2 )</th>
<th>Quadratic</th>
<th>Gaussian</th>
</tr>
</thead>
<tbody>
<tr>
<td>P53</td>
<td>2.0055e−03</td>
<td>1.2055e−006</td>
<td>6.2354e−007</td>
</tr>
<tr>
<td>Androgen</td>
<td>1.13046e−4</td>
<td>3.9074e−012</td>
<td>3.8891e−013</td>
</tr>
</tbody>
</table>
for calculating type I error. The results show that \( P \)-values of non-linear statistics were smaller than that of the Hotelling’s \( T^2 \).

4 DISCUSSION

A key issue in gene expression data analysis is to identify differentially expressed genetic networks. To improve the power of the Hotelling \( T^2 \) statistic for testing differential expressions of genetic networks I have developed a general statistical framework for non-linear tests, and have provided basic procedures on how to construct test statistics using non-linear transformations of means of gene expressions; I have presented two non-linear test statistics for testing differential expressions of genetic networks between two types of tissue samples.

In this report, I have derived asymptotical distributions of the non-linear test statistics under null and alternative hypotheses. To reveal the relationship between the power of linear test statistics and non-linear statistics, I have approximated the non-centrality parameter of the non-linear test statistics and showed how it depends on the measure of non-linearity of functions. I have also demonstrated that under some conditions the power of non-linear test statistics is larger than that of Hotelling \( T^2 \) statistic. The power of test statistics is a complicated issue; it depends on many parameters such as difference in population means of gene expressions using two types of tissue samples, the variance of gene expressions, the number of tissue samples and the measure of non-linearity of non-linear functions; so it is difficult to find statistics that are uniformly most powerful. To further evaluate performance of the non-linear test statistics, the proposed non-linear test statistics are applied to two real datasets.

The differential expression of genetic networks is the property of the networks as whole, owing to perhaps differential expressions of some individual genes in the network, or other factors like gene–gene interaction. This report shows that non-linear transformations provide amplified power and can more conclusively demonstrate differentiation of tumor and normal tissues, all without high rate of false positive, and thus its superiority. Because of its enhanced power, cases that might have been missed would emerge so they can be investigated further. It is a better one-step measure for testing genetic networks for differentiation.

The results in this report are thus far limited. Theoretical and empirical studies should be conducted to compare and investigate the relative strength and weakness of non-linear statistics and other existing statistics for identifying differentially expressed genetic networks. This report only presents two non-linear statistics; it is worthwhile to investigate other non-linear statistics and develop general procedures for searching optimal non-linear statistics with the highest power. Non-linear tests are powerful tools, particularly for identifying differentially expressed genetic networks. However theory for non-linear tests has not been fully developed and non-linear statistics have not been applied to large datasets. Considerable theoretic work and empirical evaluation for non-linear tests for gene expression data analysis are urgently needed.

Conflict of Interest: none declared.

REFERENCES