**Gene expression**

**Multivariate Analysis of Variance Test for Gene Set Analysis**

Chen-An Tsai¹ and James J. Chen²*  
¹ Department of Public Health & Biostatistics Center, China Medical University, Taichung, Taiwan, ² Division of Personalized Nutrition and Medicine, National Center for Toxicological Research, FDA, Jefferson, AR 72079, USA  
Associate Editor: Prof. John Quackenbush

**ABSTRACT**

**Motivation:** Gene class testing (GCT) or gene set analysis (GSA) is a statistical approach to determine whether some functionally predefined sets of genes express differently under different experimental conditions. Shortcomings of the Fisher’s exact test for the over-representation analysis are illustrated by an example. Most alternative GSA methods are developed for data collected from two experimental conditions, and most is based on a univariate gene-by-gene test statistic or assume independence among genes in the gene set. A multivariate analysis of variance (MANOVA) approach is proposed for studies with two or more experimental conditions.

**Results:** When the number of genes in the gene set is greater than the number of samples, the sample covariance matrix is singular and ill-condition. The use of standard multivariate methods can result in biases in the analysis. The proposed MANOVA test uses a shrinkage covariance matrix estimator for the sample covariance matrix. The MANOVA test and six other GSA published methods, PCA, SAM-GS, ANCOVA, Global, GSEA, and MaxMean, are evaluated using simulation. The MANOVA test appears to perform the best in terms of control of Type I error and power under the models considered in the simulation. Several publicly available microarray datasets under two and three experimental conditions are analyzed for illustrations of GSA. Most methods, except for GSEA and MaxMean, generally are comparable in terms of power of identification of significant gene sets.

**Availability:** A free R-code to perform MANOVA test is available at [http://mail.cmu.edu.tw/~catsai/research.htm](http://mail.cmu.edu.tw/~catsai/research.htm)

**Contact:** JamesJ.Chen@fda.hhs.gov; catsai@mail.cmu.edu.tw

**1 INTRODUCTION**

Biological phenomena often occur through the interactions of multiple genes, via signaling pathways, networks, or other functional relationships. In microarray data analysis, after selecting the list of significant genes, investigators are often interested in knowing whether some functionally predefined classes of genes are differentially expressed. A gene class refers to a group of genes with related functions or a set of genes grouped together based on biologically relevant information, such as a metabolic pathway, protein complex, or GO (gene ontology) category. A statistical test to determine significance of a gene class is referred to as gene class testing (GCT) or gene set analysis (GSA). The common approach to the GSA is first to identify a list of genes that express differently among two groups of samples. The list of differentially expressed genes is then examined with biologically pre-defined gene sets to determine whether any set is overrepresented in the list compared with the whole list (e.g., Draghici et al., 2003; Khatri and Draghici, 2005; Rivals et al., 2007). This approach is known as the over-representation analysis (ORA).

The Fisher’s exact test is typically used to assess the significance for an over-representation (Draghici et al., 2003). The p-value for significance of gene set from the Fisher’s exact test is calculated as

\[ p(x) = \sum_{l=x}^{K} \binom{N}{l} \binom{M-N}{K-l}, \]

where M is the total number of genes in the array, N is the number of genes in the class, K is the number of genes in the significant list, and x is the number of genes in the list from the class. There are a number of shortcomings with the ORA approach (e.g., Pavlidis et al., 2004; Tian et al., 2005; Subramanian, A. et al., 2005). First, the division of genes into differential and non-differential expression groups is arbitrary, and the genes in the non-differential expression list are discarded, regardless of their p-values. Second, the Fisher’s exact test assumes the genes are independent, and it does not take the correlation structure in the gene class into consideration. Third, the approach simply counts the number of genes in the list; the order of genes is not taken into consideration. These shortcomings are highlighted by an analysis of a gene set from the diabetes dataset presented by Mootha et al. (2003) given below.

The diabetes dataset consists of 149 gene classes from 10,526 genes measured on 17 subjects with normal glucose tolerance and 18 subjects with Type 2 diabetes mellitus. Of the 149 gene classes analyzed, the gene class OXPHOS (oxidative phosphorylation) was one of the most significance gene classes identified (e.g., Mootha et al., 2003; Tian et al., 2005; Chen et al., 2007). Table 1 shows the p-values from the Fisher’s exact test according with the significant cutoff probability. It can be seen that the p-value fluctuates as the cutoff changes.

Recently, Subramanian et al. (2005) and Mootha et al. (2003) proposed the Gene Set Enrichment Analysis (GSEA) by considering the entire distribution of genes in the class, rather than individual genes. The GSEA approach is able to identify a significant gene set between the diabetic samples and normal muscles for which no single gene was found to be differentially expressed using the ORA approach. The work of Mootha et al. (2003) has inspired the development of various GSA methods for alternatives to the ORA approach. Tian et al. (2005) and Chen et al. (2007) proposed test statistics based on the two-sample t-statistics. Dinu et al. (2007) proposed a SAM-GS test based on the SAM statistic (Tusher et al., 2001) and Adewale et al. (2008) generalized the SAM-GS statistic from the framework of regression model. Their test for a gene set are an aggregate of each individual gene test statistics within the gene set. The SAM-GS test is the only GSA method to address the small variability present in microarray gene.

---

¹To whom correspondence should be addressed.
expression studies by incorporating the SAM constant [Tusher et al., 2001] into the test statistic. The small variability can lead to inflation of the t-test statistic due to very small denominator, and therefore genes whose average expressions corresponding to the two groups are extremely close can be identified as significant. Efron and Tibshirani (2007) proposed a MaxMean statistic for summarizing gene-sets, and a re-standardization for more accurate inferences. These approaches utilize an aggregate of individual statistics based on univariate per gene analysis.

Alternatively, Tomfohr et al. (2005) used a t-statistic but after reducing the dimension of the gene expression data matrix to its first principal component. Kong et al. (2006) proposed using Hotelling’s $T^2$ statistic with a similar dimensional reduction approach. Goeman et al. (2004) proposed a global score test by modeling gene expressions as random effects in a logistic regression model. Mansmann and Meister (2005) and Hummel et al. (2008) proposed an ANCOVA (analysis of covariance) test, which is similar to the Goeman et al. (2004) model except that the roles of condition and gene are exchanged in the regression models. A third approach is a meta-analysis based the individual p-values from the univariate test (e.g., Pavlidis et al., 2004). Recently, Goeman and Mansmann (2008) proposed a focus-level method for GSA analysis of GO terms. It made use of the hierarchical structure of GO graphs. They proposed using the closed testing procedure (Marcus et al., 1976) to account for multiple testing in the GSA analysis.

Table 1. An over-representation analysis of the diabetes data from the Fisher’s exact test. The gene OXPHOS (oxidative phosphorylation) with the class size of (M = 206) is one of the most significance gene classes identified. For a given cutoff, K is the number of genes in the significant list, and X is the number of gene in the list from the class. The p-value fluctuates as the cutoff changes.

<table>
<thead>
<tr>
<th>Cutoff</th>
<th>K</th>
<th>X</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>.0025</td>
<td>36</td>
<td>1</td>
<td>.2951</td>
</tr>
<tr>
<td>.0030</td>
<td>42</td>
<td>1</td>
<td>.3351↑</td>
</tr>
<tr>
<td>.0050</td>
<td>67</td>
<td>1</td>
<td>.4789↑</td>
</tr>
<tr>
<td>.0060</td>
<td>81</td>
<td>2</td>
<td>.1840↓</td>
</tr>
<tr>
<td>.0100</td>
<td>124</td>
<td>2</td>
<td>.3369↑</td>
</tr>
<tr>
<td>.0150</td>
<td>187</td>
<td>2</td>
<td>.4582↑</td>
</tr>
<tr>
<td>.0200</td>
<td>248</td>
<td>5</td>
<td>.0918↓</td>
</tr>
<tr>
<td>.0210</td>
<td>262</td>
<td>8</td>
<td>.0041↓</td>
</tr>
<tr>
<td>.0300</td>
<td>363</td>
<td>9</td>
<td>.0083↑</td>
</tr>
<tr>
<td>.0400</td>
<td>463</td>
<td>9</td>
<td>.0347↑</td>
</tr>
<tr>
<td>.0500</td>
<td>573</td>
<td>11</td>
<td>.0221↓</td>
</tr>
</tbody>
</table>

Liu et al. (2007) compared statistical performance of the global test (Goeman et al., 2004), ANCOVA test (Mansmann and Meister, 2005), and SAM-GS (Dinu et al., 2007). In the simulation experiment, they found that the asymptotic distributions of the global and ANCOVA tests can lead to an incorrect Type I error (size). A proper standardization across genes is necessary for the global and ANCOVA tests in order to obtain more accurate inference. After a proper standardization, the two tests with permutation-based inference gave a correct size. SAM-GS showed slightly higher power than the global and ANCOVA tests. In the analysis of three microarray datasets, the three methods showed similar power using permutation inference.

Tian et al. (2005) presented two fundamental hypotheses for GSA. The first hypothesis (Q1) tests the same level of association of a gene set with the given phenotype as the complement of the gene set. The second hypothesis (Q2) tests if there is no gene in the gene set associated with the phenotype. The null distribution of the statistic for Q1 was generated by permuting genes (gene sampling), and the null distribution for Q2 was generated by permuting samples (subject-sampling). Goeman and Bühlmann (2007) provided detailed discussions of the Q1 versus Q2 hypothesis and the generation of their respective null distributions via gene-sampling versus subject sampling. Nam and Kim (2008) provided a list of GSA methods with their hypotheses (either Q1 or Q2 or both) and the sampling methods, and the references.

Almost all GSA methods considered studies only for two experimental conditions, although some approaches can be generalized to more than two conditions. Many microarray experiments involve multiple experimental conditions. The different experimental conditions can be dose levels, time points, or treatment combinations. The Goeman et al. (2004) logistic model for two conditions recently has been extended to multinomial model for multiple conditions (http://www.bioconductor.org/) in R package. The ANCOVA test (Mansmann and Meister, 2005) can be directly used to analyze data from three or more conditions.

This paper proposes a modified multivariate analysis of variance (MANOVA) test for GSA for data collected from studies with two or more experimental conditions. When there are two conditions, the MANOVA test becomes the Hotelling $T^2$ test. Hotelling’s $T^2$ has already been applied to identify differentially expressed genes (Szabo et al., 2003; Kim et al., 2005; Lu et al., 2005). One commonly known challenge of the use of the MANOVA or Hotelling’s $T^2$ test for GSA is that it requires the number of samples to be larger than the number of genes in the gene set in order to avoid singularity in the inversion of the sample covariance matrix. An intuitive approach to account for the singularity is to use the generalized inverse matrix to compute the test statistic. However, this approach does not perform well (data not shown). A less known problem is that the sample covariance matrix is not a good estimate of the true population covariance matrix when the number of genes is more than the number of samples (Schafer and Strimmer, 2005). Simply using the sample covariance matrix can lead to poor performance in the subsequent analysis. This paper applies the shrinkage covariance matrix estimator (Schafer and Strimmer, 2005) to compute the MANOVA test statistic. The analysis focuses on the Q2 hypothesis since it is consistent with the conventional approach of identifying differentially expressed genes. The p-values are computed by permuting the samples that are obtained independently by the experimental design. The p-values are comparable across the gene sets and can be used for gene ranking as well as significance assessment.

2 METHOD

Consider a microarray study of m genes with c experimental conditions (phenotypes) of sample sizes $n_1, \ldots, n_c$. Without loss of gen-
erality, consider a gene set consisting of m genes. Let \( y_{ij} \) be the m-vector of intensities for simple \( j = 1,...,n_i \) in condition \( i = 1,...,c \). The MANOVA model (Johnson and Wichern, 2000) can be expressed as \( y_{ij} = \mu_i + \epsilon_{ij} \), where \( \epsilon_{ij} \) is m-vector of residuals with \( \text{Var}(\epsilon_{ij}) = \Sigma \) and \( \mu_i \) is the m-vector of means for the \( i \)-th condition. The null hypothesis of no difference in gene expressions among the \( c \) conditions is given as: \( \mu_1 = \mu_2 = ... = \mu_c \). The alternative is at least one gene express differently in at least two conditions. There are four MANOVA tests: Wilks’ \( \Lambda \), Pillai’s tract, Hotelling’s \( T^2 \), and Roy’s largest root. The four tests are equivalent to Hotelling’s \( T^2 \) when there are only two conditions. The Wilks’ \( \Lambda \) is used in this paper given as

\[
\Lambda = \prod_{i=1}^{K} \frac{1}{1 + \lambda_i}
\]

Where \( \lambda_i \) is the \( k \)-th eigenvalue of the matrix \( S (= E^{-1}H) \), and \( E \) is the within sum of squares matrix. The number of eigenvalues \( K \) is equal to the minimum of the number of genes \( (m) \) and the number of conditions minus one \( (c-1) \). When the number of genes in the gene set is greater than the number of samples, the matrix \( E \) is singular and ill-condition. The shrinkage covariance matrix estimator \( (S_{ij}) \) proposed by Schäfer and Strimmer (2005) is used to make the matrix well-condition, and is given as

\[
S_{ij}^* = \begin{cases} 
    \frac{s_{ij}}{r_{ij}} & \text{if } i = j \\
    \left( \frac{\sqrt{s_{ij}}}{\sqrt{r_{ij}}} \right) & \text{if } i \neq j
\end{cases}
\]

\[
r_{ij}^* = n_i \min \{1, \max(0,1 - \hat{\lambda}^2)\},
\]

where \( s_{ij} \) and \( r_{ij} \) respectively denote the empirical sample variance and sample correlation, and the optimal shrinkage intensity \( \hat{\lambda} \) is estimated by

\[
\hat{\lambda}^2 = \frac{\sum_{i,j} \hat{\text{Var}}(r_{ij})}{\sum_{i,j} r_{ij}^2}.
\]

The null distribution of ‘standard’ Wilks’ \( \Lambda \) (or Hotelling’s \( T^2 \)), based on the sample covariance matrix, can be approximated by a \( F \) distribution for some special cases (Johnson and Wichern, 2000). For the ‘modified’ Wilks’ \( \lambda \) using the shrinkage estimator, its distribution does not have a close form. Using a similar approach to the commonly used GSA methods, the p-values of the modified Wilk’s \( \lambda \) are computed using the permutation method.

3 SIMULATION STUDY

The proposed MANOVA approach is evaluated and compared with six published GSA methods using the Monte Carlo simulation for two experimental conditions. These six methods are: principal component analysis (PCA) by Kong et al. (2006), SAM-GS by Dinu et al. (2007), ANCOVA by Mansmann and Meister (2005), Global by Goeman et al. (2004), GSEA by Subramanian et al. (2005), and MaxMean by Efron and Tibshirani (2007). A brief description of each method is given in Supplemental 1. The simulation design was similar to that considered by Liu et al. (2007). It considered 100 genes \((m=100)\) in a gene set. In the null model, the 100 genes were simulated from a multivariate normal (MVN) distribution with a mean vector \( \mu \) and a diagonal variance-covariance matrix \( \Sigma \). The 100 elements of \( \mu \) were generated from the uniform random variables \( U[0,10] \), and the 100 diagonal elements of \( \Sigma \) were generated from the uniform random variable \( U[0.1,10] \). The first 20 genes had a common correlation coefficient of 0.5, as were the next 20 genes. That is, \( \Sigma \) had compound symmetry structures in the first 20 genes and the next 20 genes. The correlation \( \rho \) was set at 0 (to include the independent model), 0.3, 0.5, and 0.9. The alternative model had the same variance-covariance structure. But, the first 20 genes had a mean difference of \( 2\gamma \), and the next 20 genes had the mean difference of \(-2\gamma \). The mean difference \( \gamma \) was set at 0.2, 0.4, 0.6, 0.8, 1.0, and 1.2. The numbers of samples in each condition considered were 10 and 25. The simulation data were replicated 1,000 times in each model. The p-values were based on 1,000 permutations.

The R-code performing SAM-GS was downloaded from http://www.ualberta.ca/~yyasui/homepage.html, as well as R-packages, GlobalAncova and Globaltest, to respectively implement ANCOVA and Global tests were available from http://www.bioconductor.org. The R-code for GSEA was downloaded from http://www.broad.mit.edu/gsea, and MaxMean was from the GSA package in R. The MANOVA and PCA were implemented using the R statistical software. It should be noted that GSEA and MaxMean are a one-sided test to identify either up- or down regulated gene sets; they are considered in the evaluation since these two methods have been used widely.

Table 2 shows the empirical type I errors using the nominal level of 0.05. The Type I errors from Hotelling’s \( T^2 \) and ANCOVA are reasonably close to or below the nominal level. Both tests show slight conservatism for \( n_i = 10 \) and \( \rho = 0.3 \). PCA, SAM-GS, and GSEA show anti-conservatism in few cases. Global appears to be very conservative. MaxMean appears to have an overly inflated size. Figure 1 illustrates the empirical powers using the nominal level of 0.05 for \( \gamma = 0.2, 0.4, 0.6, 0.8, 1.0 \), and 1.2. It can be seen that Hotelling’s \( T^2 \) performs the best among the seven tests in all cases. The one-sided GSEA and MaxMean tests have lower powers since the alternative models consist of mixture of up-and down-regulated genes in the gene sets. For the two-sided test, the Global test has the lowest power, consistent with the results in Table 2. SAM-GS and PCA appear to be comparable. More discussion on the simulation results, the one-sided versus two-sided tests, and Q1 and Q2 hypotheses are given in Discussion.

<table>
<thead>
<tr>
<th>Method</th>
<th>( \rho = 0 )</th>
<th>( \rho = 0.3 )</th>
<th>( \rho = 0.5 )</th>
<th>( \rho = 0.9 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hotelling’s ( T^2 )</td>
<td>0.050</td>
<td>0.039</td>
<td>0.038</td>
<td>0.050</td>
</tr>
<tr>
<td>PCA</td>
<td>0.053</td>
<td>0.042</td>
<td>0.052</td>
<td>0.062</td>
</tr>
<tr>
<td>SAM-GS</td>
<td>0.046</td>
<td>0.042</td>
<td>0.038</td>
<td>0.055</td>
</tr>
<tr>
<td>ANCOVA</td>
<td>0.042</td>
<td>0.038</td>
<td>0.034</td>
<td>0.052</td>
</tr>
<tr>
<td>Global</td>
<td>0.001</td>
<td>0.009</td>
<td>0.016</td>
<td>0.034</td>
</tr>
<tr>
<td>GSEA</td>
<td>0.059</td>
<td>0.058</td>
<td>0.052</td>
<td>0.048</td>
</tr>
<tr>
<td>MaxMean</td>
<td>0.093</td>
<td>0.094</td>
<td>0.107</td>
<td>0.098</td>
</tr>
</tbody>
</table>

MANOVA test to gene set analysis
Table 2 (Continued)

<table>
<thead>
<tr>
<th>Method</th>
<th>$n_1 = n_2 = 25$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\rho = 0$</td>
</tr>
<tr>
<td>Hotelling’s $T^2$</td>
<td>0.046</td>
</tr>
<tr>
<td>PCA</td>
<td>0.052</td>
</tr>
<tr>
<td>SAM-GS</td>
<td>0.041</td>
</tr>
<tr>
<td>ANCOVA</td>
<td>0.050</td>
</tr>
<tr>
<td>Global</td>
<td>0.005</td>
</tr>
<tr>
<td>GSEA</td>
<td>0.054</td>
</tr>
<tr>
<td>MaxMean</td>
<td>0.103</td>
</tr>
</tbody>
</table>

Fig. 1. Power analysis of five GSA methods: Hotelling’s $T^2$, PCA, SAM-GS, ANCOVA, Global, SEA, and MaxMean tests.

4. APPLICATIONS

4.1 Two experimental conditions

We applied the proposed MANOVA and the six other GSA methods to three microarray studies: Gender, Leukemia, and P53. The three studies are frequently used for GSA illustrations (e.g., Subramanian et al., 2005; Dinu et al., 2007). The three studies are publicly available at the GSEA website (http://www.broad.mit.edu/gsea), each study consists of two catalogs of gene sets, chromosome and cytogenetic catalog (C1) and functional catalog (C2). The gender dataset includes 15,056 mRNA expression profiles from 15 male and 17 female samples of lymphoblastoid cell lines. The leukemia dataset is used to study lymphoid leukemia (ALL) and acute myeloid leukemia (AML) by comparing 10,056 expression profiles derived from 24 ALL patients and 24 AML patients. The p53 dataset is a study to identify targets of the transcription factor p53 from 10,100 gene expression profiles in the NCI-60 collection of cancer cell lines. The mutation status of the p53 gene has been reported for 50 of the NCI-60 cell lines with 17 normal and 33 mutation samples. Dinu et al. (2007) analyzed the three datasets Gender (C1), Leukemia (C1), and P53 (C2) in SAM-GS and GSEA comparisons. The seven GSA methods were applied to the four datasets: Gender (C1), Gender (C2), Leukemia (C1), and P53 (C2). The numbers of gene sets are 212, 308, 182, and 308, respectively.

Table 3 shows the number of gene sets with p-values less than 0.01 and 0.05 from the seven GSA methods. The p-values were used instead of the false discovery rate (FDR) because p-values are more informative for the comparison purpose. For example, for the Leukemia study the number of gene sets with an FDR $\leq 0.25$, which was used by Dinu et al. (2007), is 182 in all five two-sided tests (the number is 3 for GSEA and 0 for MaxMean). For the two-sided methods, there are some discrepancies between the simulation result and the data analysis. It seems that MANOVA, PCA and SAM-GS perform better on the simulation and on the P53 and Leukemia datasets, whereas Global and ANCOVA perform better on the two Gender datasets. For the one-sided methods, MaxMean identify slightly more gene sets than GSEA. However, MaxMean is shown to have an inflated size in the simulation analysis. The p-values and q-values (FDR) of each gene set of the four studies for the seven GSA methods are provided in Supplementary 2.

Table 3. The number of gene sets with p-values less than 0.01 and 0.05 from the Hotelling’s $T^2$, PCA, SAM-GS, ANCOVA, Global, GSEA, and MaxMean tests for the four datasets, gender (C1), gender (C2), leukemia (C1), and p53 (C2), the numbers of gene sets are 212, 308, 182, and 308, respectively.

<table>
<thead>
<tr>
<th>Methods</th>
<th>p-value ≤ 0.01</th>
<th>gender (C1)</th>
<th>gender (C2)</th>
<th>Leukemia (C1)</th>
<th>Leukemia (C2)</th>
<th>P53 (C1)</th>
<th>P53 (C2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hotelling’s $T^2$</td>
<td>0.01</td>
<td>5</td>
<td>5</td>
<td>182</td>
<td>40</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>7</td>
<td>9</td>
<td>182</td>
<td>78</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCA</td>
<td>0.01</td>
<td>3</td>
<td>1</td>
<td>180</td>
<td>47</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>3</td>
<td>1</td>
<td>181</td>
<td>78</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAM-GS</td>
<td>0.01</td>
<td>4</td>
<td>2</td>
<td>182</td>
<td>30</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>4</td>
<td>5</td>
<td>182</td>
<td>49</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANCOVA</td>
<td>0.01</td>
<td>5</td>
<td>7</td>
<td>171</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>17</td>
<td>26</td>
<td>176</td>
<td>16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Global</td>
<td>0.01</td>
<td>5</td>
<td>7</td>
<td>171</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>22</td>
<td>28</td>
<td>175</td>
<td>18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSEA</td>
<td>0.01</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>6</td>
<td>8</td>
<td>8</td>
<td>26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MaxMean</td>
<td>0.01</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>17</td>
<td>27</td>
<td>15</td>
<td>33</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The two-sided methods, in particular $T^2$, generally identify more gene sets than the one-sided GSEA test. The biggest discrepancy between the two-sided and one-sided tests is the leukemia data set in which the two-sided methods identify
MANOVA test to gene set analysis

171 (the least) out of the 182 gene sets whereas the one-sided methods identify 15 (the most). Dinu et al (2007) discussed large discrepancies between the GSEA and SAM-GS methods in their analysis of the p53 and leukemia datasets. In the p53-comparison analysis, SAM-GS identified more gene sets than GSEA, and many of those additional gene sets identified by SAM-GS were associated with p53 genes or p53 signaling. In the leukemia dataset, Dinu et al. (2007) pointed out to a high percentage of individual genes were significant, and commented that a GSA analysis should be consistent with the individual gene analysis. In the present analysis, all methods, except for GSEA and MaxMean, identified more than 90% of genes sets and are consistent with the individual gene analysis.

Table 4 illustrates the similarities and discrepancies among the five two-sided and two one-sided methods for some gene sets observed. These are selected from the top 10 ranked gene sets from each of the seven GSA methods. ANCOVA and Global appear to be very similar, as are the two one-side methods. The complete list of the 10 top ranked gene sets is given in Supplementary 3.

4.2 More than two experimental conditions

The MANOVA approach was also applied to a breast cancer dataset (van’t Veer et al., 2002) to illustrate a GSA analysis with more than two experimental conditions. This dataset was a study of gene expression profiling to predict the clinical outcome of breast cancer. A subset of normalized gene expression data consisting of 1113 genes from 96 samples without BRCA1 or BRCA2 mutations was mapped to nine cancer related pathways. The experimental conditions analyzed were three ordered levels of tumor grade, 1, 2, and 3, with the sample sizes of 11, 25, and 60, respectively. Table 5 shows the p-values of the nine pathways using the MANOVA and ANCOVA. MANOVA and ANCOVA appear to be comparable for gene sets with very small p-values. The MANOVA and ANCOVA, respectively, have the p-values of 0.0025 and 0.0545 for the Notch delta signaling pathway, and p-values of 0.0535 and 0.1165 for Tgf beta signaling pathway.

Table 5: Results of the GSA analysis of breast cancer data for 9 pathways using MANOVA and ANCOVA.

5 DISCUSSION

Genes in a gene set are functionally related and are not independent; the complex structure of gene interactions within a gene set are not fully captured using univariate approaches. The methods such as the Fisher's exact test or the ANCOVA asymptotically parametric test which calculated the p-values under the assumption of independence between genes will have incorrect Type I error if genes are in fact correlated (Goeman and Buhlmann, 2007; Liu et al., 2007). Most GSA test methods, such as SAM-GS, Global, ANCOVA, and GSEA, do not involve an estimate of the covari-
Multivariate methods, such as Tomfohr et al. (2005) and Kong et al. (2006), take the variance or correlation matrix into consideration in the analysis. The covariance matrix should be positive definite. But when the number of genes is greater than the number of samples, the sample covariance matrix is singular (non-positive definite). Furthermore, Schafer and Strimmer (2005) showed that the sample covariance matrix is ill-condition and its eigenvalues can be very different from the true eigenvalues. The use of the sample covariance matrix in a multivariate method for the analysis of data with high dimensional variables and low dimensional sample sizes is inefficient and can result in biases in the analysis (Table 2). Both Tomfohr et al. (2005) and Kong et al. (2006) applied dimensional reduction techniques to find the orthogonal linear combinations of variables that showed the greatest variability of the data. The dimensional reduction techniques define a smaller number of hybrid components that are a composite of the original variables. Since the underlying dimensionality is not known, these techniques must search for the number of components that can account for much of data variation based on some ‘subjective cut-off criterions’. The Tomfohr et al. (2005) approach simply used the largest principal component, based on the gene expression data matrix instead of the sample covariance matrix. This approach can be effective if there is a single dimensional space that can explain much of data variation sufficiently. The analysis of searching for the component that optimizes differences between conditions is considered to be a post hoc analysis in the sense that it is performed after a significance difference in the whole gene set is observed.

The MANOVA (Hotelling’s $T^2$) is a multivariate generalization of the univariate analysis of variance (or t-test). The ANOVA and t-test are used to identify differentially expressed genes, while the MANOVA and Hotelling’s $T^2$ are used to identify differentially expressed gene sets. The univariate ANOVA or t-test compares the mean differences in expression among the phenotypes for an individual gene; MANOVA or $T^2$ compares the mean differences for a set of genes simultaneously. The null hypothesis is rejected if one or more of the mean differences or some combination of mean differences among the genes in the gene set differs from zero. The parametric MANOVA (and $T^2$) test has been used for analyzing multivariate data in many applications. The test statistics were well studied and robust to the skew distribution. The proposed modified MANOVA test uses a shrinkage covariance matrix estimator (Schafer and Strimmer, 2005) to incorporate the correlations structure among the genes in the test statistic and to account for the singularity and ill-condition of the sample covariance matrix. The modified MANOVA test uses the permutation method to compute the p-values. The proposed MANOVA (or $T^2$) test to identify significant gene sets is a generalization of the univariate test such as SAM, which uses an alternative variance estimate to identify significant genes with the permutation method. Furthermore, after appropriate filtering and normalization, gene-expression data appear rather symmetrical (e.g. Irizarry, et al., 2003; Mansmann and Meister, 2005), gene expression data appear rather symmetrical. The proposed $T^2$ test appears to perform well in terms of controlling the Type I error and power as compared to the six other existing methods in simulation and data analysis. Since the comparison between two conditions is the simplest study, similar results could be expected for cases involving multiple conditions.

The proposed MANOVA test is designed to identify genes sets consisting of differentially expressed genes for two or more experimental conditions. When there are more than two conditions, the changes of gene expressions between any two conditions can be up, down, or both. In the context of two experimental conditions, the original GSEA statistic (Mootha et al., 2003) was a one-sided test to identify gene sets containing down-regulated genes in Type 2 diabetes mellitus subjects. The basic idea in this analysis is that the gene sets are closely related and, hence, will have similar expression patterns, either up or down. In GCT, the one-sided test means that the changes of gene expressions in the gene set are in one direction: either up- or down-regulation. Chen et al. (2007) recommend the ordinary least squares (OLS) statistic (O’Brien, 1984) or the standardized weighted sum statistic (LWS) (Laüter, 1996) for a one-sided test. These two global statistics used the similar concept of re-standardization of MaxMean (Efron and Tibshirani, 2007) to account for random selection of genes. These three tests can identify either up- or down-regulated gene sets, while the GSEA test performs up- and down-regulation separately. If the goal is to detect coordinated changes in one direction, then the one-sided hypothesis is appropriate. The one-sided tests can be very powerful when the changes are in the same direction, such as the pathway cell cycle arrest in the Gender C2 data and Nfl-pathway in the P53 data set.

This paper mainly considers the two-sided GSA tests, $T^2$, PCA, SAM-SG, ANVOVA, and Global. The two-sided test means that changes of gene expressions in the gene set can be both up- and down-regulation (Chen et al., 2007). In the exploratory context, it may not be possible to pre-specify how individual genes in a gene set will respond between two conditions. Nevertheless, whether to use a one-sided or two-sided test should be guided by biology.

The simulation models were based on the block compound symmetry (CS) correlation with heterogeneous variance structure. The block CS is the simplest correlation structure to model gene expression data. It provides a basic assessment of the behaviors and performances of GSA methods. Table 2 suggests that MaxMean does not seem to have correct size using the codes downloaded from the GSA package in R. The GSEA and Max-mean methods compare the given gene-set score not only to scores from permutations of the samples (Q2), but also take into account scores from sets formed by random selections of genes (Q1) (Efron and Tibshirani, 2007). Figure 1 shows that the power decreases as the correlation increases in all GSA methods. As described, the simulation models was generated containing both up- and down-regulated genes, therefore, two-sided tests are more powerful than
the one-sided tests. Among the two-sided tests, the ANCOVA and Global have the lower power than $T^2$ PCA, and SAM-GS. Since ANCOVA is based on the model of an equal variance across all genes, additional simulations with an equal variance model of $\sigma^2 = 5$ for all genes were also evaluated. Under the equal variance model, SAM-GS and ANCOVA perform the best; it follows by PCA. $T^2$ and Global have the lowest power. These results are provided in Supplementary 4.

In summary, the proposed modified MANOVA (and $T^2$) statistic takes into account some aspects of the complex multivariate correlations of genes by incorporating a shrinkage sample covariance matrix in the test statistics. The proposed MANOVA or $T^2$ test is useful for identifying differentially expressed gene sets that contain both up- and down-regulated genes under the Q2 hypothesis.

Conflict of Interest: none declared.

ACKNOWLEDGEMENTS

The views presented in this paper are those of the authors and do not necessarily represent those of the U.S. Food and Drug Administration. This work was supported partially by the National Science Council of Taiwan, R. O. C. under the grants NSC 97-2118-M-039-003 and NSC 97-3112-B-001-020 in the National Research Program for Genomic Medicine. The authors would like to thank a reviewer raising the issues of independence among genes in the GSA test statistics and three other reviewers provide valuable comments and suggestions that improved this paper considerably.

REFERENCES


MANOVA test to gene set analysis

The one-sided tests. Among the two-sided tests, the ANCOVA and Global have the lower power than $T^2$ PCA, and SAM-GS. Since ANCOVA is based on the model of an equal variance across all genes, additional simulations with an equal variance model of $\sigma^2 = 5$ for all genes were also evaluated. Under the equal variance model, SAM-GS and ANCOVA perform the best; it follows by PCA. $T^2$ and Global have the lowest power. These results are provided in Supplementary 4.

In summary, the proposed modified MANOVA (and $T^2$) statistic takes into account some aspects of the complex multivariate correlations of genes by incorporating a shrinkage sample covariance matrix in the test statistics. The proposed MANOVA or $T^2$ test is useful for identifying differentially expressed gene sets that contain both up- and down-regulated genes under the Q2 hypothesis.

Conflict of Interest: none declared.