# The Shrinkage Variance Hotelling $T^2$ Test for Genomic Profiling Studies

Grant Izmirlian, Jian-Lun Xu

June 8, 2005

**Abstract**

Designed gene expression micro-array experiments, consisting of several treatment levels with a number of replicates per level, are analyzed by applying simple tests for group differences at the per gene level. The gene level statistics are sorted and a criterion for selecting important genes which takes into account multiplicity is applied. A caveat arises in that true signals (genes truly over or under expressed) are "competing" with fairly large type I error signals. False positives near the top of a sorted list can occur when genes having very small fold-change are compensated by small enough variance to yield a large test statistic. One of the first attempts around this caveat was the development of "significance analysis of micro-arrays (SAM)", which used a modified t-type statistic thresholded against its permutation distribution. The key innovation of the modified t-statistic was the addition of a constant to the per gene standard errors in order to stabilize the coefficient of variation of the resulting test statistic. Since then, several authors have proposed the use of shrinkage variance estimators in conjunction with t-type, and more generally, ANOVA type tests at the gene level. Our new approach proposes the use of a shrinkage variance Hotelling T-squared statistic in which the per gene sample covariance matrix is replaced by a shrinkage estimate borrowing strength from across all genes. It is demonstrated that the new statistic retains the F-distribution under the null, with added degrees of freedom in the denominator. Advantages of this class of tests are (i) flexibility in that a whole family of hypothesis tests is possible (ii) the gains of the above-mentioned earlier innovations are enjoyed more fully. This paper summarizes our results and presents a simulation study benchmarking the new statistic against another recently proposed statistic.

*Keywords:* microarray; differential expression; empirical Bayes; ANOVA

# 1 Introduction

Gene expression microarrays provide a fast and systematic way to identify genes differentially expressed between two or more experimental groups of samples in a hypothesis driven study. These samples and experimental groups could be, for example, human prostate cancer cell line RNA samples treated with two or three different agents, or treated with the same agent at differing concentrations.

Right now cDNA chips contain on the order of ten thousand genes, while oligonucleotide arrays contain upwards of twelve thousand genes. In the not too distant future entire genome chips will become available. The consequence is a tremendous savings in time and resources as the per gene expense in time and resources for the preliminary screening of genes has gone down considerably. Nonetheless, the considerable cost per array results in experiments that are typically based upon few replicates. For example, an experiment consisting of two experimental conditions might have just three replicates per set of conditions.

While the shift in platfroms from cDNA arrays to oligonucleotide arrays has resulted in the reduction in various sources of within gene and extra gene variability, the reality is that there is still a great deal of endemic noise in these sorts of investigations. Given the small number of replicates, power is a primary concern. Albeit, the goal of statistical analysis in this setting is to arrive at a

relatively short list of candidate genes that warrant further investigation via a more sensitive and specific technique such as PCR. The investigator typically has aloted specific resources for the further investigation of a given number of genes and will request a "short list" of the requesite length. Therefore, the role of efficiency and power may not be completely appreciated. Clearly, however, the goal is to present the best possible list, so that the role of efficiency and power can now be understood.

A caveat arises in that true signals (genes truly over or under expressed) are "competing" with fairly large type I error signals. False positives near the top of a sorted list can occur when genes having very small fold-change are compensated by small enough variance to yield a large test statistic. One of the first attempts around this caveat was the development of "significance analysis of micro-arrays" or (SAM), [11], which used a modified t-type statistic thresholded against its permutation distribution. The key innovation of the modified t-statistic was the addition of a constant to the per gene standard errors in order to stabilize the coefficient of variation of the resulting test statistic. Since then, [12], [6], several authors have proposed the use of shrinkage variance estimator in conjunction with t-type and more generally, ANOVA type tests at the gene level. One advantage of this latter approach is that it doesn't require the computation of ad-hoc fudge constants. In the situation under study, e.g. a hypothesis driven

experiment consisting of a small number of experimental groups, a natural model is the per gene linear model on the appropriate scale, leading to a per gene ANOVA type test of the null. Recent work, [12], [2], [6] presented a model in which the per gene residual variance parameters were considered to be draws from an inverse gamma distribution, resulting in a "shrinkage variance test" that could potentially have gains in efficiency depending on the heterogeneity of the extra gene variability. The idea of using a shrinkage estimate of within group variance has also been persued by others, [1], [7], [8]. One assumption of that model which is often violated in applications is that the extra gene variability is consistent across experimental conditions. In order to circumvent this restrictive assumption, we extend that work to the multivariate setting arriving now at a whole class of hypothesis tests based upon a shrinkage variance Hotelling $T^2$. If there is any appreciable between-group correlation, this approach constitutes a more efficient use of the scarce data available per gene data. Furthermore, as we shall point out in this work, the incorporation of a shrinkage variance/covariance estimator into the usual Hotelling $T^2$ statistic accomplishes the goals of the earlier innovations to an even greater degree.

## 2 Background and Motivation: Designed Gene Expression Micro-Array experiments

The impetus for this work were two microarray studies with which the authors have been involved. The first of these was a spotted cDNA array experiment studying the effects of the isoflavone/phytoestrogen genestein on gene expression in the LnCAP cell line. Several batches of colonies were treated with either $1\mu$M, $5\mu$M, $25\mu$M, genestein or control media and allowed to grow for 24 hours. Messenger RNA (mRNA) isolated from each of the treated groups was hybridized onto the green channel of a corresponding micro-array, while mRNA isolated from the control treated colony was hybridized onto the red channel of each micro-array. This experiment was conducted independently and in identical fashion on three separate dates. Systematic variability occuring from array to array and within array were adjusted out in the manner suggested by [3]. Within each experimental replicate and for each gene, the log base two of the ratio of normalized green to red channel expression values were calculated and used in subsequent analysis. The research questions being investigated were (i) whether there was differential expression between the green and red channels under treatment with genestein at any of the three concentrations, and if so (ii) was there a trend in this effect.

The second study was an oligonucleotide micro-array experiment studying

the effects of two hormones, dehydroepiandrosterone (DHEA) and dihyrotestosterone (DHT), on gene expression in the LnCAP line. Again, several batches of colonies were treated with either DHEA, dhT, or control media and allowed to grow for 24 hours. mRNA isolated from each of the two treated colonies as well as from the control treated colony was hybridized onto one of three corresponding single channel oligonucleotide arrays. The raw image files, in CEL format, were imported into the R statistical computing platform [10]. For each gene, the probe set was summarized into a model based gene expression index [5], using the Bioconductor suite of add-on libraries for R [4]. Within each experimental replicate and for each gene, the log base two of the expression ratios of treatment to control were calculated and used in subsequent analysis. The research questions here were (i) whether there was differential expression between treatment and control under treatment with either hormone, any of the three conditions, and if so (ii) was there differential expression between the two treatments.

# 3 The Shared Hotelling $T^2$ statistic

As indicated in the introductory remarks above, the new methodologic tool introduced here is a Hotelling $T^2$ statistic for a variety test of the null which incorporates a shrinkage estimate of the per gene residual variance. Suppose that

each preprocessed microarray yields expression levels on each of $G$ genes. In the type of studies dealt with here we have a total of $n \times d$ such microarrays arising from $n$ identical replicates of an experiment having $d$ experimental conditions or "treatments". Here as is usually the case, the measurements being analyzed will be the log base two of a treatment to control ratio. For each of the $1 \leq g \leq G$ genes, we consider these measurements as an i.i.d. sequence of $d$-demensional random variables, $\{Y_{g,i} : i = 1, 2, \ldots, n_g\}$, where we allow the possibility that there may be a different number of measurements for different genes due to reading errors. We assume such missingness is completely at random. Let $\overline{Y}_g$ and $S_g$ be the $d$-dimensional sample mean and unbiased sample covariance matrix corresponding to the sample $\{Y_{g,i} : i = 1, 2, \ldots, n_g\}$. Denote by $F_{n_1,n_2}$ and $F_{n_1,n_2,\theta}$ the CDFs corresponding to central and non-central $F-$distributions, respectively, of degrees $n_1$ and $n_2$, the latter having non-centrality parameter $\theta$. The following theorem shows that, under an assumed conjugate prior, we can replace the estimated covariance matrix in the usual Hotelling $T^2$ test with a shrinkage estimate and still retain the property that the resulting test has an $F$ distribution under the null hypothesis.

**Theorem 1:** *Suppose that* $\min_g n_g > d$ *and for a given gene,* $g$*, that*

1. *conditional upon* $\Sigma_g$, $\{Y_{g,i} : i = 1, 2, \ldots, n_g\}$ *is i.i.d.* $N_d(\mu, \Sigma_g)$,

2. $\{\Sigma_g : g = 1, 2, \ldots, G\}$ *is i.i.d.* $\text{InvWishart}_d(\nu, \Lambda)$ *and independant of the above.*

*Let* $T_g^2 = n_g \, \overline{Y}_g' \left(\Lambda + (n_g - 1)S_g\right)^{-1} \overline{Y}_g$.

$$\textit{Then under } H0 : \mu = 0_d, ShHT_g^2 = \frac{\nu + n_g - 2d - 1}{d} \, T_g^2 \textit{ has the } F_{d,\nu+n_g-2d-1} \textit{ distribution.} \quad (1)$$

The model in items 1 and 2 above is called the Normal–Inverse Wishart model in the following. The above statistic has the potential for fair sized gains in efficiency. The most ideal situation occurs when the average (over genes) of the within gene variability is reasonably small but there is reasonable spread across genes in the magnitude of this variation. In such a case, the parameter $\Lambda$ would not add so much magnitude to the denomenator, while the shape parameter, $\nu$ would gives us extra degrees of freedom as if we had more replicates per experimental condition. In reality there is trade off between these two phenomena, and one checks for gain in efficiency by comparing with the standard Hotelling $T^2$.

Next, we note that, as is the case in the usual Hotelling $T^2$ statistic, a whole family of statistics arises by applying a linear transformation. We state this as a corollary to the above theorem.

**Corollary 1:** *Assume conditions (1) and (2) above except without any restriction on* $d$ *and* $n_g$ *relative to oneanother. Consider the matrix* $M$*, which is chosen to be of dimension* $q \times d$ *of rank* $r < \min_g n_g$*. Then we can replace* $\overline{Y}_g, S_g, \Lambda$ *and* $d$ *by* $M\,\overline{Y}_g, MS_gM', M\Lambda M'$ *and* $r$ *in the theorem above and the conclusions still follow.*

The above theorem and its corollary are used to test a variety of null hypotheses, $H_0 : M\mu = 0$ where $\mu = \mathbb{E}Y_1$. There are three natural choices for $M$. Call these the "zero means" contrast, $M_{\mu 0}$, the "equal means" contrast, $M_{\mu \mathrm{eq}}$,and the "no trend" contrast, $M_{\mathrm{trend0}}$. Specifically, these are given by: $M_{\mu 0} = I_d$, which requires that $n > d$, $M_{\mu \mathrm{eq}} = I_d - \frac{1}{n}\mathbf{1_d}\mathbf{1_d'}$, which requires that $n > d - 1$, and $M_{\mathrm{trend0}} = \{(u\,u')^{-1}\,u'\}_2, u = [\mathbf{1_d}, [\mathbf{0}, \mathbf{1}, \ldots, \mathbf{d-1}]']$, which requires that $n > 1$ and $d > 2$. The application of these results to testing hypotheses in the analysis of both cDNA and oligonucleotide arrays will be clearly laid out in the section which follows.

Notice in the definition of the statistics $\mathrm{ShHT}_g^2$ given above in 1, the parameter matrix, $\Lambda$, and the shape parameter, $\nu$ arising in the prior distribution of $\Sigma_g$ are assumed to be "known". The next result is used to estimate $\Lambda$ and $\nu$ via maximum likelihood using the data the $S_g$, $g = 1, \ldots, G$ which under our model are i.i.d. draws from the density given below in 2.

**Theorem 2:** *Under the conditions of theorem 1,* $A_g = (n-1)S_g$ *has density function equal to*

$$f(A) = \frac{\Gamma_d\left(\frac{\nu+n_g-d-2}{2}\right)}{\Gamma_d\left(\frac{n_g-1}{2}\right)\Gamma_d\left(\frac{\nu_g-d-1}{2}\right)} \frac{|\Lambda|^{(\nu-d-1)/2}\,|A|^{(n_g-d-2)/2}}{|\Lambda+A|^{(\nu+n_g-d-2)/2}}. \quad (2)$$

# 4 Comparison with other approaches–simulation study

We conducted a simulation study in order to compare the operating characteristics of the proposed shared variance Hotelling $T^2$ statistic ($\text{ShHT}^2$) in expression 1 with those of three other statistics. In all cases the test was relative to the null hypothesis of group means identically zero, with two groups. We consider a shared univariate $T^2$ ($\text{ShUT}^2$) statistic based on the assumption of uncorrelated errors and common group variances which was the topic of [12], and similar in nature to statistics considered in [7], [12], [2], and [6]. Additionally, the standard versions of the above two shared variance statistics will also be considered: the the standard Hotelling $T^2$ ($HT^2$), [9], and the standard univariate $T^2$, ($\text{UT}^2$). The spirit of the comparison was to determine the consequences of the two major features of the statistic: its multivariate nature and its shrinkage estimate of the variance/covariance matrix. For sake of completeness, we make brief mention of each of the three statistics being compared. To this end, recall the notation used above in theorem 1 and its corollary. In addition to the quantities pre-

sented there, we write $Y_g$ for the $nd$ dimensional column vector containing the observations $\{Y_{g,i,k} : i = 1, \ldots, n_g, k = 1, \ldots, d\}$ stacked by replicate within component, and $R_g = (n-1)\sum_{k=1}^{d} S_{g,k,k}$ for the total within group sum of squares. The other three statistics being benchmarked against our own $\mathrm{ShHT}^2$ statistic are:

$$\mathrm{HT}_g^2 = \frac{n_g - d}{d} \frac{n_g}{n_g - 1} \overline{Y}_g' \, S_g^{-1} \, \overline{Y}_g \tag{3}$$

$$\mathrm{UT}_g^2 = \frac{d(n_g - 1)}{n_g d} \frac{Y_g' Y_g}{R_g} \tag{4}$$

$$\mathrm{ShUT}_g^2 = \frac{2s + d(n_g - 1)}{n_g d} \frac{Y_g' Y_g}{2r + R_g} \tag{5}$$

Under the assumption that $\{Y_{g,i} : i = 1, \ldots, n_g\}$ are multivariate normally distributed having zero mean vector and constant covariance matrix, $\Sigma_g$, the $\mathrm{HT}_g^2$ statistic in expression 3 is distributed as $F_{d,n_g-1}$, [9]. Under the assumption that $\{Y_{g,i,k} : i = 1, \ldots, n_g, k = 1, \ldots, d\}$ are normally distributed having mean zero and common variance $\sigma_g^2$, it follows from elementary results regarding ratios of chi-squared variables that the $\mathrm{UT}^2$ statistic in expression 4 is distributed as $F_{n_g d, d(n_g-1)}$. Finally, if the last mentioned assumption of normality is made conditionally upon $\sigma_g^2$, and $\{\sigma_g^2 : g = 1, \ldots, G\}$ are assumed to be i.i.d. draws from an inversge gamma distribution with shape parameter, $s$, and rate parameter, $r$, the $\mathrm{ShUT}^2$ statistic in expression 5 is distributed as $F_{n_g d, 2s+d(n_g-1)}$. Note that our version of the $\mathrm{ShUT}^2$ statistic is a test of the null that all group

means are zero, which differs from that presented in [12], but the only difference between the two statistics is in the numerator. Thus the proof contained in [12] carries over to the present setting intact, with a modification in the numerator degrees of freedom. The parameters $s$ and $r$ in the prior distribution are estimated as in [12], by fitting the observations $\{\sigma_g^2 : g = 1, \ldots, G\}$ to an inverse gamma distribution via maximum likelihood.

The first simulation study was conducted by generating data from the Normal–Inverse Wishart model with $d = 2$ groups and $n_g = 3$ replicated observations for each of G=12625 genes, using values for $\Lambda$ and $\nu$ that were obtained in the analysis of the oligonucleotide array data (see below for further details). One hundred of the genes were designated as "true positives" by giving them non-zero group specific means that were chosen in the following way. First, a value of $\theta$ was chosen so that

$$0.90 = F_{6,4,3\theta}(F_{6,4}^{-1}(1 - 0.0026))$$

i.e., so that the $\mathrm{UT}^2$ statistic would have power 90% at a type I error of 0.26% to reject the null hypothesis of zero group means. This value of $\theta = 7.5$ was then multiplied by the average per group standard deviation calculated under the Normal–Inverse Wishart model, i.e. $\frac{1}{\nu - 2d - 2}\mathrm{diag}\big[\Lambda\big]$ to arrive at the two group specific means applied identically to each of the ten designated genes.

In order to study the robustness of the test statistic to lack of model assumptions, a second simulation study was conducted using a Normal-2 component mixed inverse Wishart distribution. Specifically, the data are i.i.d. multivariate normal but the prior distribution on the random variance/covariance matrix is a mixture of two inverse Wisharts, having shape parameters $\nu_1$ and $\nu_2$ and common rate matrix $\lambda$. The mixing proportion, $f$ and shape parameters $\nu_1$ and $\nu_2$ were chose so that the expected value of $S_g$, the per gene empirical covariance matrix, would remain identical its value under the normal/inverse Wishart model used previously, $\frac{\Lambda}{\nu-2d-1}$. The values used were $f = 0.2, \nu_1 = 18.4067$, and $\nu_2 = 6.77542$. Once again, one hundred genes were designated as "true positives" by assigning means as above.

The simulation results were summarized in two ways. The first method, shown in tables 4 and 4, used the Benjamini-Hochberg FDR stepdown procedure to set the significance criterion. In each simulation replicate, the four listed statistics and corresponding p-values were calculated for each of the 12625 genes. Next, for each statistic, the list was sorted on corresponding p-value and the row containing the largest p-value not exceeding $(rank)FDR/12625$ and all rwos above it were marked significant. The true positive rate was derived as the number of genes called significant as a proportion of those truely differentially expressed, i.e. 100. The false positive rate was derived as the number of

genes called significant not among those 100. These were averaged over simulation replicates yielding emprical true positive rates ($eTPR$) and empirical false positive rate ($eFPR$). In table 4 is shown results for the data simulated from the normal/Inverse Wishart model. The leftmost column is the nominal false discovery rate, $FDR$, used in setting the significance criterion. The next eight columns are the empirical true positive and false positive rates for each of the four benchmarked statistics.

Results corresponding to data simulated from the normal/inverse Wishart model are shown in table 4. In the case of the proposed statistic, $ShHT2$, the $eFPR$ coincides within simulation error with the FDR. That is because the p-values are derived via the F-distribution listed in theorem 1, which assumes the data arise from a normal/inverse Wishart distribution. Notice as well that the $eTPR$ is quite high in the 90's at the low FDR of 0.05. The other three statistics benchmarked a clearly inferior. First, $HT2$, thestandard hotelling $T^2$, is nearly uninformative, displaying an $eTPR$ of 100% at all values of $FDR$ with correspondingly high $eFPR$ ranging upwards from 85%. The shrinkage variance F-statistic, $ShUT2$, is overly concervative, with $eFPR$ equal to zero within simulation error and $eTPR$ ranging from 20% to 50%. Finally, the ordinary F-statistic, $UT2$, is overly conservative at the lower FDR's of 5% and 10%, but then uninformative at the higher FDR's of 15%, 20% and 25%.

The results corresponding to data simulated from the normal/mixed inverse Wishart model are shown in 4. The only notable difference relateive to remarks made above is that control over the FDR is now lost, as the $eFPR$ no longer agrees with the $FDR$. Still, if the simulation model can be considered an extreme departure from the model assumptions then use of the FDR=5% which gives $eFPR = 12\%$ and $eTPR = 94\%$ should be acceptible.

On the other hand one may wish to dispesne with any attempts at controling the false discovery rate at all, and instead, rely on the statistic's ability to provide a more informative ordering. In this case, we simply decide how many genes we wish to call significant and draw the line there. For the second method of summarizing the simulation results the $eTPR$ and $eFPR$ were derived this time using, consecutively, each of the values of the statistic as the significance criterion. The results for data obeying model assumptions are shown in figures 1, and for data not obeying model assumptions in figure 2.

Table 1

| | ShHT2 | | HT2 | | ShUT2 | | UT2 | |
|---|---|---|---|---|---|---|---|---|
| FDR | TPF | FPF | TPF | FPF | TPF | FPF | TPF | FPF |
| 0.05 | 0.929 | 0.045 | 1.00 | 0.857 | 0.204 | 0.00 | 0.014 | 0.00 |
| 0.10 | 0.964 | 0.093 | 1.00 | 0.927 | 0.341 | 0.00 | 0.079 | 0.00 |
| 0.15 | 0.976 | 0.141 | 1.00 | 0.951 | 0.424 | 0.00 | 1.00 | 0.992 |
| 0.20 | 0.983 | 0.192 | 1.00 | 0.963 | 0.480 | 0.00 | 1.00 | 0.992 |
| 0.25 | 0.987 | 0.242 | 1.00 | 0.970 | 0.526 | 0.00 | 1.00 | 0.992 |

Table 2

| | ShHT2 | | HT2 | | ShUT2 | | UT2 | |
|---|---|---|---|---|---|---|---|---|
| FDR | TPF | FPF | TPF | FPF | TPF | FPF | TPF | FPF |
| 0.05 | 0.944 | 0.123 | 1.00 | 0.863 | 0.424 | 0.00 | 0.343 | 0.000 |
| 0.10 | 0.968 | 0.223 | 1.00 | 0.929 | 0.556 | 0.00 | 0.593 | 0.000 |
| 0.15 | 0.976 | 0.307 | 1.00 | 0.951 | 0.626 | 0.00 | 1.000 | 0.992 |
| 0.20 | 0.981 | 0.378 | 1.00 | 0.963 | 0.671 | 0.00 | 1.000 | 0.992 |
| 0.25 | 0.985 | 0.442 | 1.00 | 0.970 | 0.706 | 0.00 | 1.000 | 0.992 |

# 5 Application: Two Case Studies

Table 5

| | Gene | dhea | dht | stat | p-val | FDR=0.10 |
|---|---|---|---|---|---|---|
| 1 | 34319_at | 1.690 | 4.400 | 273.0 | 1.902e-07 | 7.129e-06 |
| 2 | 36658_at | 2.440 | 2.790 | 90.9 | 8.665e-06 | 1.426e-05 |
| 3 | 33998_at | 0.519 | 0.275 | 85.1 | 1.084e-05 | 2.139e-05 |
| 4 | 38827_at | 1.310 | 1.840 | 64.1 | 2.836e-05 | 2.851e-05 |

# References

[1] P. Baldi and A. D. Long, *A bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes*, Bioinformatics **17** (2001), 509–519.

[2] X.-G. Cui, J.-T. G. Hwang, J. Qui, N. J. Blades, and Churchill G. A., *Improved statistical tests for differential gene expression by shrinking variance components estimates*, Biostatistics (2005).

[3] S. Dudoit, Y.-H. Yang, M. J. Callow, and T. P. Speed, *Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments*, Tech. report, UCB statistics, 2000.

[4] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y.-C. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y.-H. Yang, and J.-H. Zhang, *Bioconductor: Open software development for computational biology and bioinformatics*, Genome Biology (2004).

[5] C. Li and W. H. Wong, *Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection*, PNAS (2001).

[6] I. Lönnstedt, R. Rimini, and P. Nilsson, *Empirical bayes microarray ANOVA and grouping cell lines by equal expression levels*, Statistical Applications in Genetics and Molecular Biology **4** (2005), Article 7.

[7] I. Lönnstedt and Speed T., *Replicated microarray data*, Statistica Sinica **12** (2002), 31–46.

[8] R. Menezes, *Heirarchical modeling to handle heteroscedacisity in microarray data*, Presentation at 3-day Workshop on Statistical Analysis of Gene Expression Data, Richardson,S. and Brown,P. Organizers, 2003.

[9] R. J. Muirhead, *Aspects of multivariate statistical theory*, John Wiley & Sons, Hoboken, NJ, 1982.

[10] R Development Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2004, ISBN 3-900051-00-3.

[11] V. G. Tusher, R. Tibshirani, and G Chu, *Significance analysis of microarrays applied to the ionizing radiation response*, PNAS (2001).

[12] G. W. Wright and R. M. Simon, *A random variance model for differential gene detection in small sample microarray experiments*, Bionformatics **19** (2003), 2448–2455.

# Appendix Proofs of theorems

**Proof of theorem 1:** In the following, for any symmetric matrix with spectral decomposition $A = QDQ'$, let $A_{\mathrm{s}}^{\frac{1}{2}} = QD^{\frac{1}{2}}Q'$ be the symmetric square root of $A$. $A_{\mathrm{s}}^{-\frac{1}{2}}$ is the symmetric square root of $A^{-1}$. Square root matrices without the subscript s are considered Cholesky square roots, but will not appear in this

manuscript. First, rewrite $T^2$ as follows:

$$\begin{aligned} T^2 &= n\left(\Sigma_s^{-\frac{1}{2}}\,\bar{Y}\right)\left(\Sigma_s^{-\frac{1}{2}}\Lambda\Sigma_s^{-\frac{1}{2}} + (n-1)\Sigma_s^{-\frac{1}{2}}S\Sigma_s^{-\frac{1}{2}}\right)^{-1}\left(\Sigma_s^{-\frac{1}{2}}\,\bar{Y}\right) \\ &\stackrel{\mathcal{D}}{=} n\left(\Sigma_s^{-\frac{1}{2}}\,\bar{Y}\right)\left(\Lambda_s^{\frac{1}{2}}\Sigma^{-1}\Lambda_s^{\frac{1}{2}} + (n-1)\Sigma_s^{-\frac{1}{2}}S\Sigma_s^{-\frac{1}{2}}\right)^{-1}\left(\Sigma_s^{-\frac{1}{2}}\,\bar{Y}\right), \end{aligned}$$

where equality in distribution follows from the fact that because $\Sigma_s^{-\frac{1}{2}}\Lambda\Sigma_s^{-\frac{1}{2}}$ and $\Lambda_s^{\frac{1}{2}}\Sigma^{-1}\Lambda_s^{\frac{1}{2}}$ are both positive definite and symetric, it follows from theorem A9.9 of [9] that they are an orthogonal similarity transformation of eachother and since the latter has a Wishart distribution (see below), equality in distributions follows from the invariance of the Wishart distribution to orthogonal similarity transformations.

Next we make the following observations:

1. $(n-1)\Sigma_s^{-\frac{1}{2}}S\Sigma_s^{-\frac{1}{2}}$ has the $\text{Wishart}_d(n-1, I_d)$ distribution and is therefore, independent of $\Sigma$. This is because the conditional distribution of $(n-1)S$ given $\Sigma$ is $\text{Wishart}_d(n-1, \Sigma)$.

2. $\Lambda_s^{\frac{1}{2}}\Sigma^{-1}\Lambda_s^{\frac{1}{2}}$ has the $\text{Wishart}_d(\nu-d-1, I_d)$ distribution, because $\Sigma^{-1}$ has the $\text{Wishart}_d(\nu-d-1, \Lambda^{-1})$.

Therefore, the sum,

$$V = \Lambda_s^{\frac{1}{2}}\Sigma^{-1}\Lambda_s^{\frac{1}{2}} + (n-1)\Sigma_s^{-\frac{1}{2}}S\Sigma_s^{-\frac{1}{2}}$$

has the $\text{Wishart}_d(\nu+n-d-2, I_d)$ distribution. Next, put $Z = \Sigma_s^{-\frac{1}{2}}\,\bar{Y}$ and

rewrite $T^2$ as

$$T^2 = nZ'V^{-1}Z = \frac{nZ'Z}{\frac{Z'Z}{Z'V^{-1}Z}}.$$

Notice that since $Z$ has been rescaled, it is independant of $\Sigma$. Next, because $Z$ is the sample mean, it is independant of the sample covariance matrix, $S$. Thus $Z$ and $V$ are independant. Next, it follows from theorem 3.2.12 of [9], the denominator is distributed $\chi^2_{\nu+n-2d-1}$ and independent of the $Z$. Because the numerator is $\chi^2_d$, it follows that $T^2 \ \frac{\nu+n-2d-1}{d}$ has the $F_{d,\nu+n-2d-1}$ distribution.

**Proof of theorem 2:** As stated above, the conditional distribution of $A = (n-1)S$ given $\Sigma$ is $\text{Wishart}_d(n-1, \Sigma)$ which has density:

$$f^W_{d,n-1,\Sigma}(A) = 2^{-d(n-1)/2}\Gamma_d\left(\frac{n-1}{2}\right)^{-1}\frac{|A|^{(n-d-2)/2}}{|\Sigma|^{(n-1)/2}}\text{etr}\left(-\frac{1}{2}\Sigma^{-1}A\right)$$

while $\Sigma$ has the $\text{InvWishart}_d(\nu, \Lambda)$ distribution, which has density:

$$f^{W^{-1}}_{d,\nu,\Lambda}(\Sigma) = 2^{-d(\nu-d-1)/2}\Gamma_d\left(\frac{\nu-d-1}{2}\right)^{-1}\frac{|\Lambda|^{(\nu-d-1)/2}}{|\Sigma|^{\nu/2}}\text{etr}\left(-\frac{1}{2}\Sigma^{-1}\Lambda\right)$$

Taking the product of the two above densities and reorganizing factors yeilds:

$$f^W_{d,n-1,\Sigma}(A)f^{W^{-1}}_{d,\nu,\Lambda}(\Sigma) = 2^{-d(\nu+n-d-2)/2}\Gamma_d\left(\frac{\nu+n-d-2}{2}\right)^{-1}\frac{|\Lambda+A|^{(\nu+n-d-2)/2}}{|\Sigma|^{(\nu+n-1)/2}}\text{etr}\left(-\frac{1}{2}\Sigma^{-1}(\Lambda+A)\right)$$
$$\frac{\Gamma_d\left(\frac{\nu+n-d-2}{2}\right)}{\Gamma_d\left(\frac{n-1}{2}\right)\Gamma_d\left(\frac{\nu-d-1}{2}\right)}\frac{|\Lambda|^{(\nu-d-1)/2}\,|A|^{(n-d-2)/2}}{|\Lambda+A|^{(\nu+n-d-2)/2}}.$$

Thus, the posterior distribution of $\Sigma$ given $A = (n-1)S$ is

$\text{InvWishart}_d(\nu+n-1, \Lambda+(n-1)S)$, and so the distribution of $A = (n-1)S$ is the one given in expression 2.

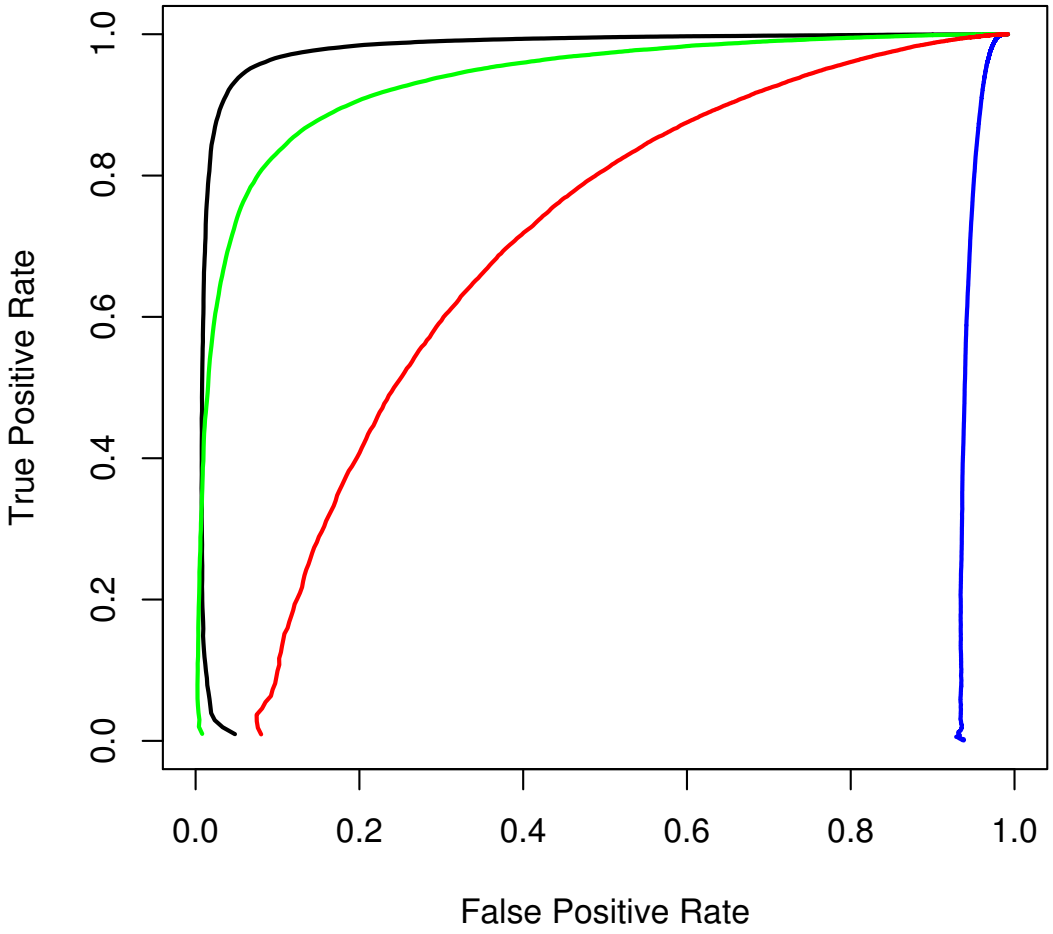


Figure 1: In data simulated from the normal/Inverse Wishart model, the true positive and false positive rates for the four benchmarked statistics as cutoff ranges through all possible values of the statistic. ShHT2=black, ShUT2=green, UT2=red, HT2=blue

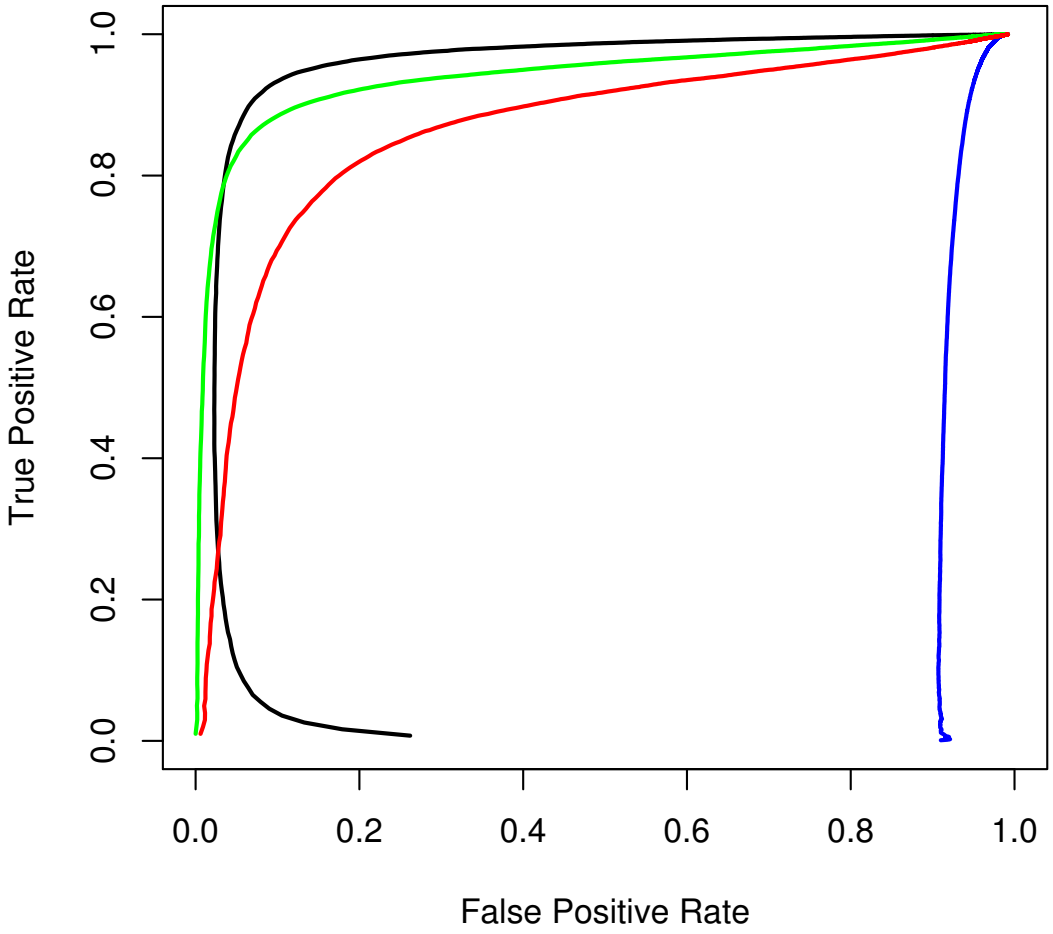


Figure 2: In data simulated from the normal/mixed Inverse Wishart model, the true positive and false positive rates for the four benchmarked statistics as cutoff ranges through all possible values of the statistic. ShHT2=black, ShUT2=green, UT2=red, HT2=blue