BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–7

The Shrinkage Variance Hotelling T-Squared Test for Small Sample Micro-Array Profiling Studies

Grant Izmirlian* and Jian-Lun Xu

National Cancer Institute; Executive Plaza North, Suite 3131; 6130 Executive Blvd, MSC 7354; Bethesda, MD 20892-7354

ABSTRACT

Motivation: 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 false positives near the top of the sorted list can occur when genes having very small fold-change are compensated by small enough variance to yield a large test statistic. Several authors have proposed the incorporation of shrinkage estimators to stabilize the estimated per gene variance paramters. We propose 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.

Results: Benchmarking simulation study comparing the performance of the proposed statistic against another recently proposed statistic was conducted using data drawn from (i) our model assumptiions: conditionally multivariate normal with inverse Wishart variance/covariance (ii) the model assumptions of the recently proposed statistic: conditionally i.i.d. normal with inverse gamma variance and (iii) a model violating both sets of assumptions. The previously proposed statistic performs slightly better only under its more restrictive set model assumptions. In all other cases our statistic performs as well when expected group variances are the same and better when the expected group variances differ.

Availability: An R package, SharedHT2, capable of performing all calculations mentioned in this manuscript and more is available from the R project website, www.r-project.org, in the contributed packages section.

Contact: izmirlian@nih.gov

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 appreciated in these terms.

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), Tusher *et al*., 2001, 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, (Wright and Simon, 2003, Lönnstedt and Speed, 2002) 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, (Wright and Simon, 2003, Cui *et al*., 2005, Lönnstedt *et al*., 2005), presented a model in which the per gene residual variance parameters were considered to be draws from an inverse gamma distribution, resulting in a "(equal variance) 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 such as Baldi and Long, 2001, and Menezes, 2003. A common assumption made is that, conditional upon

*to whom correspondence should be addressed

a per gene random variance parameter, that the experimental group expression values are independant and have common variances. 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-squared. 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-squared 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 1μM, 5μM, 25μM, 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 Duduoit *et al*., 2000. 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 (R Development Core Team, 2004). For each gene, the probe set was summarized into a model based gene expression index (Li and Wong, 2001), using the Bioconductor suite of add-on libraries for R (Bioconductor Development Team, 2004). 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. These questions were investigated at two levels: a stringent criterion produced four significantly differentially expressed genes, and then a less stringent criterion was used to form a longer list for cross-referencing with simple gene ontology genelists for purposes of identifying potentially important pathways. We return to discussion of these two applications in a later section.

3 THE SHARED HOTELLING T-SQUARED STATISTIC

As indicated in the introductory remarks above, the new methodologic tool introduced here is a Hotelling T-squared statistic for a variety test of the null which incorporates a shrinkage estimate of the per gene residual variance/covariance matrix. 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 $\bar{Y}_g$ 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 θ. The following theorem shows that, under an assumed conjugate prior, we can replace the estimated covariance matrix in the usual Hotelling T-squared test with a shrinkage estimate and still retain the property that the resulting test has an F distribution under the null hypothesis with the usual degrees of freedom incremented by the shape parameter corresponding to the prior.

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

1. *conditional upon* Σ_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.* $\mathrm{InvWishart}_d(\nu, \Lambda)$ *and independant of the above.*

Let $T_g^2 = n_g \bar{Y}_g' \Big((n_g - 1) S_g + \Lambda \Big)^{-1} \bar{Y}_g$.

$$\textit{Then under } H0 : \mu = 0_d, \quad \mathrm{ShHT2}_g = \frac{n_g + \nu - 2d - 1}{d} T_g^2$$
$$\textit{has the } F_{d, n_g + \nu - 2d - 1} \textit{ distribution.} \tag{1}$$

The model in items 1 and 2 above is called the multivariate normal/inverse Wishart (MVN/IW) 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 Λ would not add so much magnitude to the denomenator, while the shape parameter, ν 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-squared.

Next, we note that, as is the case in the usual Hotelling T-squared 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 $\bar{Y}_g, S_g, \Lambda$ and d by $M\bar{Y}_g, MS_gM', M\Lambda M'$ 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, $M_{\mu 0} = I_d$, the "equal means" contrast, $M_{\mu \text{eq}} = I_d - \frac{1}{n} J_{d,d}$,and the "no trend" contrast, $M_{\text{trend0}} = \{(u\,u')^{-1}\,u'\}_2$. Here, I_d is the d-dimensional identity matrix, $J_{p,q}$ is a $p \times q$ dimensional matrix of ones, and $u = [J_{d,1}, [0, 1, \ldots, d-1]']$. Note the requirement in corollary 1 above that the rank of M be less than the number of replicates, n. This results in the following requirements for each of the three above mentioned contrasts: (i) zero means: $n > d$, (ii) equal means: $n > d - 1$, (iii) no trend: $n > 1$ 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 ShHT2_g given above in 1, the parameter matrix, Λ, and the shape parameter, ν arising in the prior distribution of Σ_g are assumed to be "known". The next result is used to estimate Λ and ν via maximum likelihood using data on the per gene empirical variance/covariance matrices, 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*

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

4 SOFTWARE: R PACKAGE SHAREDHT2

An R package for conducting analyses using the methods of this paper has been created and is available for download at the CRAN website. One of the most desirable facets of this package is that it is entirely coded in C with minimal processing done in R. The main function "EB.Anova" fits the MVN/IW model to micro-array data and calculates the per gene $ShHT2$ statistics shown in formula 1 in theorem 1 when the argument "Var.Struct" is set to "general". In this case, the ordinary HT2 statistics are calculated as well. In addition, the same function can be used to fit the normal/inverse gamma model of Wright and Simon, 2003 and calculate the $ShAnF$ statistics shown in formula 4 in the preceding section by setting the argument "Var.Struct" to "simple". In this case, the ordinary AnF statistics are calculated as well. In both cases, the models are fit using maximum likelihood estimation. There is flexibility in the choice of hypothesis test via setting the argument "H0" to one of the following choices. Under the "general" variance structure option, (i) if $n > d$, the H0="zero.means" null may be tested, (ii) if $n > d - 1$, the H0="equal.means" null may be tested, and (iii) if $n > 1$, the H0="no.trend" null may be tested, but of course this only makes sense if $d > 2$. Also, (iv) the user may also set H0 to an custom contrast matrix of dimension $r \times d$ and of rank r. Under the "simple" variance structure option, any of the above null hypotheses may be tested as long as $n > 1$. By default, "H0" is set to "equal.means". The package uses S3 classes, has defined 'print', 'update' and 'as.data.frame' for the class "fit.n.data", which contains a model fit and a data-frame component. There are other nice features. For example, if the user specifies "na.action=na.pass" then NA's are treated as missing-at-random as long as the minimum number of replicates per group meets the above requirements. Next, the genelist, sorted on p-value corresponding to either of the two computed statistics, can be browsed in the html viewer. If the data comes from an affy experiment and the rows are named after the affy gene identifiers, then the rows of the genelist in the html viewer are linked to the Weizmann Institute's "GeneCards" database, or an alternate genecards database of the users choice. Additionally, the simulation study presented in the following section may be repeated using included functions. It is worth mention here that these simulations were only made possible by migrating the entire procedure including the loop over simulation replicates, into C. The interested reader is encouraged to browse the documentation.

5 OTHER STATISTICS UNDER STUDY

In anticipation of the simualation study presentation which is to follow, we now present three other statistic against which we benchmarked our proposed ShHT2 statistic, together with corresponding distributions under the null. As mentioned in the introduction, the results presented here generalize results in Wright and Simon (2003). Thus the main comparison will be with the statistic proposed in that work, which we shall call the shared ANOVA F-statistic (ShAnF). The ShHT2 and ShAnF are each themselves shrinkage-variance modifications of standard textbook statistics, namely the Hotelling T-squared (HT2) and anova F (AnF) statistics. Thus we will benchmark our proposed ShHT2 statistic agasint the ShAnF, the HT2 and the AnF. In our benchmarking comparisons, we restricted attention to tests of the null hypothesis that all group means are zero. This is of course no loss of generality as tests of all other nulls amount to a mere linear transformation of the data. The following expressions are based upon the notation of theorem 1 and its corollary above, only with the dependence on gene, g, suppressed. In addition to the quantities presented there, Y will indicate the vector of measurements on a single gene stacked first by replicate and then by group, and we write $R = (n-1)\sum_{k=1}^{d} S_{k,k}$ for the total within group sum of squares.

- Anova F Statistic

$$\text{AnF} = \frac{n-d}{d}\frac{n^2\bar{Y}'\bar{Y}}{R}. \quad (3)$$

  If the vector Y is independent normal with group specific means and common variance, v, then under the null hypothesis that the group means are zero, the AnF statistic has the F distribution with d and $n-d$ degrees of freedom.

- Shared Anova F Statistic

$$\text{ShAnF} = \frac{2s+n-d}{d}\frac{n^2\bar{Y}'\bar{Y}}{R+2r}. \quad (4)$$

  If, conditional upon v, Y is independant normal with group specific means and common variance, and v is distributed as

an inverse gamma with shape and rate parameters $2s$ and $2r$ (Norm/IG), then under the null hypothesis that the group means are zero, the ShAnF statistic has the F distribution with d and $n - d + 2s$ degrees of freedom, (Wright and Simon, 2003).

- Hotelling T-squared

$$\mathrm{HT2} = \frac{n-d}{d} n \bar{Y}' \left((n-1)S\right)^{-1} \bar{Y}, \tag{5}$$

If the sequence of random vectors $\{Y_i : i = 1, \ldots, n\}$ is independent multivariate normal having group specific means and common variance/covariance matrix $\mathbb{\Sigma}$, then under the null hypothesis that the group means are zero, the $\mathrm{HT2}$ statistic has the F distribution with d and $n - d$ degrees of freedom, (Muirhead, 1982).

- Shared Hotelling T-squared
We reiterate the results from theorem 1 above.

$$\mathrm{ShHT2} = \frac{\nu + n - 2d - 1}{d} n \bar{Y}' \left(\Lambda + (n-1)S\right)^{-1} \bar{Y}, \tag{6}$$

If, conditional upon $\mathbb{\Sigma}$, the sequence of random vectors $\{Y_{g,i} : i = 1, \ldots, n_g\}$ is MVN/IW, i.e. conditionally independent multivariate normal given $\mathbb{\Sigma}$, having group specific means and common variance/covariance matrix $\mathbb{\Sigma}$, and $\mathbb{\Sigma}$ is distributed as an inverse Wishart with shape parameter, ν, and rate matrix parameter, Λ, then under the null hypothesis that the group means are zero, the $\mathrm{ShHT2}$ statistic has the F distribution with d and $\nu + n - 2d - 1$ degrees of freedom.

6 BENCHMARKING SIMULATION STUDY

In our simulation study, we compared performance of our proposed ShHT2 statistic with that of the ShAnF, HT2, and AnF statistics. The simulation study was designed around the concept that the generated datasets would share basic characteristics with the oligonuceotide array data encountered in our substantive work at the stage of analysis. Thus, the datasets being simulated were meant to be three measurements on the logged ratios between the RNA expression of LnCAP cells under treatment with each of the two hormones and that under treatment with a placebo. Each resulting dataset was therefore a matrix of dimension 12625 by 6, with rows corresponding to genes in the oligonucleotide array and columns corresponding to the three replicated measurements on each of two groups. We conducted a series of simulation studies. In each of these, the number of simulation replicates was 500, and the gene array data was given a group specific mean plus a mean zero noise. Again, in all cases, the group specific means were identically zero in all cases but the first 100 of the 12625 rows (true positives). These groups were given nonzero means large enough to be detected by the AnF statistic with fairly high probability. Specifically, a value of θ was chosen so that

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

i.e., so that the AnF 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 from the simulation distribution. We conducted five separate simulation studies, each using a different simulation distribution for the mean zero noise in order to compare the robustness to lack of model assumptions of our ShHT2 statistic with that of the ShAnF. The first was taken from the Norm/IG family, under which the ShAnF has the null distribution presented below formula 4, the second and third were taken from the MVN/IW family, under which the ShHT2 has the null distribution presented in formula 1 in theorem 1. The fourth and fifth were taken from a third family consisting of conditionally multivariate normal with random covariance drawn from a mixed inverse Wishart distribution (MVN/mxIW). This latter mixture was a two component mixture over the shape parameter of two inverse Wisharts having identical rate matrix. For sake of consistency, parameters were chosen so that the expected value of within group variances were roughly the same from family to family. Particular values of the parameters used in all five cases are displayed in table 1. Note in particular that the pairs of simulations, (2,3) and (4,5) from the MVN/IW and MVN/mxIW families, respectively, consist of one member with equal expected group variances with roughly 0.67 expected correlation and a second member with unequal expected group variances and nearly zero expected correlation. We refer to these as -a and -b respectively.

We assessed the performance of the four statistics, ShHT2, HT2, ShAnF and AnF, under each of the five simulation distributions, Norm/IG, MVN/IW-a, MVN/IW-b, MVN/mxIW-a, and MVN/mxIW-b, in two different ways. The first method of assessment used the nominal distribution corresponding to each statistic to select important genes, within each simulation replicate, using the Benjamini-Hochberg (B-H) false discovery rate (FDR) procedure, (Benjamini and Hochberg, 1995). This algorithm selects important genes from a list sorted on p-values and marks all genes having p-value not greater than the largest not exceeding its rank times the FDR divided by the number of tests. Their theorem guarantees that if the nominal null distribution used is correct then the expected proportion of false discoveries is less than FDR. B-H FDR thresholding was done at each of the nominal false discovery rates (FDR's): 1%, 5%, 10%, 15%, 20%, and 25%. At each of these FDR's, the empirical true discovery and false discovery rates (eTDR and eFDR, respectively) were computed by counting within each simulation replicate, the number of genes selected that were and were not among the 100 true positive genes, respectively, and then dividing the first of these by 100 and the second of these by the number of genes selected. These per simulation replicate values were then averaged over 500 simulations. These are displayed in tables 2 (Norm/IG noise), 3 (MVN/IW-a noise), 4 (MVN/IW-b noise), 5 (MVN/mxIW-a noise), and 6 (MVN/mxIW-b noise). First, observe that the eFDR's for the ShHT2 when the simulation distribution comes from the MVN/IW family (column 3 in tables 3 and 4) are within simulation error of the nominal FDR values. This is not surprising as it is exactly what the main theorem of Benjamini and Hochberg, 1995, guarantees. The same general tendency is the true in the case of the eFDR's for the ShAnF when the simulation distribution comes from the Norm/IG family (column 7 in table 2), although not to the same degree of precision. The reason for this is not entirely clear. In both cases however, when model assumptions are met one can feel confident that the FDR is being controlled at the nominal rate. Next we observe what hapens to this level of control over the observed FDR when model assumptons are not true. Note that for the proposed ShHT2 statistic, the eFDR's in tables 2 (Norm/IG noise), 5 (MVN/mxIW-a noise), and 6 (MVN/mxIW-b noise) is inflated relative to the nominal FDR by a factor ranging from 2 to 3 in all cases except the first line of table 2. This is in contrast to the ShAnF statistic. Although the eFDR is inflated by roughly the same amount when the data are distributed according to

the SimMVN/IW-a family, which has correlated groups with equal variance on average, this 2 fold factor is exacerbted to roughly 4 fold under the SimMVN/IW-b data, which has nearly uncorrelated groups but unequal group variances. When the data comes from the SimMVN/mxIW (both -a and -b) the observed eFDR's are inflated by a factor between 4 and 5 in most all cases. On one hand it could be argued that this is not a fair comparison since our proposed statistic is judged under one restricted model departure and one relaxed model departure, while the ShAnF is judged under two levels of relaxed model departures. However, a counter-argument is our practical experience, which indicates that reality lies somewhere closer to the less restricted families of models. Next, there is not that much to be said regarding comparison of the eTDR's as all simulation studies were conducted under essentially a single alternative. It is the case that when one compares lines for which the eFDR's for the proposed ShHT2 and the ShAnF are nearly equal, the ShAnF enjoys a slightly higher eTDR. This gain, due to the single per gene variance parameter in the ShAnF as opposed to three in the ShHT2, is however fairly small. All things considered, in terms of having a reasonable level of control over eFDR with an acceptibly strong eTDR under all cases considered, the proposed ShHT2 seems to be preferable. Notice, incidentally, that at the alternative considered the HT2 and AnF are underpowered.

Next, if one is willing to part with control over type I errors and leave issues of validity up to the confirmation to bench science or cross-referencing via pathways, then we don't require the use of null distributions at all and consider instead, only the statistics validity in ranking genes. In order to address this issue, the second method of assessment selected important genes within each simulation replicate and for each of the four statistics, by thresh-holding on the ranks of the statistic. At each resulting thresh-hold criterion, the eTDR's and eFDR's were computed as described above and then averaged over the 500 simulations. For each of the four statistics considered we display plots of the resulting values of the eTDR versus values of the eFDR in figures 1 (Norm/IG noise), 2 (MVN/IW-a noise), 3 (MVN/IW-b noise), 4 (MVN/mxIW-a noise), and 5 (MVN/mxIW-b noise). In terms of validity in ranking, the ShHT2 outperforms the ShAnF by a modest amount when the expected group variances are unequal, regardless of whether the data obeys the model assumptions as in figure 3 (MVN/IW-b noise) or does not as in figure 5 (MVN/mxIW noise). The situation is reversed when the data obeys the model assumptions corresponding to the ShAnF as seen in figure 1. When the data exhibit correlated groups with equal expected group variances as in figures 2 (MVN/IW-a noise) and 4 (MVN/mxIW-b noise) it appears to be too close to call. Finally, note that the textbook statistics, AnF and HT2 lag behind quite substantially, although with at least moderately respectable showing from the AnF. This is due to its noted robustness properties. The reason why the HT2 performs so poorly is due to the small number of replicates which causes instability in the estimated per gene variance/covariance matrices. This is precisely what the shrinkage variance estimate is able to circumvent.

7 APPLICATION: TWO CASE STUDIES

We now turn to the analysis of the two motivating case studies mentioned in the introduction. The first of these was a spotted cDNA array experiment seeking to determine the presence of differential gene expression, relative to control, of varying concentrations of genestein, and if so, whether or not this effect exhibited a dose response phenomenon at the concentrations studied. In this case there were three experimental groups (1μM, 5μM and 25μM genestein) each with a matched control. After preprocessing as discussed in the introduction, logged base 2 of the expression ratios of treatment to control yielded three replicates in each of three groups. Since the number of replicates is equal to the number of groups, the ShHT2 tests of zero group means were infeasible. Instead, we conducted ShHT2 tests of equal group means and the ShHT2 tests of zero trend. In addition, ShAnF tests of zero group means, equal group means and zero trend were calculated. Unfortunately there were no significant results to report at FDR less than the high 90's at which all genes were marked. However, since as we have shown, the ranks of the shrinkage variance statistics (both the ShHT2 and the ShAnF) can be considered more reliable than the naive text-book variants, the following observations are worth highlighting. The result of the intersection of the top 100 genelists corresponding to each of the abovementioned feasible tests produced twelve genes listed in table **??**. Of course since the selected inferential procedure turned up nothing not much weight can be assigned to the findings. Secondly, the agreement of several differing analyses does not constitute validity in these findings, (**?**, Ransohoff, 2005). However, it is interesting to note that of the top five listed genes in table 7 are a gene associated with sperm maturation and apoptosis, (**?**), a gene encoding a RAS-like nuclear G protein, (Hayashi *et al*., 1995), a gene differentially expressed in human gastric cancer, (Nishigaki, 2005), and a growth hormone signaling androgen receptor gene, (Weiss-Messer, 2004). This confirms the findings of our simulation study that orderings based upon either the ShHT2 or ShAnF are more reliable than an ordering based upon the text-book variants. A second study using affymetrix oligonucleotide arrays to investigate the effects of genestein on LnCAP gene expression used the Wilcoxon rank sum test at p ¡ 0.01 to perform the initial screening for potentially important genes (Takahashi *et al*., 2004), followed by the text-book AnF statistic. While it is reasonable to expect that a non-parametric test will produce more reliable results than the AnF and HT2, one can only speculate on the reliability of such a screen based upon an unstabilized non-parametric statistic being used in the setting of such a small sample. Recall that the efficiency of the Wilcoxon test relative to the AnF is 0.87.

The second study used Affymetrix oligonucleotide micro-arrays to study the effects of two hormones, dehydroepiandrosterone (DHEA) and dihyrotestosterone (DHT), on gene expression in the LnCAP cell line. After preprocessing as discussed in the introduction, logged base 2 of the expression ratios of treatment to control yielded three replicates in each of two groups. ShHT2 tests for zero group means at a FDR of 10% identified four significatly differentially expressed genes. Two of these four results were due to significant differential expression between the two treatments, as nested ShHT2 tests of equal group means revealed at p=0.01. Next, for purposes of identifying potentially important pathways, HT2 tests of zero means at a FDR of 50% produced a list of 83 genes. Several important pathways were identified by cross-referencing this list with the simple gene ontology gene lists, marking a pathway if it contained four or more genes from the above-mentioned list of 83. The specific details of these findings will appear in a forthcoming paper.

8 DISCUSSION

We have seen from the simulation study that the ShAnF performs slightly better only under its more restrictive set model assumptions. In all other cases the ShHT2 performs as well when expected group variances are the same and better when the expected group variances differ. The main message is that both variance stablized statistics perform substantially better than their text-book counterparts, the AnF and the HT2 statistics. This gain in performance has been demonstrated both from the standpoint of statistical inference based upon the B-H FDR procedure, as well as from the standpoint of reliability in the ordering itself. As we learned from simulations generated from the MVN/mxIW distributions, the shrinkage estimation procedure and statistic are fairly robust to lack of model assumptions. For these reasons and due to the broad scope of applicability of shrinkage estimation (Xu and Izmirlian., 2005), and the robustness of ANOVA type tests, it can be argued that variance stabilization should always be performed in small sample micro-array studies. Moreover in the interest of reliability of findings, we advocate the use of the ShHT2 test within the B-H FDR procedure for gene selection starting with otherwise unfiltered genelist.

REFERENCES

K. A. Baggerly, J. S. Morris, S. R. Edmonson, and K. R. Coombes. Signal in noise: evaluating reported reproducibility of serum proteomics tests for ovarian cancer. *J Natl Cancer Inst*,97:307–309, 2005.

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*, 17:509–519, 2001.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J Royal Statist Soc B*, 57:289–300, 1995.

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*, 6(1):59–75, 2005.

S. Dudoit, Y.-H. Yang, M. J. Callow, and T. P. Speed. Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. Technical Report 578, UCB statistics, 2000.

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*, 5(10):R80, 2004.

N. Hayashi, N. Yokoyama, T. Seki, Y. Azuma, T. Ohba, and T. Nishimoto. RanBP1, a Ras-like nuclear G protein binding to Ran/TC4, inhibits RCC1 via Ran/TC4. *Mol Gen Genet.*, 247(6):661–669, 1995.

C. Li and W. H. Wong. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *PNAS*, 98(1):31–36, 2001.

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(1):Article 7, 2005.

I. Lönnstedt and Speed T. Replicated microarray data. *Statistica Sinica*, 12:31–46, 2002.

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

R. J. Muirhead. *Aspects of Multivariate Statistical Theory*. John Wiley & Sons, Hoboken, NJ, 1982.

R. Nishigaki, M. Osaki, M. Hiratsuka, T. Toda, K. Murakami, K.-T. Jeang, H. Ito, T. Inoue, and M. Oshimura. Proteomic identification of differentially-expressed genes in human gastric carcinomas. *Proteomics*, 5(12):3205p–13, 2005.

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.

D. F. Ransohoff. Lessons from Controversy: Ovarian Cancer Screening and Serum Proteomics. *J Natl Cancer Inst*,97: 315–319, 2005.

Y. Takahashi, J. A. Lavigne, S. D. Hursting, V. R. G. Chandramouli, S. N. Perkins, J. C. Barrett, and T. T. Y. Wang. Using DNA microarray analyses to elucidate the effects of genistein in androgen-responsive prostate cancer cells: Identification of novel targets. *MOL CARCINOGEN*, 41(2):108–119, 2004.

V. G. Tusher, R. Tibshirani, and G Chu. Significance analysis of microarrays applied to the ionizing radiation response. *PNAS*, 98:5116–5121, 2001.

E. Weiss-Messer, O. Merom, A. Adi, R. Karry, M. Bidosee, R. Ber, A. Kaploun, A. Stein, and R. J. Barkey. Growth hormone (GH) receptors in prostate cancer: gene expression in human tissues and cell lines and characterization, GH signaling and androgen receptor regulation in LNCaP cells. *Mol Cell Endocrinol*, 220(1-2):109–123, 2004.

P. Wong, D. Taillefer, J. Lakins, J. Pineault, G. Chader, and M. Tenniswood. Molecular characterization of human TRPM-2/clusterin, a gene associated with sperm maturation, apoptosis and neurodegeneration. *Eur J Biochem*, 221(3):917-25, 2004.

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

J.-L. Xu, and G. Izmirlian. Estimation of location parameters for spherically symmetric distributions. *J Multivar Ann*, In press, 2005.

APPENDIX PROOFS OF THEOREMS

Proof of theorem 1: All square root matrices used in the following are considered to be the symetric square root. That is, if A is a symetric matrix with spectral decomposition $A = QDQ'$, then the symetric square root of A is $A^{\frac{1}{2}} = QD^{\frac{1}{2}}Q'$, and $A^{-\frac{1}{2}}$ is the symetric square root of A^{-1}. With clarity in syntax in mind, here we suppress the dependence upon gene, g. First, we rewrite T^2 as follows:

$$T^2 = n\bar{Z}'V^{-1}\bar{Z}$$

where $\bar{Z} = \mathfrak{V}^{-\frac{1}{2}}\bar{Y}$ and

$$V = V_I + V_{II} = (n-1)\mathfrak{V}^{-\frac{1}{2}}S\mathfrak{V}^{-\frac{1}{2}} + \mathfrak{V}^{-\frac{1}{2}}\Lambda\mathfrak{V}^{-\frac{1}{2}}.$$

We make the following observations.

- Because, the distribution of $(n-1)S$ given $\mathfrak{V}$ is $\text{Wishart}_d(n-1, \mathfrak{V})$ it follows that V_I has the $\text{Wishart}_d(n-1, I_d)$ distribution and is independent of $\mathfrak{V}$.
- Next, since $\mathfrak{V}^{-1}$ has the $\text{Wishart}_d(\nu - d - 1, \Lambda^{-1})$ distribution, it follows that $\widetilde{V}_{II} = \Lambda^{\frac{1}{2}}\mathfrak{V}^{-1}\Lambda^{\frac{1}{2}}$ has the $\text{Wishart}_d(\nu - d - 1, I_d)$ distribution.
- Now, because both V_{II} and $\widetilde{V}_{II}$ are symetric and positive definite, it follows from theorem A9.9 of Muirhead, 1982 that they are orthogonal similarity transformations of oneanonther.
- Since $\widetilde{V}_{II}$ is Wishart with identity rate parameter matrix, it is spherically symetric and therefore, invariant to orthogonal transformations. Hence, $V_{II} \overset{\mathbf{D}}{=} \widetilde{V}_{II}$.

Therefore the sum $V = V_I + V_{II}$ has the $\text{Wishart}_d(\nu + n - d - 2, I_d)$ distribution. Next, we rewrite T^2 as

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

Notice that since $\bar{Z}$ has been rescaled, it is independant of $\mathfrak{V}$. Next, because $\bar{Z}$ is the sample mean, it is independant of the sample covariance matrix, S. Thus $\bar{Z}$ and V are independant. Next, it follows from theorem 3.2.12 of Muirhead, 1982, the denominator is distributed $\chi^2_{\nu+n-2d-1}$ and independent of the $\bar{Z}$. Because the numerator is χ^2_d, it follows that $T^2\ \frac{\nu+n-2d-1}{d}$ has the $F_{d,\nu+n-2d-1}$ distribution.

Table 1. Parameters used in Simulation Study

	Norm/IG	MVN/IW		MVN/mxIW	
		a	b	a	b
shape_1	1.9360	9.1020	8.75800	18.407	17.0330
shape_2				6.7750	6.68900
$\text{rate}_{1,1}$	0.0402	0.1280	0.15700	0.1280	0.15700
$\text{rate}_{1,2}$		0.0846	-0.00845	0.0846	-0.00845
$\text{rate}_{2,2}$		0.1240	0.05150	0.1240	0.05150
mixing				0.2000	0.20000
avg. var.	0.0430	0.0405	0.03790	0.0405	0.03790

Items not applicable are left blank

Table 2. eTDR's and eFDR's for 4 statistics–Norm/IG noise

	ShHT2		HT2		ShAnF		AnF	
FDR	eTDR	eFDR	eTDR	eFDR	eTDR	eFDR	eTDR	eFDR
0.01	0.921	0.040	0.000	0.000	0.769	0.000	0.000	0.000
0.05	0.979	0.134	0.000	0.000	0.953	0.001	0.000	0.000
0.10	0.988	0.229	0.000	0.000	0.994	0.028	0.000	0.000
0.15	0.992	0.312	0.000	0.000	1.000	0.089	0.000	0.000
0.20	0.994	0.386	0.001	0.000	1.000	0.161	0.000	0.000
0.25	0.995	0.452	0.003	0.000	1.000	0.242	0.000	0.000

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

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

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

$$g(\mathfrak{Sigma}) = \Gamma_d\left(\frac{\nu-d-1}{2}\right)^{-1} \frac{|\Lambda|^{(\nu-d-1)/2}}{2^{d(\nu-d-1)/2}\left|\mathfrak{Sigma}\right|^{\nu/2}} \text{etr}\left(-\frac{1}{2}\mathfrak{Sigma}^{-1}\Lambda\right)$$

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

$$\begin{aligned} f(A)g(\mathfrak{Sigma}) &= \frac{|\Lambda+A|^{(\nu+n-d-2)/2}\,\text{etr}\left(-\frac{1}{2}\mathfrak{Sigma}^{-1}(\Lambda+A)\right)}{\Gamma_d\left(\frac{\nu+n-d-2}{2}\right)2^{d(\nu+n-d-2)/2}\left|\mathfrak{Sigma}\right|^{(\nu+n-1)/2}} \\ &\quad \frac{\Gamma_d\left(\frac{\nu+n-d-2}{2}\right)}{\Gamma_d\left(\frac{n-1}{2}\right)\Gamma_d\left(\frac{\nu-d-1}{2}\right)} \frac{|\Lambda|^{(\nu-d-1)/2}\,|A|^{(n-d-2)/2}}{|\Lambda+A|^{(\nu+n-d-2)/2}}. \end{aligned}$$

Thus, the posterior distribution of $\mathfrak{Sigma}$ 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.

Table 3. eTDR's and eFDR's for 4 statistics–MVN/IW-a noise

	ShHT2		HT2		ShAnF		AnF	
FDR	eTDR	eFDR	eTDR	eFDR	eTDR	eFDR	eTDR	eFDR
0.01	0.546	0.009	0.000	0.000	0.859	0.000	0.000	0.000
0.05	0.928	0.048	0.000	0.000	1.000	0.138	0.000	0.000
0.10	0.964	0.094	0.000	0.000	1.000	0.375	0.000	0.000
0.15	0.977	0.142	0.001	0.000	1.000	0.591	0.000	0.000
0.20	0.984	0.193	0.001	0.000	1.000	0.761	0.000	0.000
0.25	0.987	0.240	0.002	0.000	1.000	0.867	0.000	0.000

Table 4. eTDR's and eFDR's for 4 statistics–MVN/IW-b noise

	ShHT2		HT2		ShAnF		AnF	
FDR	eTDR	eFDR	eTDR	eFDR	eTDR	eFDR	eTDR	eFDR
0.01	0.937	0.008	0.000	0.000	0.918	0.001	0.000	0.000
0.05	0.985	0.042	0.000	0.000	1.000	0.171	0.000	0.000
0.10	0.993	0.090	0.000	0.000	1.000	0.400	0.000	0.000
0.15	0.995	0.139	0.001	0.000	1.000	0.603	0.000	0.000
0.20	0.996	0.189	0.001	0.000	1.000	0.765	0.000	0.000
0.25	0.997	0.241	0.002	0.000	1.000	0.866	0.000	0.000

Table 5. eTDR's and eFDR's for 4 statistics–MVN/mxIW-a noise

	ShHT2		HT2		ShAnF		AnF	
FDR	eTDR	eFDR	eTDR	eFDR	eTDR	eFDR	eTDR	eFDR
0.01	0.767	0.033	0.000	0.000	1.000	0.162	0.194	0.000
0.05	0.944	0.127	0.000	0.000	1.000	0.497	0.722	0.000
0.10	0.967	0.225	0.000	0.000	1.000	0.701	0.931	0.007
0.15	0.977	0.308	0.001	0.000	1.000	0.815	0.995	0.085
0.20	0.982	0.382	0.002	0.000	1.000	0.878	1.000	0.189
0.25	0.985	0.445	0.003	0.000	1.000	0.914	1.000	0.280

Table 6. eTDR's and eFDR's for 4 statistics–MVN/mxIW-b noise

	ShHT2		HT2		ShAnF		AnF	
FDR	eTDR	eFDR	eTDR	eFDR	eTDR	eFDR	eTDR	eFDR
0.01	0.940	0.029	0.000	0.000	1.000	0.164	0.153	0.000
0.05	0.981	0.119	0.000	0.000	1.000	0.475	0.709	0.000
0.10	0.989	0.215	0.000	0.000	1.000	0.669	0.902	0.002
0.15	0.993	0.297	0.001	0.000	1.000	0.787	0.983	0.041
0.20	0.994	0.365	0.002	0.000	1.000	0.859	0.999	0.121
0.25	0.995	0.428	0.003	0.000	1.000	0.903	1.000	0.199

Table 7. Genes from Wang Arrays

CLU–clusterin (complement lysis inhib)
RANBP1–RAN binding protein 1
FN14–type I transmembrane protein Fn1
ECH1–enoyl Coenzyme A hydratase 1, pe
GHR–growth hormone receptor
SELENBP1–selenium binding protein 1
CRYAA–crystallin, alpha A
ESTs.621
ALDOC–aldolase C, fructose-bisphospha
H.sapiens HCG II mRNA
Homo sapiens mRNA; cDNA DKFZp434I0812
FMOD–fibromodulin

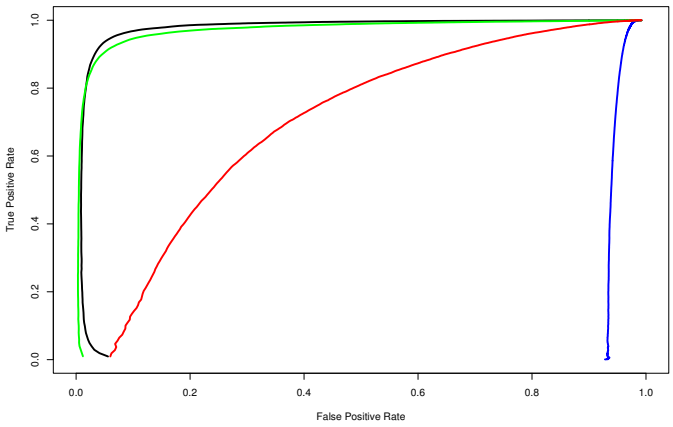

Fig. 2. Performance of four statistics using data simulated from the multivariate normal/Inverse Wishart distribution having equal group variances but with correlated group errors. The plot shows the empirical true positive rate versus the empirical false positive rate for each of the four benchmarked statistics. ShHT2=black, ShAnF=green, AnF=red, HT2=blue

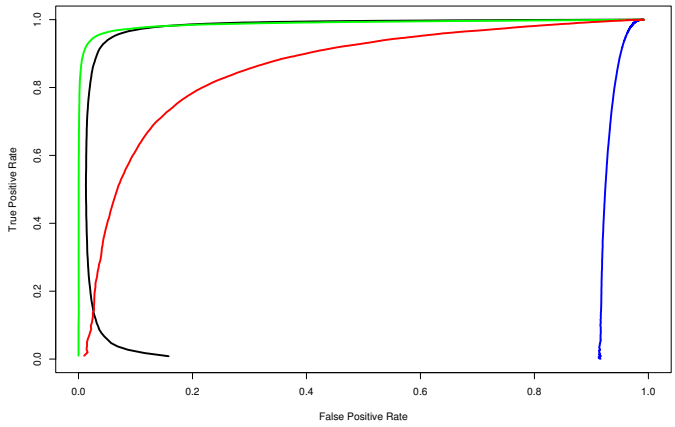

Fig. 1. Performance of four statistics using data simulated from the normal/Inverse gamma distribution. The plot shows the empirical true positive rate versus the empirical false positive rate for each of the four benchmarked statistics as cutoff ranges through all possible values of the statistic. ShHT2=black, ShAnF=green, AnF=red, HT2=blue

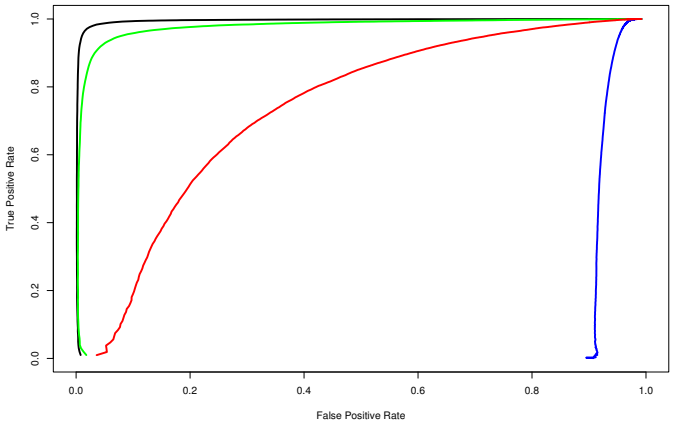

Fig. 3. Performance of four statistics using data simulated from the multivariate normal/Inverse Wishart distribution having uncorrelated group errors but with unequal group variances. The plot shows the empirical true positive rate versus the empirical false positive rate for each of the four benchmarked statistics. ShHT2=black, ShAnF=green, AnF=red, HT2=blue

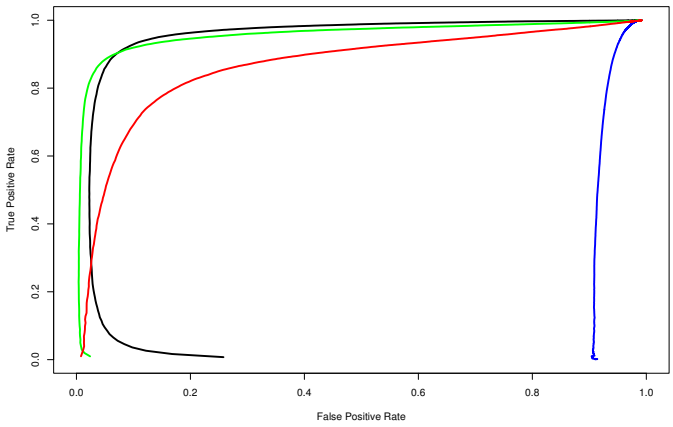

Fig. 4. Performance of four statistics using data simulated from the multivariate normal/mixed Inverse Wishart distribution having equal group variances but with correlated group errors. The plot shows the empirical true positive rate versus the empirical false positive rate for each of the four benchmarked statistics. ShHT2=black, ShAnF=green, AnF=red, HT2=blue

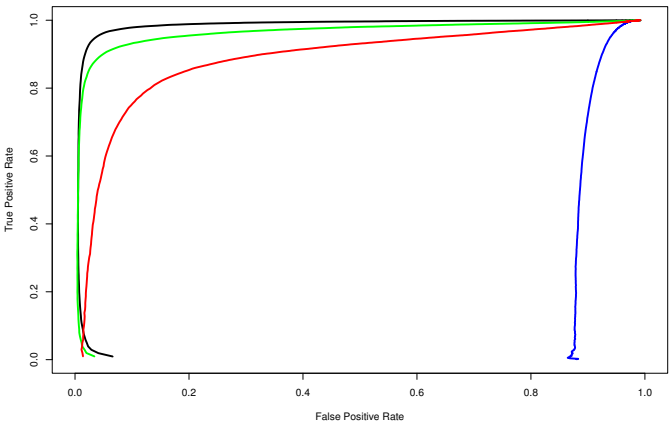

Fig. 5. Performance of four statistics using data simulated from the multivariate normal/mixed Inverse Wishart distribution having uncorrelated group errors but with unequal group variances. The plot shows the empirical true positive rate versus the empirical false positive rate for each of the four benchmarked statistics. ShHT2=black, ShAnF=green, AnF=red, HT2=blue