Shrinkage concept in microarray study

Q. Dat Nguyen

Institut für Informatik, Martin-Luther-Universität Halle-Wittenberg

Supervisor: Prof. Dr. Ivo Grosse

April 29, 2011

Contents

1. Introduction

DNA microarrays since its first introduction has become a very promising biotechnology which is one of the most powerful and widely used technology. Microarrays were being applied almost everywhere in life sciences, both in molecular biology and in medicine to address a wide range of problems, from bacterial infections to the classification of tumors (Boldrick et al., 2002; Golub et al., 1999), even from the detection of alternative splicing to the hybridization of comparative genomes.
With the power of measuring the mRNA abundance at the genomic level, the identifying of differential expressed genes become the most interested question in microarray data analysis, that is, the question which genes have the expression levels are changed during important biological processes (e.g. cellular replication (time), treatment/control cell type, dose of a drug), or varied across collections of related samples (e.g. tumor samples from patients with cancer, from patients after a specify drug treatment).
The biological question of which gene is differential expressed can be formulated as the statistical test of the null hypothesis: there's no association between the expression levels of the gene and the responses.
It seems simple with many powerful techniques from statistic, e.g. Student's t-test, but many problems arisen from the fact that DNA microarrays experiments deliver ton of informations. One of the problems is the multiple testing problem was considered in (Dudoit et al., 2003). We consider in this paper another problem, well, let cast aside the multiple problem, we assume that most of classical statistical tests perform well to identify the genes, assumed that we have enough samples, in other words, a huge dataset is what we need to perform those tasks. At the first sight, all of microarray datas are huge, but unfortunately, it is huge because of the large number of genes represented in their array (The human genome has approximately about 23.000 genes, most of model organisms was used in research like Arabidopsis thaliana has around 25.000 genes, some higher plants have even more genes, the fewest lies around at thousand genes, but it's still a huge number).
Typically because of the costly production of microarray, for each gene the number of RNA samples assayed is small. Therefore, the commonly used approach of treating the differential expression of one gene at a time as a solely population often has low power. The Student's t-test have been shown with very low power when the samples size falls under 10. Some could consider the assumption that all genes have the same variance (to stabilize the variance estimator, but in the practice often not this case) to increase the power of detection but also increase the risk of generating false detection if the assumption is false. More complicated, the measured datas are often non-normally distributed and have non-identical and dependent distributions between genes.
Because of those problems, in the last few years various advanced approaches have been suggested, from modifying the simple fold change to the classical ordinary t statistic. The idea of modifying the classical fold change method (Kadota et al., 2008) bases simply on empirical observation from real data, although it doesn't base on any statistical concept, surprisingly performs well with real datas. The modification of Student's t-test based mostly on modifying the estimators of variance. The widely used SAM t-test (Tusher et al., 2001) for example adds a small constant to the gene-specific variance estimator to stabilize the small variance. Another used the empirical Bayesian approaches to have a more stabilized estimator of the

variance. One important point to mention here is, the parallel nature of the inference in microarrays (variation because of technical noises) and in biology (genes in the same tissues tend to have same expression profile) allows some possibilities for borrowing information from the ensemble of genes which can assist in inference about each gene individually. That "prior" knowledge about microarray should be considered in any Bayesian approaches.
The advantage of Bayesian approaches is, because they allow the information sharing between genes, which is essential when the number of sample is small. The disadvantage of the Bayesian approaches is that they could become quite computationally and analytically expensive, even if the assumption of normally distributed for the data is true. One of the approaches has been suggest recently was from (Opgen-Rhein and Strimmer, 2007). This "shrinkage" approach was shown to be simple as the SAM test but performs well like any another full Bayesian approaches, yet even could be derived fully analytic, requires no computer-intensive procedures, and makes no prior assumptions about the distribution of data.
We know that most of microarrays experiments focus on the two factorial question, means just two responses are compared, e.g. one need to compare tumor cell samples with tumor cell after a drug treatment samples to detect which genes are differential expressed after the treatment. Sometimes but more than two factors are concerned, e.g. the drug treatment after one specific time interval (24h, 48h), or the same tumor cell but with various drugs. In this case the simple fold change fails, but what happens with the well-known t-test? The t-test appears at the first look just an extent of fold change with the consideration of the variance. Fortunately, another well-known extension of the Student t-test, F-test or ANOVA (**AN**alysis **O**f **VA**riance) can handle any desirable factorial questions.

The purpose of this paper is further examination of the "good" shrinkage approach when applied in practice. First we shall give the readers a fundamental understanding in microarrays technology, which quite useful for readers who are new to microarray. Then a briefly review of Student t-test and understand the original shrinkage idea of James & Stein, which crucial to understand our work in this paper. Third is our simulation setup to test the theory of the shrinkage approach in the practice.

2. Background

2.1. Microarray

First are some backgrounds on microarrays in general. According to the knowledge nowadays we all know that all the living organisms are composed of one or more cells, and cells are basic units of structure and function in an organism. Cells store their hereditary information (genome) in the form of chromosomes, which are large pieces of DNA containing hundreds of thousands genes, each gene specifies the composition and structure of a protein. All living organisms are composed largely of proteins, which are polymers (or a chain) of amino acids, are the workhouse molecules of the cell, for example, form part of a cellular structure itself, catalyze almost all the biochemical reactions in cell in form of enzymes, control the activity of other proteins, produce energy and important biomolecules like DNA and proteins. The information contained in DNA is first duplicated via the replication process, then transcribed in messenger RNA (mRNA). The mRNA is eventually processed in eukaryotic cells, then

translated to synthesize new proteins.
Every cell in an organism has nearly the same set of chromosomes, and thus contains the same supply of proteins. However cells do have remarkably distinct properties, for example liver cells, skin cells or heart cells. Those distinctions of cells are the result of differences in the abundance, distribution and state of the cell proteins. The changes in protein abundance are determined in part by changes in the levels of mRNA, speaks, which protein is synthesized in different state of cell, is dependent on that gene which encoded for the protein, is expressed. Thus there is a connection between the state of cell with his mRNA quantity.
DNA microarrays are high-throughput biological assays that can quickly and efficiently measure DNA or mRNA abundance in cells for thousands of genes simultaneously. There are two major kinds of microarrays, an oligonucleotide array which mostly produced from Affymetrix (Wikipedia, 2010a), and a spotted cDNA microarray, which are very different in designs and need to be corresponding processed in the next steps before receiving the needing expression values. The details on different microarray designs will not be mentioned here but rather refer to the readers to extending sources (Wikipedia, 2010b).
After all of those preprocessing, normalization and summarization steps, from either type of microarray we obtain several thousand expression values, one or many for each gene. With the given values, we need to identify those genes which demonstrate a significant change in expression level under the impact of certain experimental conditions, such as the cancerous tumors. Those genes, which are differentially expressed in one set of samples relative to another, could be used to understand the potentially meaningful correlations between genes and specific conditions (e.g. which genes are expressed in tumor issues but not in normal state). Moreover, genes which are not differential expressed could be filtered out to reduce the effort of further details analysis. Although simple in principle, the identification of differential expressed genes could be complex in practice, as many problems arise as we already mentioned in the introduction part.

2.2. Student's t-test

To understand the reason of using the Student's t-test in differential expression data analysis, it's best to have a simple example of a typical microarray experiment. The example below was taken from (Zhang, 2007) to show the readers a real application of microarrays.
Suppose we are interested in identifying which genes are involved in multiple sclerosis diseases and after the drug treatment. The given data are samples taken from 14 multiple sclerosis patients, contain the expression levels of 4,132 genes for each patient prior to and 24 hours after interferon-β treatment. We have two groups of samples with each group contains 14 samples, the first group indicate the samples were taken before the treatment and the second group is corresponding to post-treatment condition.

How could we answer the interested question of differential expressed genes? Let observe one gene, for example gene i from the given data. We should have for gene i two groups of samples, within each group are 14 samples. If gene i was involved in some pathways of multiple sclerosis and was required to activate in the drug treatment condition, the RNA-$Polymerase$ binds at the specific transcription factor and gene i was hereby transcribed. Therefore the mRNA amount specifies for gene i was changed between two conditions, prior

and after the drug treatment. Typically, the first attempt to analyze differentially-expressed genes simply calculated the average expression over all samples in each group, then divided the two average expression values. A fixed cut-off k was chosen and if gene i have the quotient of average expression values over k, then declare that gene i is differential expressed. This so-called fold change method is very simple, just considering those genes which demonstrate a significant change between the experiment samples of particular interest (in this example was the interferon-β treatment for multiple sclerosis) and control. A k threshold was typically chosen as a two or three-fold change, dependent on experiments, and often arbitrary and did not take into account the overall distribution of the samples.

Definition Let's denote $\overline{x}_1$ the average expression values from the control group and $\overline{x}_2$ the average expression values from the treatment group. The fold change ratio is defined as $\frac{\overline{x}_1}{\overline{x}_2}$.

In microarrays data analysis, where the measurements mostly needed to transform to have the nearly same property of a normal distribution, another widely used variant is using logarithm transformation of the data, then using the average log fold change difference.

Let's denote $\overline{y}_1$ the average *log* expression values from the control group and $\overline{y}_2$ the average *log* expression values from the treatment group. The average fold change difference is then defined as $\overline{y}_1 - \overline{y}_2$.

It worths to mention that the two variants of fold change (fold change ratio and average log fold change) are sometimes misused as exchangeable in microarray analysis, but they have difference in their mathematical definition, and also differ in results as well. Nevertheless their difference in definition is small and the difference in results is also not significant as well. An excellent work of Tibshirani (2007) concerns about this problem and was highly recommended. Someone with statistical knowledge should already know those problems were yet considered in statistic. The Student's t-test is a standard statistical hypothesis test for detecting significant change of a variable between repeated measurements in two groups. The test follows a Student's t distribution if the null hypothesis is supported. It is most commonly applied when the test statistic would follow a normal distribution if the value of a scaling term in the test statistic were known. When the scaling term is unknown and is replaced by an estimate based on the data, the test statistic (under certain conditions) follows a Student's t distribution. To understand why the t-test is more sophisticated than the simply fold-change method, we begin with the derivation of t-test for the most simple case, the one-sample t-test.

2.2.1. *One-sample t-test*

Definition Suppose we have samples from a normally distributed population. We are interested in testing the null hypothesis that the population mean is equal to a specified value μ_0. Let's denote N samples of the population as $x_1, x_2, ..., x_N$.
The assumption is that x_i independent normally distributed with mean μ and variance σ^2 for all i: $x_i \sim \mathcal{N}(\mu, \sigma^2)\, \forall i$.

The sample mean and the unbiased sample variance of the population are defined as follow:

$$\overline{\mu} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

$$\hat{\sigma}^2_{N-1} = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \overline{\mu})^2$$

The t-statistic is defined as:

$$t = \frac{\overline{\mu} - \mu_0}{\hat{\sigma}^2_{N-1}}\sqrt{N} \tag{1}$$

If the null hypothesis is true ($H_0 : \mu = \mu_0$), then the t-statistic is t-distributed with (N-1) degrees of freedom.

Proof. As already shown in Appendix D, the sample mean has the following distribution (Equation 45)

$$\Rightarrow \overline{\mu} \sim \mathcal{N}(\mu, \frac{\sigma^2}{N})$$

$$\Rightarrow \frac{(\overline{\mu} - \mu)\sqrt{N}}{\sigma} \sim \mathcal{N}(0, 1)$$

Analog with the unbiased sample variance (Equation 49)

$$(N-1)\frac{\hat{\sigma}^2_{N-1}}{\sigma^2} \sim \chi^2_{N-1}$$

Let Z $= \frac{(\overline{\mu}-\mu)\sqrt{N}}{\sigma} \sim \mathcal{N}(0,1)$
Let V $= (N-1)\frac{\hat{\sigma}^2_{N-1}}{\sigma^2} \sim \chi^2_{N-1}$
Let

$$t = \frac{Z}{\sqrt{V/(N-1)}} = \frac{\frac{(\overline{\mu}-\mu)\sqrt{N}}{\sigma}}{\sqrt{(N-1)\frac{\hat{\sigma}^2_{N-1}}{\sigma^2}/(N-1)}} = \frac{\overline{\mu} - \mu}{\hat{\sigma}^2_{N-1}}\sqrt{N}$$

If the null hypothesis is true ($H_0 : \mu = \mu_0$), we see that is the same t from Equation 1, and this statistic is t-distributed with (N-1) degrees of freedom. (Equation 41) □

This one-sample t-test could be used for paired data, e.g. the data set from the above example, which has *biological replicates*, that means each data points has a pair of measurements, one prior to (control) and one after the treatment, The pair of measurements combine to a single log ratio. In this case, the observed measurements for each gene form one vector of log ratios, and the paired t-test is reduced to a one-sample t-test. Here, the null hypothesis that the gene is not differential-expressed is the same as that the mean of the log ratios μ equals to 0, denoted by $H_0 : \mu = 0$. The t-statistic is calculated as follow:

$$t = \frac{\overline{\mu}}{\hat{\sigma}_{N-1}}\sqrt{N}$$

where $\overline{\mu}$ is the average of the log ratios, $\hat{\sigma}_{N-1}$ is the (unbiased) standard deviation from the log ratios, and N is the number of samples.
A p-value and a t-value can then be obtained by looking up a t-distribution with (N-1) degrees of freedom, and the null-hypothesis is rejected if the t-statistic is larger then the obtained t-value, based on the p-value.

2.2.2. *Two-sample t-test*

The two-sample t-test tests the null hypothesis that the means of two normally distributed populations are equal. As two populations show up, two-sample t-test may has *equal-variance* and *unequal-variance*. All such tests are usually called Student's t-test, though *strictly* speaking the name Student's t-test originally was used if the variances of the two populations are also assumed to be equal; and the form of the test used when the variance is unequal is sometimes called *Welch's t-test*. These tests are often referred to as "unpaired" or "independent samples" t-tests, as they are typically applied when the statistical units underlying the two samples being compared are non-overlapping. Since *Welch's t-test* is a generalization of Student's t-test, we'll introduce this test here.

Definition The t-statistic for all two-sample t-test could be calculated as follow:

$$t = \frac{\overline{\mu}_1 - \overline{\mu}_2}{\sqrt{\frac{\hat{\sigma_1}^2}{N_1} + \frac{\hat{\sigma_2}^2}{N_2}}} \tag{2}$$

where analog in the one-sample case, $\overline{\mu}_1$ and $\overline{\mu}_2$ are the means, $\hat{\sigma_1}^2$ and $\hat{\sigma_2}^2$ are the variances, and N_1 and N_2 are the samples sizes of two populations, respectively.
This t-statistic is then t-distributed with df degrees of freedom. In the case of unequal variance, the *degrees of freedom* need to be adjusted as:

$$df = \frac{\left(\frac{\hat{\sigma_1}^2}{N_1} + \frac{\hat{\sigma_2}^2}{N_2}\right)^2}{\frac{(\frac{\hat{\sigma_1}^2}{N_1})^2}{N_1 - 1} + \frac{(\frac{\hat{\sigma_2}^2}{N_2})^2}{N_2 - 1}} \tag{3}$$

Note that the value of df is usually not an integer and need to be truncated.
The derivation of the two-sample t-test will not be shown in this paper.

In the most cases of microarray data analysis, the samples sizes is typically very small ($N_1, N_2 < 5$), then the *pooled sample variance* $\hat{\sigma_p}^2$ is often used to estimate the sample variability instead of the variances for each group. The Equation 2 reduces to:

$$t = \frac{\overline{\mu}_1 - \overline{\mu}_2}{\sqrt{\hat{\sigma_p}^2 \left(\frac{1}{N_1} + \frac{1}{N_2}\right)}} \tag{4}$$

where

$$\hat{\sigma_p}^2 = \frac{(N_1 - 1)\hat{\sigma_1}^2 + (N_2 - 1)\hat{\sigma_2}^2}{N_1 + N_2 - 2} \tag{5}$$

This t-statistic from Equation 4 is then t-distributed with $df = (N_1 + N_2 - 2)$ degrees of freedom.
In the case of the same sample sizes of two groups ($N_1 = N_2 = N$), the t-statistic reduces to:

$$t = \frac{\overline{\mu}_1 - \overline{\mu}_2}{\hat{\sigma_p}^2}\sqrt{N} \tag{6}$$

As we can see in comparing the above equation of the simplified t-statistic with the average log fold change difference, the two definitions distinguish only in the denominator and a constant term ($\sqrt{N}$). The t-test is hence more sophisticated than the fold change method. The significance of the differentially expressed genes depends not only on the average difference but also on both the population variability and the number of individuals in the populations. In general, the accuracy of the successful determination increases with the number of samples.

2.3. ANalysis Of VAriance

2.3.1. *2-groups One-Way ANOVA*

There are so many introductions and motivations for ANOVA in the classic mathematical literature, in this work we'll try to introduce the $One-Way$ ANOVA approach as an extension of the standard t-test. As we already understand the standard t-test, it'll be easier to see the $One-Way$ ANOVA is simply an extension of t-test when we have more than two groups (or $levels$ in the ANOVA language). Another requirement of ANOVA which discriminative from t-test is in ANOVA is assumed that the variances are the same for all groups, which is often not the case in microarray analysis. But with some heuristic improvement, the ANOVA approach is still very useful in the praxis even without the same variances assumption. Besides that, the unequal variance problem in ANOVA approach was considered as a not trivial problem and it'll be not covered in this paper.
Let's first review the two-samples t-test case to see which distribution has the $squared$ t-statistic. Let's denote x_{ij} the j-te measurement of i-te group. Group 1 and 2 has N_1 and N_2 samples respectively. We assume that those measurements of two groups come from a normal distribution with mean μ_1, μ_2 and same variance σ^2. The tested null hypothesis is $H_0 : \mu_1 = \mu_2$. From Equation 2 we have

$$t = \frac{\overline{\mu}_1 - \overline{\mu}_2}{\sqrt{\frac{\hat{\sigma_1}^2}{N_1} + \frac{\hat{\sigma_2}^2}{N_2}}} \tag{7}$$

with $\hat{\sigma_1}^2$ and $\hat{\sigma_2}^2$ are the $unbiased$ sample variance of each group

$$\hat{\sigma_1}^2 = \frac{\sum_{j=1}^{N_1}(x_{1j} - \overline{\mu}_1)^2}{N_1 - 1}$$
$$\hat{\sigma_2}^2 = \frac{\sum_{j=1}^{N_2}(x_{2j} - \overline{\mu}_2)^2}{N_2 - 1}$$

To ensure the same variances assumption, we define the *pooled sample variance* $\hat{\sigma_p}^2$

$$\hat{\sigma_p}^2 = \frac{(N_1 - 1)\hat{\sigma_1}^2 + (N_2 - 1)\hat{\sigma_2}^2}{N_1 - 1 + N_2 - 1} \tag{8}$$

The t-statistic becomes

$$t = \frac{\overline{\mu}_1 - \overline{\mu}_2}{\sqrt{\hat{\sigma_p}^2 \left(\frac{1}{N_1} + \frac{1}{N_2}\right)}} \tag{9}$$

We're interested in the squared t-statistic, which provide

$$\begin{aligned}
\Rightarrow t^2 &= \frac{(\overline{\mu}_1 - \overline{\mu}_2)^2}{\hat{\sigma_p}^2 \left(\frac{1}{N_1} + \frac{1}{N_2}\right)} \\
&= \frac{\frac{(\overline{\mu}_1 - \overline{\mu}_2)^2}{\sigma^2 \left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}{\frac{\hat{\sigma_p}^2}{\sigma^2}} \\
&= \frac{A}{B}
\end{aligned}$$

The next step is to see which distribution has A and B. With A we have

$$\begin{aligned}
E(\overline{\mu}_1 - \overline{\mu}_2) &= \mu_1 - \mu_2 = 0 \\
Var(\overline{\mu}_1 - \overline{\mu}_2) &= Var(\overline{\mu}_1) + Var(\overline{\mu}_2) = \frac{\sigma^2}{N_1} + \frac{\sigma^2}{N_2} \\
\Rightarrow (\overline{\mu}_1 - \overline{\mu}_2) &\sim \mathcal{N}(0, \sigma^2(\frac{1}{N_1} + \frac{1}{N_2})) \\
\Rightarrow \sqrt{A} = \frac{(\overline{\mu}_1 - \overline{\mu}_2)}{\sqrt{(\frac{\sigma^2}{N_1} + \frac{\sigma^2}{N_2})}} &\sim \mathcal{N}(0, 1) \\
\Rightarrow A &\sim \chi_1^2
\end{aligned}$$

With B

$$\begin{aligned}
B = \frac{\hat{\sigma_p}^2}{\sigma^2} &= \frac{\frac{(N_1-1)\hat{\sigma_1}^2 + (N_2-1)\hat{\sigma_2}^2}{N_1-1+N_2-1}}{\sigma^2} \\
&= \frac{(N_1 - 1)\frac{\hat{\sigma_1}^2}{\sigma^2} + (N_2 - 1)\frac{\hat{\sigma_2}^2}{\sigma^2}}{N_1 - 1 + N_2 - 1}
\end{aligned}$$

Since already known that $(N_1 - 1)\frac{\hat{\sigma_1}^2}{\sigma^2} \sim \chi^2_{N_1-1}$, $(N_2 - 1)\frac{\hat{\sigma_2}^2}{\sigma^2} \sim \chi^2_{N_2-1}$ (Equation 49)

$$\begin{aligned}
&\Rightarrow (N_1 - 1)\frac{\hat{\sigma_1}^2}{\sigma^2} + (N_2 - 1)\frac{\hat{\sigma_2}^2}{\sigma^2} \sim \chi^2_{N_1+N_2-2} \text{ (Equation 40)} \\
&\Rightarrow \frac{\hat{\sigma_p}^2}{\sigma^2} \sim \chi^2_{N_1+N_2-2}/(N_1 + N_2 - 2)
\end{aligned}$$

From Equation 42

$$t^2 = \frac{A}{B} = \frac{\chi_1^2/1}{\chi_{N_1+N_2-2}^2/(N_1+N_2-2)} \sim \mathcal{F}_{1,N_1+N_2-2}$$

That's it! This above derivation could be considered as the construction of One-Way ANOVA for two groups. We'll also introduce some new definitions which are very well-known in the ANOVA concept, the *sum of squares* due to treatments (sum of squares "between groups") the sum of squares error (sum of squares "within groups" or "residuals") and their *mean squares*, after all rewrite the above derivation in the "real ANOVA language".
From the individual sample mean of each group $\overline{\mu}_1$, $\overline{\mu}_1$, let's denote the *grand mean* of all groups as follow

$$\overline{\mu} = \frac{\sum_{j=1}^{N_1} x_{1j} + \sum_{j=1}^{N_2} x_{2j}}{N_1+N_2} = \frac{N_1}{N_1+N_2}\overline{\mu}_1 + \frac{N_2}{N_1+N_2}\overline{\mu}_2$$

The sum of squares *due to treatments* (or "between groups") is defined as follow

$$SS_{Treatment} = SST = \frac{(\overline{\mu}_1 - \overline{\mu}_2)^2}{\frac{1}{N_1}+\frac{1}{N_2}} = N_1(\overline{\mu}_1 - \overline{\mu})^2 + N_2(\overline{\mu}_2 - \overline{\mu})^2$$

The mean squares of treatments is defined as the sum of squares divided by the *degrees of freedom* of treatments, which is the number of groups (in this case $= 2$) minus one.

$$MS_{Treatment} = MST = \frac{SS_{Treatment}}{2-1}$$

The sum of squares *error* (or "within groups")

$$SS_{Error} = SSE = (N_1-1)\hat{\sigma_1}^2 + (N_2-2)\hat{\sigma_2}^2 = \sum_{j=1}^{N_1}(x_{1j}-\overline{\mu}_1)^2 + \sum_{j=1}^{N_2}(x_{2j}-\overline{\mu}_2)^2$$

Accordingly the mean squares error

$$MS_{Error} = MSE = \frac{SS_{Error}}{N_1-1+N_2-1}$$

The denominator of the above formula is the *degrees of freedom* of error itself. It's worths to mention that another term, the total sum of squares could be showed as the sum of the above two sums of squares.

$$S_{Total} = SS_{Treatment} + SS_{Error}$$

The F-statistic of the 2-groups ANOVA is defined as

$$F = \frac{MS_{Treatment}}{MS_{Error}} = t^2 \sim \frac{\chi_1^2/1}{\chi_{N_1+N_2-2}^2/(N_1+N_2-2)} \sim \mathcal{F}_{1,N_1+N_2-2}$$

2.3.2. *K-groups One-Way ANOVA*

It's not too hard to extend the One-Way ANOVA approach from two groups to K-groups. With each group i has their individual number of sample N_i itself, we define first the total number of measurements of the experiment

$$N = \sum_{i=1}^{K} N_i \tag{10}$$

The grand mean

$$\overline{\mu} = \frac{\sum_{i=1}^{K}\sum_{j=1}^{N_i} x_{ij}}{\sum_{i=1}^{K} N_i} = \frac{\sum_{i=1}^{K}\sum_{j=1}^{N_i} x_{ij}}{N} \tag{11}$$

Another terms are analog to the two-groups case.

$$SS_{Treatment} = \sum_{i=1}^{K} N_i(\overline{\mu}_i - \overline{\mu})^2 \tag{12}$$

$$MS_{Treatment} = \frac{SS_{Treatment}}{K-1} \tag{13}$$

$$SS_{Error} = \sum_{i=1}^{K}(N_i - 1){\hat{\sigma}_i}^2 \tag{14}$$

$$MS_{Error} = \frac{SS_{Error}}{\sum_{i=1}^{K}(N_i - 1)} = \frac{SS_{Error}}{N-K} \tag{15}$$

Again the formula's still valid

$$S_{Total} = SS_{Treatment} + SS_{Error}$$

Similarly the F-statistic of the K-groups ANOVA is defined as

$$F = \frac{MS_{Treatment}}{MS_{Error}} \sim \frac{\chi^2_{K-1}/K-1}{\chi^2_{N-K}/(N-K)} \sim \mathcal{F}_{K-1,N-K} \tag{16}$$

The true null hypothesis is $H_0 : \mu_1 = \mu_2 = ... = \mu_K$, then the F-statistic is F-distributed with $(K-1)$ and $(N-K)$ degrees of freedom.

3. Shrinkage concept

3.1. James & Stein shrinkage rule

3.1.1. Introduction

We all know the sample variance estimator (biased or unbiased) from a population . This could be even shown that the unbiased sample variance converges toward the real variance if the samples size N is large enough. In microarray analysis, the number of individuals using

in the experiment is mostly limited, because of costs, biological limit of individual, ethics limit... In almost cases, the samples size in microarray experiments is around 3, which is very "dangerous" to use in any statistical framework. The classical estimator in statistic becomes very unstable in this setting.
The original idea of James & Stein rule (Stein, 1955) is to solve the problem of estimating the mean of a P-dimensional multivariate normal distribution from a single (N = 1) vector-valued observation. This is an extreme example of the "small N, large P" setting. In this case the widely used maximum-likelihood estimator equals the vector of observations, and of course unstable. That so-called "Stein phenomenom" was demonstrated by Stein in 1956, even shown in the high-dimensional inference problems it is often possible to improve upon the maximum likelihood estimator. James & Stein proposed a new rule to improve the stability of the maximum-likelihood estimator. Opgen-Rhein and Strimmer (2007) even shown this rule can be use to construct a procedure for improving *any* variance estimator. In the following text we are trying to explain the James and Stein rule from the paper of Opgen-Rhein and Strimmer (2007).

Definition Suppose that we have an unregularized estimator $\hat{\theta}$ from the true parameter θ, e.g. it could be the maximum-likelihood estimator vector from the above example.
The shrinkage estimator is then written as

$$\begin{aligned} \delta^{\lambda} &= \hat{\theta} - \lambda\Delta \\ &= \hat{\theta} - \lambda(\hat{\theta} - \hat{\theta}^{Target}) \\ &= \lambda\hat{\theta}^{Target} + (1-\lambda)\hat{\theta} \end{aligned} \tag{17}$$

In other words, the shrinkage estimator δ^{λ} is the linear combination of the original estimator $\hat{\theta}$ and a target estimator $\hat{\theta}^{Target}$. The parameter λ is the *shrinkage intensity* which determines how those two estimators are "pooled" together. If $\lambda = 1$ then the target dominates completely, and if $\lambda = 0$ no shrinkage occurs.
In James & Stein estimation the search for the optimal λ is considered from a decision theoretic perspective. First, a loss function is selected (e.g. the squared error). Second, λ is chosen such that the corresponding risk (the expectation of the loss with respect to the data) of δ^{λ} is minimized. (e.g. the mean squared error). In our context, λ is estimated from the data regarding to minimize the mean squared error of δ^{λ}.

3.1.2. Derivation of shrinkage rule

Using this rule (Equation 17), the mean squared error function (MSE) of δ^{λ} could be calculated even *without* knowing the true value of θ:

Proof.

$$\begin{aligned}
MSE(\delta^\lambda) &= MSE(\hat{\theta} - \lambda(\hat{\theta} - \hat{\theta}^{Target})) \\
&= E[(\hat{\theta} - \lambda(\hat{\theta} - \hat{\theta}^{Target}) - \theta)^2] \\
&= E[(\hat{\theta} - \theta - \lambda(\hat{\theta} - \hat{\theta}^{Target}))^2] \\
&= E[(\hat{\theta} - \theta)^2 - 2\lambda(\hat{\theta} - \hat{\theta}^{Target})(\hat{\theta} - \theta) + \lambda^2(\hat{\theta} - \hat{\theta}^{Target})^2] \\
&= E(\hat{\theta} - \theta)^2 - 2\lambda E[(\hat{\theta} - \hat{\theta}^{Target})(\hat{\theta} - \theta)] + \lambda^2 E(\hat{\theta} - \hat{\theta}^{Target})^2 \\
&= MSE(\hat{\theta}) - 2\lambda E[(\hat{\theta} - \hat{\theta}^{Target})(\hat{\theta} - \theta)] + \lambda^2 E(\hat{\theta} - \hat{\theta}^{Target})^2
\end{aligned}$$

Examine further the second term without the λ:

$$\begin{aligned}
&E[(\hat{\theta} - \hat{\theta}^{Target})(\hat{\theta} - \theta)] \\
&= E[\hat{\theta}^2 - \hat{\theta}\hat{\theta}^{Target} - \theta(\hat{\theta} - \hat{\theta}^{Target})] \\
&= E(\hat{\theta}^2) - E(\hat{\theta}\hat{\theta}^{Target}) - E[\theta(\hat{\theta} - \hat{\theta}^{Target})] \\
&= E(\hat{\theta}^2) - E^2(\hat{\theta}) + E^2(\hat{\theta}) - E(\hat{\theta}\hat{\theta}^{Target}) + E(\hat{\theta})E(\hat{\theta}^{Target}) - E(\hat{\theta})E(\hat{\theta}^{Target}) - E[\theta(\hat{\theta} - \hat{\theta}^{Target})] \\
&= Var(\hat{\theta}) + E^2(\hat{\theta}) - Cov(\hat{\theta}, \hat{\theta}^{Target}) - E(\hat{\theta})E(\hat{\theta}^{Target}) - E[\theta(\hat{\theta} - \hat{\theta}^{Target})] \\
&= Var(\hat{\theta}) - Cov(\hat{\theta}, \hat{\theta}^{Target}) + E^2(\hat{\theta}) - E(\hat{\theta})E(\hat{\theta}^{Target}) - E[\theta(\hat{\theta} - \hat{\theta}^{Target})] \\
&= Var(\hat{\theta}) - Cov(\hat{\theta}, \hat{\theta}^{Target}) + E(\hat{\theta})[E(\hat{\theta}) - E(\hat{\theta}^{Target})] - E[\theta(\hat{\theta} - \hat{\theta}^{Target})] \\
&= Var(\hat{\theta}) - Cov(\hat{\theta}, \hat{\theta}^{Target}) + E(\hat{\theta})E(\hat{\theta} - \hat{\theta}^{Target}) - \theta E(\hat{\theta} - \hat{\theta}^{Target}) \\
&= Var(\hat{\theta}) - Cov(\hat{\theta}, \hat{\theta}^{Target}) + E(\hat{\theta} - \hat{\theta}^{Target})(E(\hat{\theta}) - \theta) \\
&= Var(\hat{\theta}) - Cov(\hat{\theta}, \hat{\theta}^{Target}) + E(\hat{\theta} - \hat{\theta}^{Target})Bias(\hat{\theta}) := a
\end{aligned}$$

Let's denote
$b := E(\hat{\theta} - \hat{\theta}^{Target})^2$
$c := MSE(\hat{\theta})$
then

$$\begin{aligned}
MSE(\delta^\lambda) &= MSE(\hat{\theta}) + \lambda^2 E(\hat{\theta} - \hat{\theta}^{Target})^2 \\
&\quad - 2\lambda[Var(\hat{\theta}) - Cov(\hat{\theta}, \hat{\theta}^{Target}) + E(\hat{\theta} - \hat{\theta}^{Target})Bias(\hat{\theta})] \\
&= c + \lambda^2 b - 2\lambda a
\end{aligned}$$

□

The MSE function for shrinkage estimator δ^λ turns out to be a quadratic function of λ whose parameters a, b and c are completely dependent on merely two estimators $\hat{\theta}$ and $\hat{\theta}^{Target}$ (note that the MSE($\hat{\theta}$) = c diminishes and has no influence in the risk curve of MSE(δ^λ)). In other words, with $\hat{\theta}$ and $\hat{\theta}^{Target}$ which could be completely estimated from data, a optimal shrinkage intensity λ can always be calculated to minimize the MSE. In comparing to MSE of the unregularized estimator $\hat{\theta}$, the risk improment of δ^λ determined only by a and b. Since the MSE function of shrinkage estimator δ^λ is a quadratic function of λ, the optimal shrinkage

intensity $\lambda^\star$ is simply calculated as

$$\lambda^\star = \frac{a}{b} = \frac{Var(\hat{\theta}) - Cov(\hat{\theta}, \hat{\theta}^{Target}) + E(\hat{\theta} - \hat{\theta}^{Target})Bias(\hat{\theta})}{E(\hat{\theta} - \hat{\theta}^{Target})^2} \tag{18}$$

Some interesting properties of Equation 18 could be reviewed as follow:

- The smaller the variance of $\hat{\theta}$, the smaller becomes the numerator of Equation 18, leads to smaller $\lambda^\star$. With small $\lambda^\star$, the influence of the target $\hat{\theta}^{Target}$ was lessened. Therefore, with increasing sample size the shrinkage estimator grows more stabilized against the target estimator.
- The $\lambda^\star$ also decreases when the mean squared difference $E(\hat{\theta} - \hat{\theta}^{Target})^2$ increases (the denominator of Equation 18). It worths to mention that this characteristic of $\lambda^\star$ automatically protects the shrinkage estimator δ^λ against a misspecified target $\hat{\theta}^{Target}$.
- If $\hat{\theta}$ is an *unbiased* estimator of θ, that means $E(\hat{\theta}) = \theta$ and $Bias(\hat{\theta}) = 0$, this equation reduces to
$$\lambda^\star = \frac{Var(\hat{\theta}) - Cov(\hat{\theta}, \hat{\theta}^{Target})}{E(\hat{\theta} - \hat{\theta}^{Target})^2} \tag{19}$$
- If the unregularized estimator $\hat{\theta}$ is *biased*, and the bias points already towards the target, then the $\lambda^\star$ is correspondingly reduced.
- The $\lambda^\star$ depends on the correlation between estimation error of $\hat{\theta}$ and $\hat{\theta}^{Target}$. If both are positively correlated then the weight from the second term in the numerator of Equation 18 put on the shrinkage target decreases.

3.1.3. Estimation of the optimal shrinkage intensity

In practical application of the shrinkage rule (Equation 17) one needs to obtain an estimate $\hat{\lambda}^\star$ of the optimal shrinkage intensity by estimated from the data. Many ideas arises to estimate $\lambda^\star$ with the given Equation 18, from Thompson and Thompson (1968) in the univariate case to Ledoit and Wolf (2003) with the multivariate case. We take the suggestion from Schafer and Strimmer (2005) which estimate the optimal $\lambda^\star$ by replacing all expectations, variances, and covariances in Equation 18 by their *unbiased* empirical counterparts. This leads to

$$\hat{\lambda}^\star = \frac{\hat{a}}{\hat{b}} = \frac{\hat{V}ar(\hat{\theta}) - \hat{C}ov(\hat{\theta}, \hat{\theta}^{Target}) + (\hat{\theta} - \hat{\theta}^{Target})\hat{B}ias(\hat{\theta})}{(\hat{\theta} - \hat{\theta}^{Target})^2} \tag{20}$$

Of course in case of *unbiased* $\hat{\theta}$ this reduces again to

$$\hat{\lambda}^\star = \frac{\hat{V}ar(\hat{\theta}) - \hat{C}ov(\hat{\theta}, \hat{\theta}^{Target})}{(\hat{\theta} - \hat{\theta}^{Target})^2} \tag{21}$$

Some additional points to improve the efficiency of the shrinkage rule could be considered as well

- The construction of the above shrinkage estimator assumes nothing about a normal or any other distribution of the data. Furthermore, a simple theorem derived from Ledoit and Wolf (2003) show that the optimal shrinkage intensity $\lambda^\star$ from Equation 18 guarantees minimal the mean square error *without* the need of having to specify any underlying distributions, and could be computed *without* requiring computationally expensive procedures such as Monte Carlo Markov Chain, the bootstrap, or cross-validation. A part of this theorem was already shown from us in the derivation part.

- In finite samples $\hat{\lambda}^\star$ may exceed the value one, and in some cases it may even become negative. By *truncating* the estimated $\hat{\lambda}^\star$ using $\hat{\lambda}^\star = \max(0, \min(1, \hat{\lambda}^\star))$, it could avoid over shrinkage or negative shrinkage when compute the estimated $\hat{\lambda}^\star$.

- Another very useful application of this rule is that allows *multiple shrinkage intensities*, by having multiple targets, and each has its own different optimal shrinkage intensity. We will emphasize this idea again when considering the simulation setup.

3.1.4. Example

Consider again the old problem of Stein from the introduction part. Suppose we need to estimate the mean of a P-dimensional multivariate normal distribution with unit-diagonal covariance matrix from a single (N = 1) vector-valued observation $\underline{x}$. The maximum-likelihood estimator in this case equals the vector of observation, i.e. $\hat{\underline{\theta}}^{ML} = \underline{x}$. Since the estimator has P dimensions, that means $\hat{\theta}_k^{ML} = x_k \forall k = 1..P$. In this extreme case which none of any statistical frameworks could be helpful with only one observation, however, the efficiency over the ML-estimator could be improved with the shrinkage rule.
We choose the target is zero ($\hat{\theta}^{Target} = 0$), the covariances are zeros too because no correlation between the dimensions ($\text{Cov}(x_k, x_l) = 0$) and the variances are ones ($\text{Var}(x_k) = 1$). Insert into Equation 18 and Equation 20 results

$$\lambda^\star = \frac{Var(\hat{\underline{\theta}}^{ML})}{E(\hat{\underline{\theta}}^{ML})^2} = \frac{\sum_k^P Var(\hat{\theta}_k^{ML})}{\sum_k^P E(\hat{\theta}_k^{ML})^2} = \frac{\sum_k^P Var(x_k)}{\sum_k^P E(x_k)^2} = \frac{P}{\sum_k^P E(x_k)^2}$$
$$\Rightarrow \hat{\lambda}^\star = \frac{P}{\sum_k^P x_k^2}$$

The James-Stein shrinkage estimator is easily constructed from the shrinkage rule (Equation 17)

$$\hat{\theta}_k^{JS} = \hat{\lambda}^\star \hat{\theta}_k^{Target} + (1 - \hat{\lambda}^\star)\hat{\theta}_k^{ML} = \left(1 - \frac{P}{\sum_k^P x_k^2}\right) x_k$$

Another target followed from Lindley et al. (1972) is the mean across dimensions, i.e. $\hat{\theta}^{Target} = \overline{x} = \frac{1}{P}\sum_k^P x_k$, we get $\hat{\lambda}^\star = \frac{P-1}{\sum_k^P (x_k - \overline{x})^2}$ and obtain

$$\begin{aligned}
\hat{\theta}_k^{EM} &= \frac{P-1}{\sum_k^P (x_k - \overline{x})^2}\overline{x} + \left(1 - \frac{P-1}{\sum_k^P (x_k - \overline{x})^2}\right) x_k \\
&= x_k + \frac{P-1}{\sum_k^P (x_k - \overline{x})^2}(x_k - \overline{x}) \\
&= \overline{x} + \left(1 - \frac{P-1}{\sum_k^P (x_k - \overline{x})^2}\right)(x_k - \overline{x})
\end{aligned}$$

Dependent on what kind of estimator one needs to improve, many targets could be considered and the choice of the target could have a huge impact on efficiency. Schafer and Strimmer (2005) provide a fairly extensive review of the target choice in case of estimating the covariance matrix.

3.2. The shrinkage-t statistic

In context of microarray study, the most well-known problem is how to obtain a stable estimator of gene-specific variances. We'll show in detail how the shrinkage rule is applied in microarray data analysis.
From given data of one group with P genes, let's denote x_{ij} the j-th measurement of gene i $(i = 1..P, j = 1..N)$. In statistical context, that means x_{ij} is the j-th observation of the random variable X_i. The corresponding sample mean of gene i is defined as $\overline{x}_i = \frac{1}{N}\sum_j^N x_{ij}$. Now set $w_{ij} = (x_{ij} - \overline{x}_i)^2$ and $\overline{w}_i = \frac{1}{N}\sum_j^N (x_{ij} - \overline{x}_i)^2 = \frac{1}{N}\sum_j^N w_{ij}$. All the w_{ij} can be seen as j-th sample of the random variable W_i with $W_i = (X_i - \overline{X})^2$ and their sample mean is $\overline{w}_i$. The unbiased empirical variances v_i of each gene i equals

$$v_i = Var(X_i) = \frac{1}{N-1}\sum_j^N (x_{ij} - \overline{x}_i)^2 = \frac{N}{N-1}\overline{w}_i \tag{22}$$

$\forall i = 1..P$

In the standard t-test, these variances were used in calculating the t-statistic (Equation 4). However, in microarray data where the sample size is limited (typically smaller than 10 samples per group) those variances were shown to be very unstable which lead to false detection of differential-expressed genes (variance which too small in the denominator of t-statistic leads to large test statistic). The shrinkage rule improves the classical variance estimator as follow: These variances provide the components for the unregularized estimator $\hat{\theta}$ for the shrinkage rule (Equation 17). The next step is chosen an appropriate shrinkage target. Similar to the example case, one could consider to shrink towards zero or towards the mean of the empirical variances. However, these two alternative targets are either less efficient (zero target) or less robust (mean target). We taken the suggestion from Opgen-Rhein and Strimmer (2007) and

shrink towards the median value v_{median} of all v_i.
The optimal estimated shrinkage intensity was calculated as follow

$$\begin{aligned}\hat{\lambda}^{\star} &= \frac{\sum_{i=1}^{P} \widehat{Var}(\hat{\theta}_i) - \widehat{Cov}(\hat{\theta}_i, \hat{\theta}_i^{Target}) + (\hat{\theta}_i - \hat{\theta}_i^{Target})\widehat{Bias}(\hat{\theta}_i)}{\sum_{i=1}^{P}(\hat{\theta}_i - \hat{\theta}_i^{Target})^2} \\ &= \frac{\sum_{i=1}^{P} \widehat{Var}(v_i) - \widehat{Cov}(v_i, v_{median}) + (v_i - v_{median})\widehat{Bias}(v_i)}{\sum_{i=1}^{P}(v_i - v_{median})^2}\end{aligned}$$

Since the empirical variance v_i is unbiased, $Bias(v_i) = 0$. Furthermore, the approximation $\widehat{Cov}(v_i, v_{median}) \approx 0$ could be use as well for simplicity. We obtain

$$\hat{\lambda}^{\star} = \frac{\sum_{i=1}^{P} \widehat{Var}(v_i)}{\sum_{i=1}^{P}(v_i - v_{median})^2} \tag{23}$$

If $\hat{\lambda}^{\star}$ has a small value, that means little shrinkage occurs, in opposition when $\hat{\lambda}^{\star}$ is large, more shrinkage takes place. As we could see from the numerator of the Equation 23, it could determine whenever $\hat{\lambda}^{\star}$ small or large. The numerator is the sum of all $\widehat{Var}(v_i)$, which is small if the empirical variances v_i could be reliably determined from the data, consequently there will be little shrinkage. In opposite, if $\widehat{Var}(v_i)$ is large which means the estimation of the empirical variances is not reliable, then the shrinkage procedure pooling across genes will take place. Furthermore, the denominator is an estimate of the mis-specification between the target and the v_i. If the target is incorrectly chosen, the denominator tends to very large which consequently minimize the $\hat{\lambda}^{\star}$ towards zero, hence, no shrinkage will occur.
A sample version of $\widehat{Var}(v_i)$ could be calculated as follow

$$\begin{aligned}\widehat{Var}(v_i) &= \widehat{Var}(\frac{N}{N-1}\overline{w}_i) = \frac{N^2}{(N-1)^2}\widehat{Var}(\overline{w}_i) \\ &= \frac{N^2}{(N-1)^2}\widehat{Var}(\frac{1}{N}\sum_{j}^{N} w_{ij}) = \frac{1}{(N-1)^2}\widehat{Var}(\sum_{j}^{N} w_{ij}) \\ &= \frac{1}{(N-1)^2}\sum_{j}^{N}\widehat{Var}(w_{ij})\end{aligned}$$

Since $\widehat{Var}(w_{ij})$ could be calculated as the unbiased sample variance of W_i, ($= \frac{1}{N-1}\sum_{k=1}^{N}(w_{ik} - \overline{w}_i)^2, \forall i = 1..N$), we obtain

$$\widehat{Var}(v_i) = \frac{1}{(N-1)^2}\sum_{j}^{N}\left(\frac{1}{N-1}\sum_{k=1}^{N}(w_{ik} - \overline{w}_i)^2\right) = \frac{N}{(N-1)^3}\sum_{k=1}^{N}(w_{ik} - \overline{w}_i)^2$$

Insert in Equation 23

$$\Rightarrow \hat{\lambda}^{\star} = \frac{\frac{N}{(N-1)^3}\sum_{i=1}^{P}\sum_{k=1}^{N}(w_{ik} - \overline{w}_i)^2}{\sum_{i=1}^{P}(v_i - v_{median})^2} \tag{24}$$

The shrinkage estimator for sample variance for the i-th gene is straightforward to compute

$$v_i^\star = \hat{\lambda}^\star v_{median} + (1 - \hat{\lambda}^\star) v_i \tag{25}$$

With the above framework, the new statistic, the "shrinkage t" statistic, is obtained from the suggestion of Opgen-Rhein and Strimmer (2007) by simply plugging the shrinkage variance estimator (Equation 24 and Equation 25 into the ordinary statistic.
Let's denote N_1 and N_2 as the sample sizes of group 1 (e.g. treatment group) and group 2 (e.g. control group), we could compute the shrinkage variance estimator for each group with two different shrinkage intensities $\hat{\lambda}_1^\star$ and $\hat{\lambda}_2^\star$, let's denote as $v_i^\star$ and $u_i^\star$ respectively. The sample mean of each group are $\overline{x}_i$ and $\overline{y}_i$. The shrinkage t-statistic for detecting differential-express of gene i is given by

$$t_i^\star = \frac{\overline{x}_i - \overline{y}_i}{\sqrt{\frac{v_i^\star}{N_1} + \frac{u_i^\star}{N_2}}} \tag{26}$$

Similar to the standard t statistic, one can consider to have a pooling shrinkage variance estimator for both groups (i.e. with one common shrinkage intensity) and plug it in the above equation as well.

3.3. The shrinkage-ANOVA

We're given the data of K-groups. An example from microarray analysis context might be 3 groups data with one control group, two treatments groups in which one for drug treatment and another one for another drug treatment. Each group should have their own sample size N_k.
Let's denote $x_{ij}^{(k)}$ the j-th measurement of gene i from group k $(i = 1..P, j = 1..N_k, k = 1..K)$. We define those following terms:

- Sample mean of gen i for group k

$$\overline{\mu}_i^{(k)} = \frac{1}{N_k} \sum_j^{N_k} x_{ij}^{(k)}$$

- The grand mean of gen i for all groups.

$$\overline{\mu}_i = \frac{\sum_{k=1}^K \sum_{j=1}^{N_k} x_{ij}^{(k)}}{\sum_{k=1}^K N_k} = \frac{\sum_{k=1}^K \sum_{j=1}^{N_k} x_{ij}^{(k)}}{N}$$

- To calculate $v_i^{(k)}$ the unbiased sample variance of gene i for group k, we first define

$$w_{ij}^{(k)} = (x_{ij}^{(k)} - \overline{\mu}_i^{(k)})^2$$
$$\overline{w}_i^{(k)} = \frac{1}{N_k}\sum_j^{N_k}(x_{ij}^{(k)} - \overline{\mu}_i^{(k)})^2 = \frac{1}{N_k}\sum_j^{N_k} w_{ij}^{(k)}$$

  Then

$$v_i^{(k)} = \frac{1}{N_k - 1}\sum_j^{N_k}(x_{ij}^{(k)} - \overline{\mu}_i^{(k)})^2 = \frac{N_k}{N_k - 1}\overline{w}_i^{(k)}$$

- $v_{target}^{(k)}$ as the target of shrinkage rule of $v_i^{(k)} \forall i$, for group k.

The estimated optimal shrinkage intensity for group k was computed according to Equation 21 as follow

$$\begin{aligned}\hat{\lambda}^{(k)\star} &= \frac{\sum_{i=1}^P \widehat{Var}(v_i^{(k)}) - \widehat{Cov}(v_i^{(k)}, v_{target}^{(k)})}{\sum_{i=1}^P (v_i^{(k)} - v_{target}^{(k)})^2} \\ &\approx \frac{\sum_{i=1}^P \widehat{Var}(v_i^{(k)})}{\sum_{i=1}^P (v_i^{(k)} - v_{target}^{(k)})^2}\end{aligned} \tag{27}$$

The computation of $\widehat{Var}(v_i^{(k)})$ for each group k is analog to the shrinkage-t approach and was being separately computed for each group.

$$\begin{aligned}\widehat{Var}(v_i^{(k)}) &= \widehat{Var}(\frac{N_k}{N_k - 1}\overline{w}_i^{(k)}) = \frac{N_k^2}{(N_k - 1)^2}\widehat{Var}(\overline{w}_i^{(k)}) \\ &= \frac{N_k^2}{(N_k - 1)^2}\widehat{Var}(\frac{1}{N_k}\sum_j^{N_k} w_{ij}^{(k)}) = \frac{1}{(N_k - 1)^2}\widehat{Var}(\sum_j^{N_k} w_{ij}^{(k)}) \\ &= \frac{1}{(N_k - 1)^2}\sum_j^{N_k}\widehat{Var}(w_{ij}^{(k)}) \\ &= \frac{1}{(N_k - 1)^2}\sum_j^{N_k}\left(\frac{1}{N_k - 1}\sum_{k=1}^{N_k}(w_{ik}^{(k)} - \overline{w}_i^{(k)})^2\right) \\ &= \frac{N_k}{(N_k - 1)^3}\sum_{k=1}^{N_k}(w_{ik}^{(k)} - \overline{w}_i^{(k)})^2\end{aligned}$$

Finally we have

$$\Rightarrow \hat{\lambda}^{(k)\star} = \frac{\frac{N_k}{(N_k-1)^3}\sum_{i=1}^P\sum_{k=1}^{N_k}(w_{ik}^{(k)} - \overline{w}_i)^2}{\sum_{i=1}^P (v_i^{(k)} - v_{target}^{(k)})^2} \tag{28}$$

Insert this estimated shrinkage intensity in Equation 24 we receive the shrinkage estimator for sample variance for the i-th gene of group k

$$v_i^{(k)\star} = v_{target}^{(k)} \hat{\lambda}^{(k)\star} + (1 - \hat{\lambda}^{(k)\star}) v_i^{(k)} \tag{29}$$

After obtain the shrinkage variance estimators as above, we could use those to plug in the formulas for F-statistic of One-Way k-groups ANOVA. The definitions are similar

$$SST_i = \sum_{k=1}^{K} N_k (\overline{\mu}_i^{(k)} - \overline{\mu}_i)^2 \tag{30}$$

$$MST_i = \frac{SST_i}{K-1} \tag{31}$$

$$SSE_i^\star = \sum_{k=1}^{K} (N_k - 1) v_i^{(k)\star} \tag{32}$$

$$MSE_i^\star = \frac{SSE_i^\star}{N-K} \tag{33}$$

The only thing that changes in comparison with the ordinary One-Way ANOVA is the sum of squares error, which could improve by the shrinkage rule.
The shrinkage F-statistic for detecting differential-express of gene i is given by

$$F_i^\star = \frac{MST_i}{MSE_i^\star} = \frac{SST_i/(K-1)}{SSE_i^\star/(N-K)} \tag{34}$$

4. Shrinkage-t simulation

As we already saw, the shrinkage rule is not simple to understand, even the derivation of this rule isn't easy either. Nevertheless, the application of this rule to extend the ordinary t-statistic is simply straight forward.
To examine how good the "naive" application of James-Stein shrinkage rule to the ordinary t-test effects the determination of differentially expressed genes, we set up some simulations to see how good the new rule could improve to estimate the variance and how stable the new rule was when it was adapted to various conditional settings.

4.1. Simulation settings

We would like to examine some properties and qualities of the James-Stein rule as follow

- Since the James-Stein rule modifies the variance of the standard t-test, the new estimated variance is under suspicion not Chi-squared distributed anymore, thus we're not hoping that the new null hypothesis distribution is still exactly a t-distribution. It is still worth to see that under the given true null hypothesis, what kind of distribution could deliver the new rule, as well as how different the new distribution to the standard t-distribution of the ordinary case.

- Since the application of the James-Stein rule in microarray analysis is to improve the estimating of genes-specific variances, the difference between the new improved estimated variances and the true variances was detailed examined as well.

We considered 4 simulation settings in the hope to cover some known cases which could happens in a real microarray study. The simulated data were generated from a normal distribution with the null hypothesis is $\mu = 0$ and the settings for variance of each group as follow

(I) The first simulation option is very simple, consider the null hypothesis is $\sigma^2 = 1$ for both groups. The standard t-test should be "behaved" perfect in this case.

(II) This simulation reflects the standard behavior of two groups microarray analysis. One group has a small variance while another group has a larger variance. Nonetheless the mean values of two groups are equal. The null hypothesis is where the variance for group 1 is $\sigma_1^2 = 1$ and the variance for group 2 is $\sigma_2^2 = 100$. The standard t-test should still have a perfect curve in this case.

(III) All the genes from two groups have a same variance $\sigma^2 = 1$, but some genes may behave like *outliers* in comparison with another genes, and has a larger variance $\sigma^2 = 100$. Those genes with this behavior were randomly chosen and the number of those *outliers* in the total number of genes is fixed with $p = 1, 2, 5, 10, 20, 50\%$.

(IV) In this setting each group has his own variance like in setting (II) ($\sigma_1^2 = 1$, $\sigma_2^2 = 100$), but two groups also have outliers as the same in setting (III). This case can be considered as a combination of setting (II) and (III).

4.2. Choices of the target

A typical microarray experiment delivers an informative scene as follow: after some "tendious" preprocessing routines we receive the mRNA-abundant measurements in form of real values, contain mostly in two groups, one for the control and one for the treatment. Dependent on what kinds of biological host we're dealing with, those values contains for thousands to ten thousands genes.
Let's denote x_{ij} and y_{ij} the j-th measurement of gene i from the control group and treatment group, respectively. $(i = 1..P, j = 1..N)$
We chosen the number of genes P = 5000 and the number of samples N = 3 for our simulations same as the simulation which be used in Opgen-Rhein and Strimmer (2007). Of course each group could have an own sample size, for convenience we assume that two groups have the same sample size N. The case with two different sample sizes could be analog derived.
The sample mean of each group is defined as follow

$$\overline{x}_i = \frac{1}{N} \sum_k^N x_{ik}$$

$$\overline{y}_i = \frac{1}{N} \sum_k^N y_{ik}$$

The unbiased sample variances of two groups are v_i and w_i respectively

$$\hat{v}_i = \frac{1}{N-1}\sum_{k}^{N}(x_{ik} - \overline{x}_i)^2$$
$$\hat{w}_i = \frac{1}{N-1}\sum_{k}^{N}(y_{ik} - \overline{y}_i)^2$$

Because of the flexibility of shrinkage rule, we're trying not to limit any possibilities of extending the rule. All of evaluating constructions below are merely our ideas how to choose wisely the target of the shrinkage rule. Any another freely ideas could be considered as well, but as we mentioned in the construction part of the James-Stein rule, a target could be wisely or badly chosen, and although the new rule has his mechanism to automatically prevent over- or undershrink, a wisely chosen construction for the target should be always taken place before.

1. This case assumes that two groups of microarray data have for each group a distinctive variance (same as the $unequal$-variance case of the ordinary t-test), therefore for each group exists a separate target (multiple targets), and for each new estimated variance exists a distinct shrinkage intensity λ (multiple shrinkage intensities). Let's denote the targets as

$$v_{target} = median(\hat{v}_1, \hat{v}_2, ..., \hat{v}_N)$$
$$w_{target} = median(\hat{w}_1, \hat{w}_2, ..., \hat{w}_N)$$

   The shrinkage intensity for each group was defined as (Equation 24)

$$\hat{\lambda}_1 = \frac{\frac{N}{(N-1)^3}\sum_{i=1}^{P}\sum_{k=1}^{N}\left((x_{ik} - \overline{x}_i)^2 - \frac{1}{N}\sum_{j=1}^{N}(x_{ij} - \overline{x}_i)^2\right)^2}{\sum_{i=1}^{P}(\hat{v}_i - v_{target})^2}$$
$$\hat{\lambda}_2 = \frac{\frac{N}{(N-1)^3}\sum_{i=1}^{P}\sum_{k=1}^{N}\left((y_{ik} - \overline{y}_i)^2 - \frac{1}{N}\sum_{j=1}^{N}(y_{ij} - \overline{y}_i)^2\right)^2}{\sum_{i=1}^{P}(\hat{w}_i - w_{target})^2}$$

   The new improved estimated variance for each group was defined as (Equation 25)

$$v_i^\star = (1 - \hat{\lambda}_1)\hat{v}_i + \hat{\lambda}_1 v_{target}$$
$$w_i^\star = (1 - \hat{\lambda}_2)\hat{w}_i + \hat{\lambda}_2 w_{target}$$

   The shrinkage t-statistic for detecting differential-express of gene i is simple (Equation 26)

$$t_i^{(1)} = \frac{\overline{x}_i - \overline{y}_i}{\sqrt{\frac{v_i^\star + w_i^\star}{N}}}$$

After our inspection in Opgen-Rhein and Strimmer (2007) work, this case turns out to be their default option of calculating the shrinkage variance when two groups have different variances in the st-package, which was programmed in R-Framework.

2. This case uses the same assumption as the first case, instead of using different targets, we use the $same$ target for two shrinkage variances. This target was chosen as

$$v_{target} = median(\hat{v}_1, \hat{v}_2, ..., \hat{v}_N, \hat{w}_1, \hat{w}_2, ..., \hat{w}_N)$$

Others terms are defined as the above case.

3. In this case, the idea is to treat two groups as one group, and we define a new variable as follow

$$\hat{u}_i = \frac{\hat{v}_i + \hat{w}_i}{2} \tag{35}$$

The target was chosen with the new variable

$$u_{target} = median(\hat{u}_1, \hat{u}_2, ..., \hat{u}_N)$$

The multiple shrinkage intensities was eliminated in this case, turn out to have one target with one shrinkage intensity. Therefore the data of two groups x_{ij} and y_{ij} need to combine to one group z_{ij} as follow

$$z_{ij} = x_{ij} \forall i = 1..P, \forall j = 1..N$$
$$z_{ij} = y_{i,(j-N)} \forall i = 1..P, \forall j = (N+1)..2N$$

Let's denote M = 2N, the new dataset has the new sample mean and unbiased sample variance as well

$$\overline{z}_i = \frac{1}{M} \sum_k^M z_{ik}$$
$$\hat{z}_i = \frac{1}{M-1} \sum_k^M (z_{ik} - \overline{z}_i)^2$$

The $shared$ shrinkage intensity was calculated as

$$\hat{\lambda} = \frac{\frac{M}{(M-1)^3} \sum_{i=1}^P \sum_{k=1}^M \left((z_{ik} - \overline{z}_i)^2 - \frac{1}{M} \sum_{j=1}^M (z_{ij} - \overline{z}_i)^2 \right)^2}{\sum_{i=1}^P (\hat{u}_i - u_{target})^2}$$

The shrinkage variance and the shrinkage t-statistic for this case are simply analog

$$u_i^\star = (1 - \hat{\lambda}) \hat{u}_i + \hat{\lambda} u_{target}$$
$$t_i^{(3)} = \frac{\overline{x}_i - \overline{y}_i}{\sqrt{\frac{2u_i^\star}{N}}}$$

4. This case uses the same idea of common target of the 2. case. The target was chosen in this case is however the u_{target} from the 3. case. That means we have the multiple shrinkage intensities as follow

$$\hat{\lambda}_1 = \frac{\frac{N}{(N-1)^3}\sum_{i=1}^{P}\sum_{k=1}^{N}\left((x_{ik}-\overline{x}_i)^2 - \frac{1}{N}\sum_{j=1}^{N}(x_{ij}-\overline{x}_i)^2\right)^2}{\sum_{i=1}^{P}(\hat{v}_i - u_{target})^2}$$
$$\hat{\lambda}_2 = \frac{\frac{N}{(N-1)^3}\sum_{i=1}^{P}\sum_{k=1}^{N}\left((y_{ik}-\overline{y}_i)^2 - \frac{1}{N}\sum_{j=1}^{N}(y_{ij}-\overline{y}_i)^2\right)^2}{\sum_{i=1}^{P}(\hat{w}_i - u_{target})^2}$$

The shrinkage variances were obvious separately calculated

$$v_i^\star = (1-\hat{\lambda}_1)\hat{v}_i + \hat{\lambda}_1 u_{target}$$
$$w_i^\star = (1-\hat{\lambda}_2)\hat{w}_i + \hat{\lambda}_2 u_{target}$$

The shrinkage t-statistic for detecting differential-express of gene i is analog (but obviously not the same) with the 1. case

$$t_i^{(4)} = \frac{\overline{x}_i - \overline{y}_i}{\sqrt{\frac{v_i^\star + w_i^\star}{N}}}$$

5. This case uses the assumption that the two groups come with the same variance. We discovered that the st-package from Opgen-Rhein and Strimmer (2007) uses the same approach for compute the shrinkage t-statistic for the same variance case.
This case uses the combined dataset z_{ij} from the 3. case. The target was chosen as the median of all unbiased sample variances of z.

$$z_{target} = median(\hat{z}_1, \hat{z}_2, ..., \hat{z}_N)$$

The $combined$ shrinkage intensity was also calculated completely from z

$$\hat{\lambda} = \frac{\frac{M}{(M-1)^3}\sum_{i=1}^{P}\sum_{k=1}^{M}\left((z_{ik}-\overline{z}_i)^2 - \frac{1}{M}\sum_{j=1}^{M}(z_{ij}-\overline{z}_i)^2\right)^2}{\sum_{i=1}^{P}(\hat{z}_i - z_{target})^2}$$

The $combined$ shrinkage variance and the shrinkage t-statistic for this case are simply straightforward

$$z_i^\star = (1-\hat{\lambda})\hat{z}_i + \hat{\lambda} z_{target}$$
$$t_i^{(5)} = \frac{\overline{x}_i - \overline{y}_i}{\sqrt{\frac{2z_i^\star}{N}}}$$

4.3. Methods of evaluation

There are some properties of the James-Stein shrinkage rule when applied in microarray study that we're interested in

i. How well could the new rule do in term of estimating the variance?

ii. Under the assumption of a true null hypothesis, which distribution deliver the shrinkage t-statistic? What is the differences of this new distribution in comparing with the standard t-distribution?

iii. In case that the new distribution is a t-distribution, which degrees of freedom has the distribution and that degrees could be estimated as well?

To answer the i.question, one could use many methods for evaluating as well. We decided to measure the distance of the new estimated variance with the true variance since we know the true variance from simulations. Because there are 5 cases to estimate the variance, the score function for each case should be well defined as well. In some cases which we have two targets to measure (two variances in the 1. case for example), some other cases which we have merely one target (one shared variance in the 3. case). Therefore we'll define two score functions to cover the interested question for every case. Let's denote σ_1^2, σ_2^2 are the true known variances ($\sigma_1^2 = \sigma_2^2$ for the one variance case) and $\sigma^2 = \frac{\sigma_1^2+\sigma_2^2}{2}$. We define for each case from the above 5.cases two score functions as follow

(1) This case has two estimated variances for each gene i: $v_i^\star$ and $w_i^\star$

$$score^{(1)} = \sum_i^P \left((v_i^\star - \sigma_1^2)^2 + (w_i^\star - \sigma_2^2)^2\right)$$
$$\widetilde{score}^{(1)} = \sum_i^P \left(\frac{v_i^\star + w_i^\star}{2} - \sigma^2\right)^2$$

(2) This case has also two estimated variances for each gene i: $v_i^\star$ and $w_i^\star$

$$score^{(2)} = \sum_i^P \left((v_i^\star - \sigma_1^2)^2 + (w_i^\star - \sigma_2^2)^2\right)$$
$$\widetilde{score}^{(2)} = \sum_i^P \left(\frac{v_i^\star + w_i^\star}{2} - \sigma^2\right)^2$$

(3) This case has merely one estimated variance $u_i^\star$ for each gene i. Therefore the first score function could not be defined.

$$\widetilde{score}^{(3)} = \sum_i^P \left(u_i^\star - \sigma^2\right)^2$$

(4) This case is similar to 1. and 3.case, has also two estimated variances for each gene i: $v_i^\star$ and $w_i^\star$

$$score^{(4)} = \sum_i^P \left((v_i^\star - \sigma_1^2)^2 + (w_i^\star - \sigma_2^2)^2\right)$$
$$\widetilde{score}^{(4)} = \sum_i^P \left(\frac{v_i^\star + w_i^\star}{2} - \sigma^2\right)^2$$

(5) This case has again merely one estimated variance $z_i^\star$ for each gene i.

$$\widetilde{score}^{(5)} = \sum_i^P \left(z_i^\star - \sigma^2\right)^2$$

4.4. Result

The simulation setting was set with the number of genes P = 5000 and the number of samples N = 3. All the four considered simulation options in which the third and the fourth simulation option yields for each option another six new options (for each $p = 1, 2, 5, 10, 20, 50\%$), were combined with five cases of choices for the targets, make a total number of $2 + 6 + 6 = 14$ simulations options x 5 choices of targets = 70 combinations. For each combination the two score functions for each gene was calculated according to the methods mentioned above and stored for 5000 genes. The procedure was repeated 500 times and a cumulative histogram of the two score functions for each combination was plotted. Since the (III) and (IV) simulation options are more sophisticated than the (I) and (II), the histograms of those options were shown in Figure 1 the comparison for (III) option and in Figure 2 for (IV) option.

Similar for the $\widetilde{score}$ functions which merely reduce the number of outliers to p = 2%, the histograms of $\widetilde{score}$ functions were shown in Figure 3 and Figure 4

To examine the properties of the new shrinkage t-distribution, the same procedure was performed for the calculation of the shrinkage t-statistic. To simply the problem, we assume that the new shrinkage-t distribution is already a t-distribution. We're interested in which degrees of freedom (dof) has the new shrinkage t-distribution, so there are some approaches to estimate the degrees of freedom for the new distribution

- The theoretical dof of a t-distribution for the two samples case of t-test is $(N+N-2)$.
- Since it's been conversant that the variance and the degrees of freedom of a t-distribution is connected, that is $var = \frac{dof}{dof-2}$. That means if we know about the variance of a t-distribution, we could backwards calculate the dof as follow

  $$dof = \frac{2var}{var - 1}$$

  With this formula, we calculate all the values of the new shrinkage t-distribution as

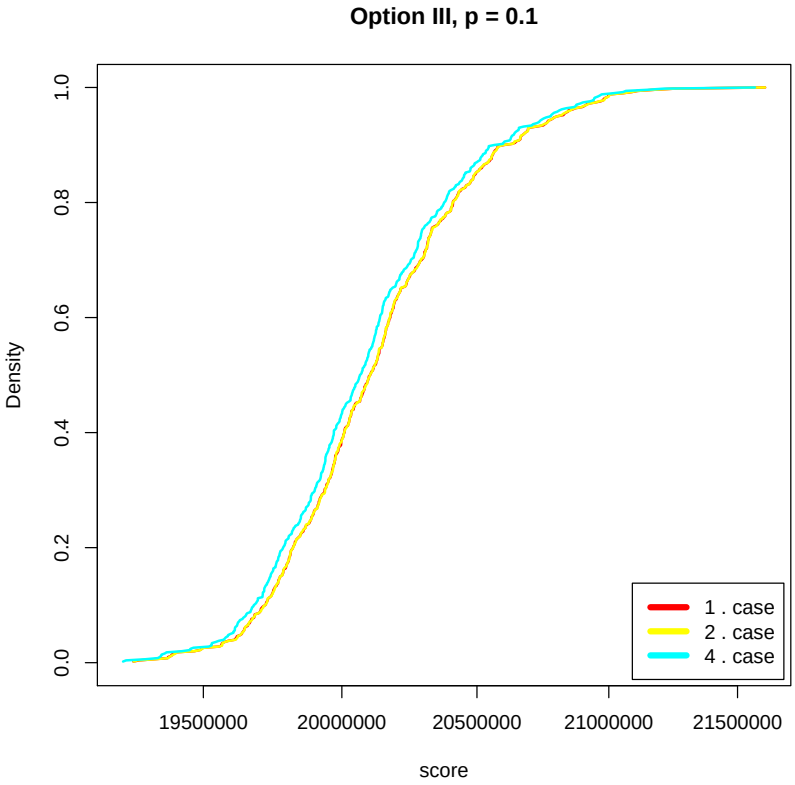

Figure 1: Cumulative histograms of score for the (III) simulation option and outliers number p = 10%

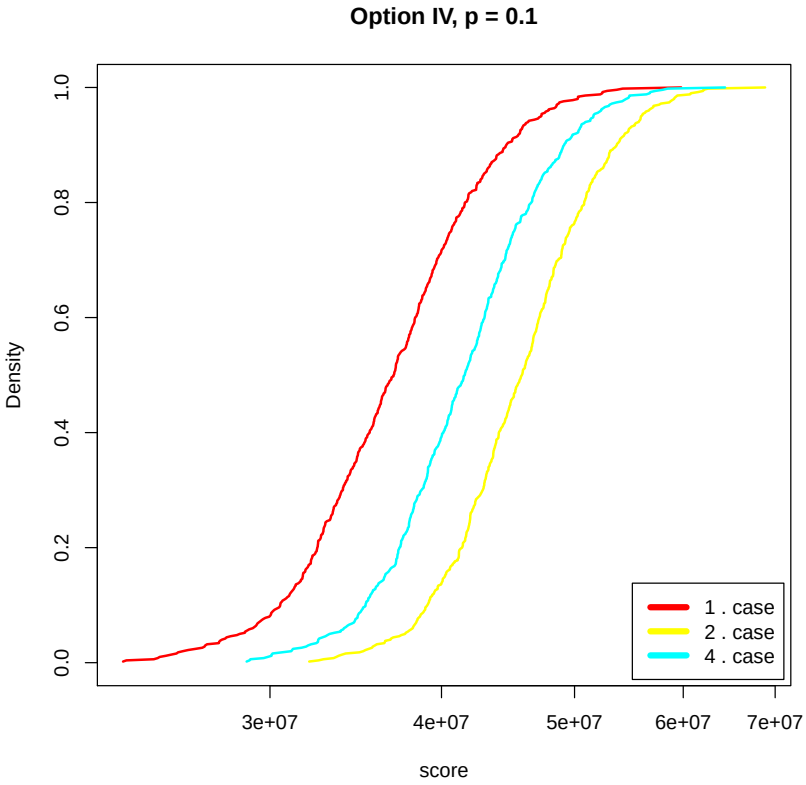

Figure 2: Cumulative histograms of score for the (IV) simulation option and outliers number p = 10%

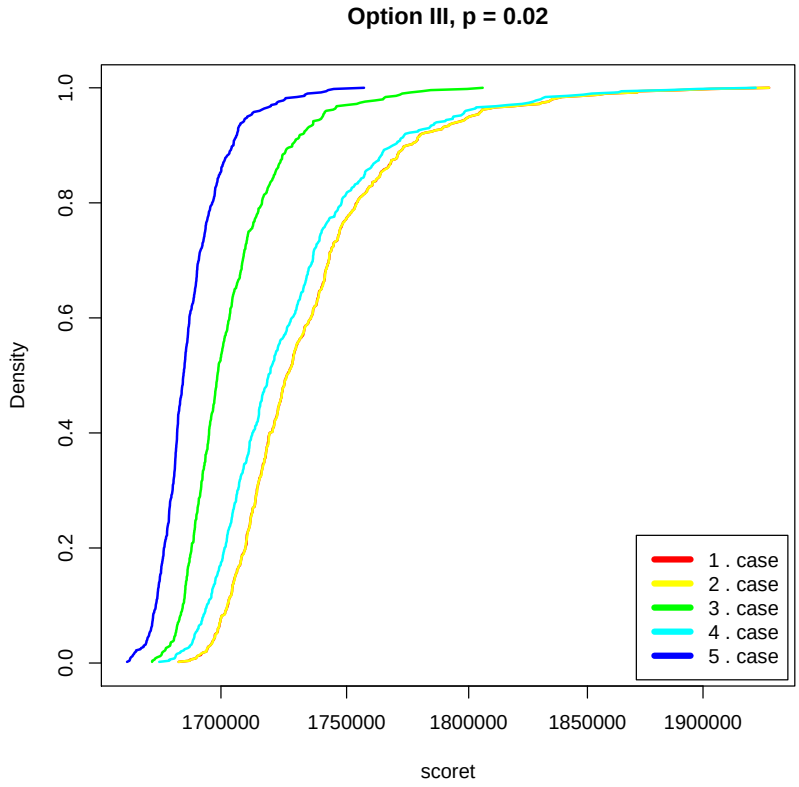

Figure 3: Cumulative histograms of $\widetilde{score}$ for the (III) simulation option and outliers number p = 2%

the methods section described above, then compute the variance of all the values and subsequently the dof.

- With a similar idea we calculate for each simulation the shrinkage t-statistic as a function of lambda and after 500 times repeated simulations the dof was estimated from the function of lambda.

The cumulative distribution function of the shrinkage t-statistic and others cumulative distribution functions of the standard t-distribution with the theoretical, from variance estimated and from lambda estimated degrees of freedom were shown in Figure 5 for the (III) simulation settings with 1% outlines and in Figure 6 for the (IV) simulation settings with 1% outlines.

5. Shrinkage-ANOVA simulation

As we already saw, the shrinkage rule is not simple to understand, even the derivation of this rule isn't easy either. Nevertheless, the application of this rule to extend the ordinary t-statistic is simply straight forward.
To examine how good the "naive" application of James-Stein shrinkage rule to the ordinary

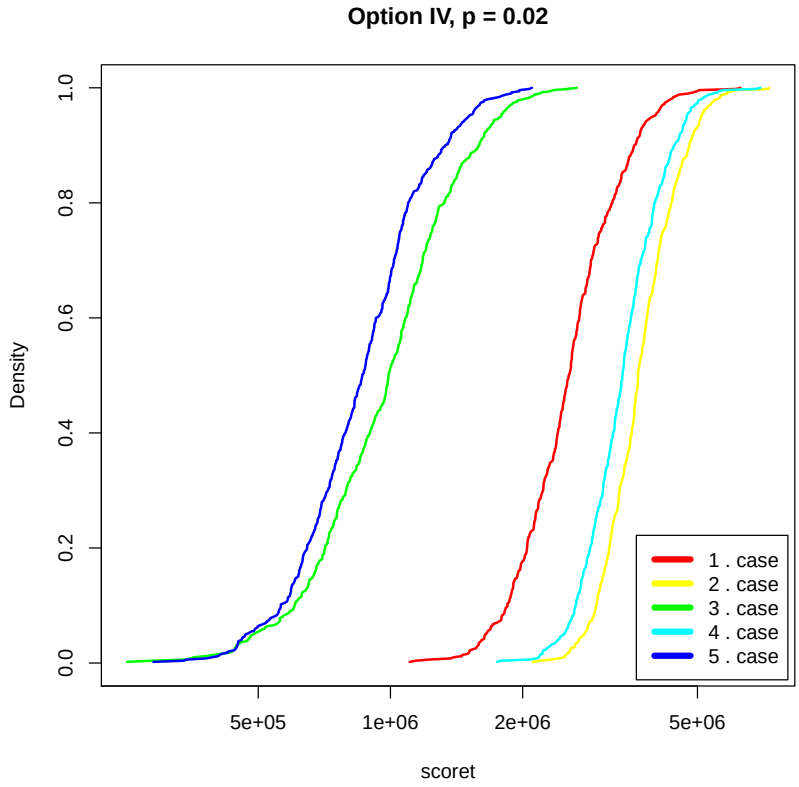

Figure 4: Cumulative histograms of $\widetilde{score}$ for the (IV) simulation option and outliers number p $=$ 2%

t-test effects the determination of differentially expressed genes, we set up some simulations to see how good the new rule could improve to estimate the variance and how stable the new rule was when it was adapted to various conditional settings.

5.1. Simulation settings

We would like to examine some properties and qualities of the James-Stein rule as follow

- Since the James-Stein rule modifies the variance of the standard t-test, the new estimated variance is under suspicion not Chi-squared distributed anymore, thus we're not hoping that the new null hypothesis distribution is still exactly a t-distribution. It is still worth to see that under the given true null hypothesis, what kind of distribution could deliver the new rule, as well as how different the new distribution to the standard t-distribution of the ordinary case.

- Since the application of the James-Stein rule in microarray analysis is to improve the estimating of genes-specific variances, the difference between the new improved estimated variances and the true variances was detailed examined as well.

We considered 4 simulation settings in the hope to cover some known cases which could happens in a real microarray study. The simulated data were generated from a normal distribution

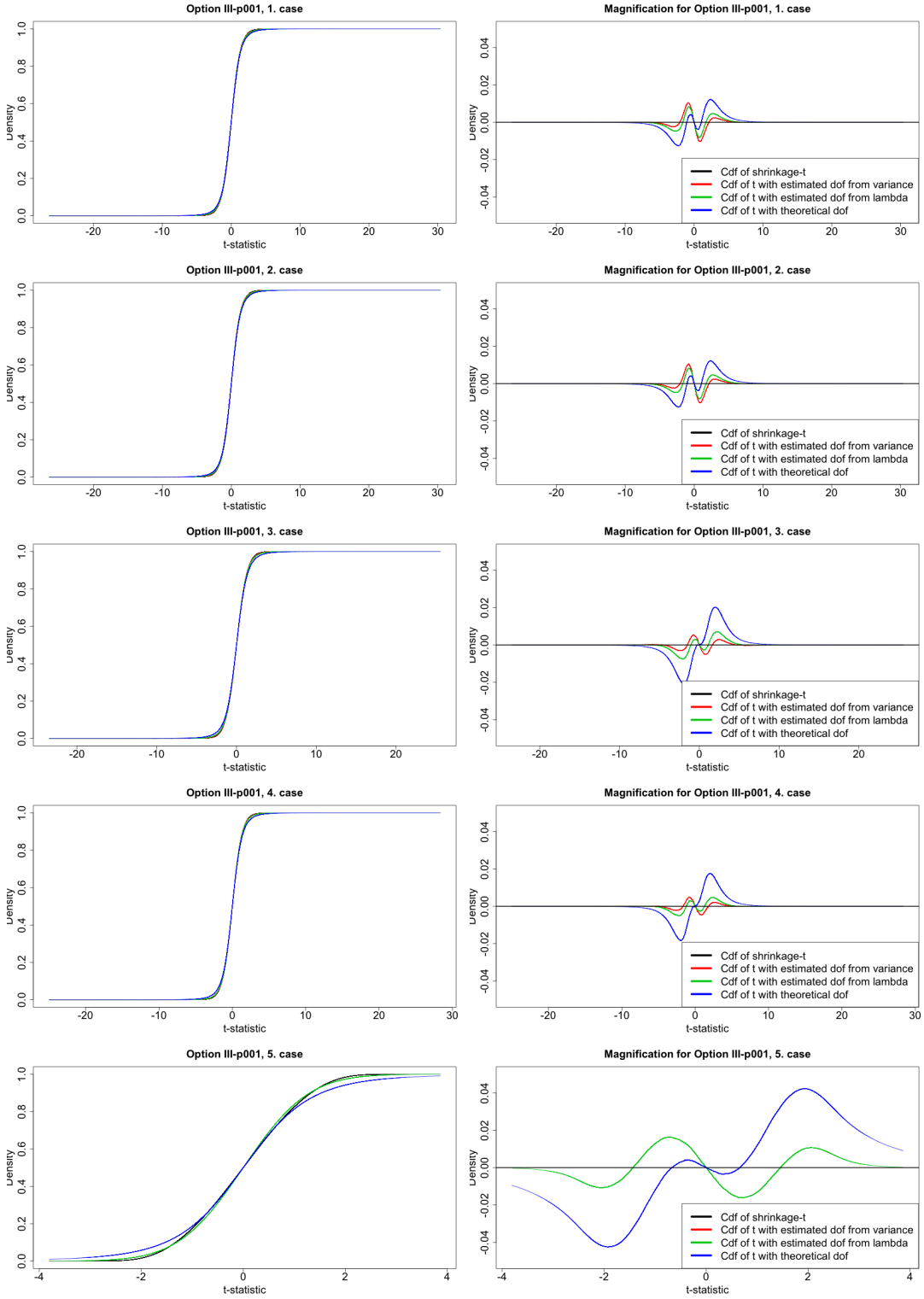

Figure 5: Cumulative distribution functions for the (III) simulation option and outliers number p = 1%

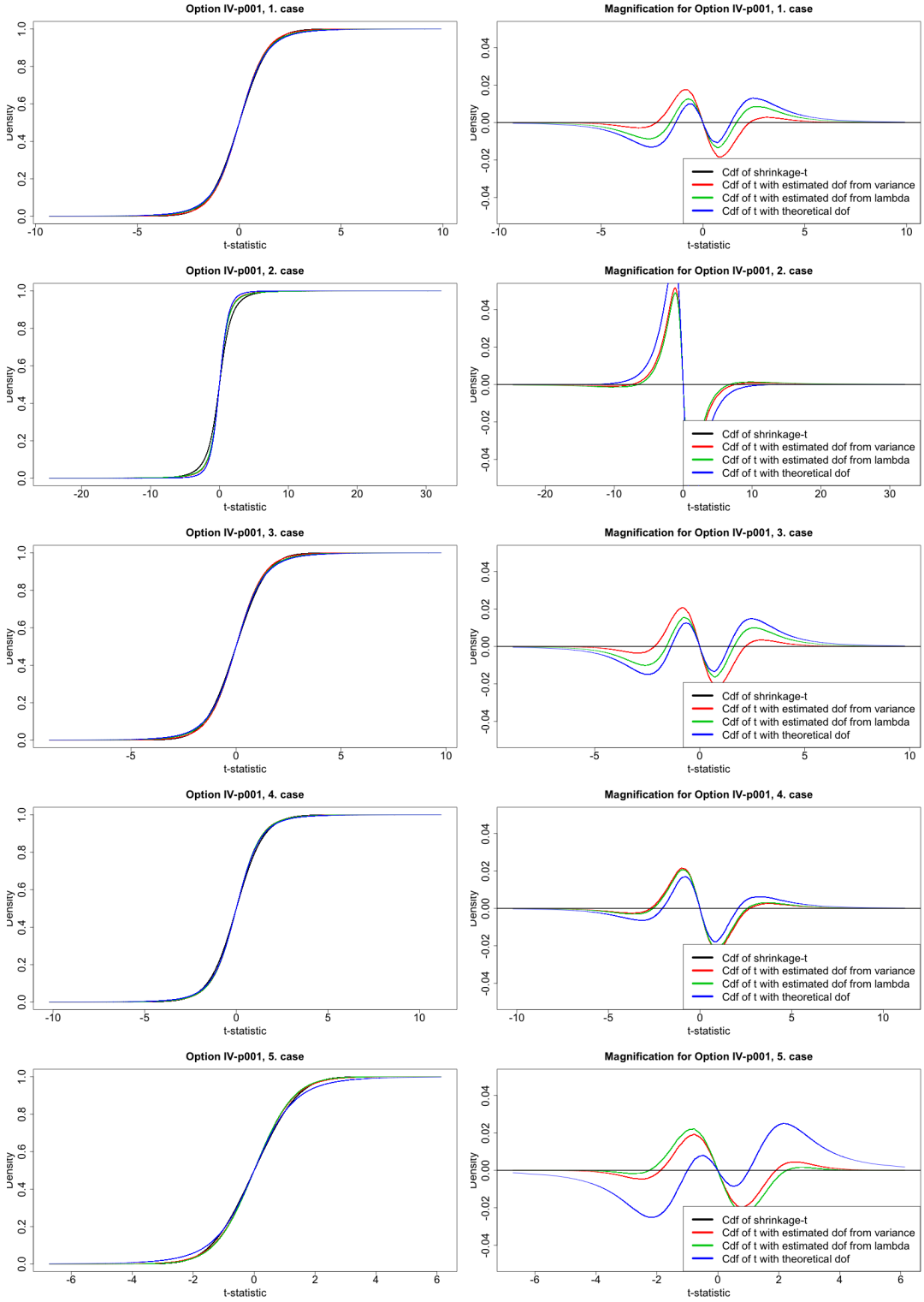

Figure 6: Cumulative distribution functions for the (IV) simulation option and outliers number p = 1%

with the null hypothesis is $\mu = 0$ and the settings for variance of each group as follow

(I) The first simulation option is very simple, consider the null hypothesis is $\sigma^2 = 1$ for both groups. The standard t-test should be "behaved" perfect in this case.

(II) This simulation reflects the standard behavior of two groups microarray analysis. One group has a small variance while another group has a larger variance. Nonetheless the mean values of two groups are equal. The null hypothesis is where the variance for group 1 is $\sigma_1^2 = 1$ and the variance for group 2 is $\sigma_2^2 = 100$. The standard t-test should still have a perfect curve in this case.

(III) All the genes from two groups have a same variance $\sigma^2 = 1$, but some genes may behave like *outliers* in comparison with another genes, and has a larger variance $\sigma^2 = 100$. Those genes with this behavior were randomly chosen and the number of those *outliers* in the total number of genes is fixed with $p = 1, 2, 5, 10, 20, 50\%$.

(IV) In this setting each group has his own variance like in setting (II) ($\sigma_1^2 = 1$, $\sigma_2^2 = 100$), but two groups also have outliers as the same in setting (III). This case can be considered as a combination of setting (II) and (III).

5.2. Choices of the target

A typical microarray experiment delivers an informative scene as follow: after some "tendious" preprocessing routines we receive the mRNA-abundant measurements in form of real values, contain mostly in two groups, one for the control and one for the treatment. Dependent on what kinds of biological host we're dealing with, those values contains for thousands to ten thousands genes.
Let's denote x_{ij} and y_{ij} the j-th measurement of gene i from the control group and treatment group, respectively. $(i = 1..P, j = 1..N)$
We chosen the number of genes P = 5000 and the number of samples N = 3 for our simulations same as the simulation which be used in Opgen-Rhein and Strimmer (2007). Of course each group could have an own sample size, for convenience we assume that two groups have the same sample size N. The case with two different sample sizes could be analog derived.
The sample mean of each group is defined as follow

$$\overline{x}_i = \frac{1}{N} \sum_k^N x_{ik}$$

$$\overline{y}_i = \frac{1}{N} \sum_k^N y_{ik}$$

The unbiased sample variances of two groups are v_i and w_i respectively

$$\hat{v}_i = \frac{1}{N-1}\sum_{k}^{N}(x_{ik} - \overline{x}_i)^2$$
$$\hat{w}_i = \frac{1}{N-1}\sum_{k}^{N}(y_{ik} - \overline{y}_i)^2$$

Because of the flexibility of shrinkage rule, we're trying not to limit any possibilities of extending the rule. All of evaluating constructions below are merely our ideas how to choose wisely the target of the shrinkage rule. Any another freely ideas could be considered as well, but as we mentioned in the construction part of the James-Stein rule, a target could be wisely or badly chosen, and although the new rule has his mechanism to automatically prevent over- or undershrink, a wisely chosen construction for the target should be always taken place before.

1. This case assumes that two groups of microarray data have for each group a distinctive variance (same as the $unequal$-variance case of the ordinary t-test), therefore for each group exists a separate target (multiple targets), and for each new estimated variance exists a distinct shrinkage intensity λ (multiple shrinkage intensities). Let's denote the targets as

$$v_{target} = median(\hat{v}_1, \hat{v}_2, ..., \hat{v}_N)$$
$$w_{target} = median(\hat{w}_1, \hat{w}_2, ..., \hat{w}_N)$$

The shrinkage intensity for each group was defined as (Equation 24)

$$\hat{\lambda}_1 = \frac{\frac{N}{(N-1)^3}\sum_{i=1}^{P}\sum_{k=1}^{N}\left((x_{ik} - \overline{x}_i)^2 - \frac{1}{N}\sum_{j=1}^{N}(x_{ij} - \overline{x}_i)^2\right)^2}{\sum_{i=1}^{P}(\hat{v}_i - v_{target})^2}$$
$$\hat{\lambda}_2 = \frac{\frac{N}{(N-1)^3}\sum_{i=1}^{P}\sum_{k=1}^{N}\left((y_{ik} - \overline{y}_i)^2 - \frac{1}{N}\sum_{j=1}^{N}(y_{ij} - \overline{y}_i)^2\right)^2}{\sum_{i=1}^{P}(\hat{w}_i - w_{target})^2}$$

The new improved estimated variance for each group was defined as (Equation 25)

$$v_i^\star = (1 - \hat{\lambda}_1)\hat{v}_i + \hat{\lambda}_1 v_{target}$$
$$w_i^\star = (1 - \hat{\lambda}_2)\hat{w}_i + \hat{\lambda}_2 w_{target}$$

The shrinkage t-statistic for detecting differential-express of gene i is simple (Equation 26)

$$t_i^{(1)} = \frac{\overline{x}_i - \overline{y}_i}{\sqrt{\frac{v_i^\star + w_i^\star}{N}}}$$

After our inspection in Opgen-Rhein and Strimmer (2007) work, this case turns out to be their default option of calculating the shrinkage variance when two groups have different variances in the st-package, which was programmed in R-Framework.

2. This case uses the same assumption as the first case, instead of using different targets, we use the $same$ target for two shrinkage variances. This target was chosen as

$$v_{target} = median(\hat{v}_1, \hat{v}_2, ..., \hat{v}_N, \hat{w}_1, \hat{w}_2, ..., \hat{w}_N)$$

Others terms are defined as the above case.

3. In this case, the idea is to treat two groups as one group, and we define a new variable as follow

$$\hat{u}_i = \frac{\hat{v}_i + \hat{w}_i}{2} \tag{36}$$

The target was chosen with the new variable

$$u_{target} = median(\hat{u}_1, \hat{u}_2, ..., \hat{u}_N)$$

The multiple shrinkage intensities was eliminated in this case, turn out to have one target with one shrinkage intensity. Therefore the data of two groups x_{ij} and y_{ij} need to combine to one group z_{ij} as follow

$$z_{ij} = x_{ij} \forall i = 1..P, \forall j = 1..N$$
$$z_{ij} = y_{i,(j-N)} \forall i = 1..P, \forall j = (N+1)..2N$$

Let's denote M = 2N, the new dataset has the new sample mean and unbiased sample variance as well

$$\overline{z}_i = \frac{1}{M} \sum_k^M z_{ik}$$
$$\hat{z}_i = \frac{1}{M-1} \sum_k^M (z_{ik} - \overline{z}_i)^2$$

The $shared$ shrinkage intensity was calculated as

$$\hat{\lambda} = \frac{\frac{M}{(M-1)^3} \sum_{i=1}^P \sum_{k=1}^M \left((z_{ik} - \overline{z}_i)^2 - \frac{1}{M} \sum_{j=1}^M (z_{ij} - \overline{z}_i)^2 \right)^2}{\sum_{i=1}^P (\hat{u}_i - u_{target})^2}$$

The shrinkage variance and the shrinkage t-statistic for this case are simply analog

$$u_i^\star = (1 - \hat{\lambda}) \hat{u}_i + \hat{\lambda} u_{target}$$
$$t_i^{(3)} = \frac{\overline{x}_i - \overline{y}_i}{\sqrt{\frac{2u_i^\star}{N}}}$$

4. This case uses the same idea of common target of the 2. case. The target was chosen in this case is however the u_{target} from the 3. case. That means we have the multiple shrinkage intensities as follow

$$\hat{\lambda}_1 = \frac{\frac{N}{(N-1)^3}\sum_{i=1}^{P}\sum_{k=1}^{N}\left((x_{ik}-\overline{x}_i)^2 - \frac{1}{N}\sum_{j=1}^{N}(x_{ij}-\overline{x}_i)^2\right)^2}{\sum_{i=1}^{P}(\hat{v}_i - u_{target})^2}$$
$$\hat{\lambda}_2 = \frac{\frac{N}{(N-1)^3}\sum_{i=1}^{P}\sum_{k=1}^{N}\left((y_{ik}-\overline{y}_i)^2 - \frac{1}{N}\sum_{j=1}^{N}(y_{ij}-\overline{y}_i)^2\right)^2}{\sum_{i=1}^{P}(\hat{w}_i - u_{target})^2}$$

The shrinkage variances were obvious separately calculated

$$v_i^\star = (1-\hat{\lambda}_1)\hat{v}_i + \hat{\lambda}_1 u_{target}$$
$$w_i^\star = (1-\hat{\lambda}_2)\hat{w}_i + \hat{\lambda}_2 u_{target}$$

The shrinkage t-statistic for detecting differential-express of gene i is analog (but obviously not the same) with the 1. case

$$t_i^{(4)} = \frac{\overline{x}_i - \overline{y}_i}{\sqrt{\frac{v_i^\star + w_i^\star}{N}}}$$

5. This case uses the assumption that the two groups come with the same variance. We discovered that the st-package from Opgen-Rhein and Strimmer (2007) uses the same approach for compute the shrinkage t-statistic for the same variance case.
This case uses the combined dataset z_{ij} from the 3. case. The target was chosen as the median of all unbiased sample variances of z.

$$z_{target} = median(\hat{z}_1, \hat{z}_2, ..., \hat{z}_N)$$

The $combined$ shrinkage intensity was also calculated completely from z

$$\hat{\lambda} = \frac{\frac{M}{(M-1)^3}\sum_{i=1}^{P}\sum_{k=1}^{M}\left((z_{ik}-\overline{z}_i)^2 - \frac{1}{M}\sum_{j=1}^{M}(z_{ij}-\overline{z}_i)^2\right)^2}{\sum_{i=1}^{P}(\hat{z}_i - z_{target})^2}$$

The $combined$ shrinkage variance and the shrinkage t-statistic for this case are simply straightforward

$$z_i^\star = (1-\hat{\lambda})\hat{z}_i + \hat{\lambda} z_{target}$$
$$t_i^{(5)} = \frac{\overline{x}_i - \overline{y}_i}{\sqrt{\frac{2z_i^\star}{N}}}$$

5.3. Methods of evaluation

There are some properties of the James-Stein shrinkage rule when applied in microarray study that we're interested in

i. How well could the new rule do in term of estimating the variance?

ii. Under the assumption of a true null hypothesis, which distribution deliver the shrinkage t-statistic? What is the differences of this new distribution in comparing with the standard t-distribution?

iii. In case that the new distribution is a t-distribution, which degrees of freedom has the distribution and that degrees could be estimated as well?

To answer the i.question, one could use many methods for evaluating as well. We decided to measure the distance of the new estimated variance with the true variance since we know the true variance from simulations. Because there are 5 cases to estimate the variance, the score function for each case should be well defined as well. In some cases which we have two targets to measure (two variances in the 1. case for example), some other cases which we have merely one target (one shared variance in the 3. case). Therefore we'll define two score functions to cover the interested question for every case. Let's denote σ_1^2, σ_2^2 are the true known variances ($\sigma_1^2 = \sigma_2^2$ for the one variance case) and $\sigma^2 = \frac{\sigma_1^2+\sigma_2^2}{2}$. We define for each case from the above 5.cases two score functions as follow

(1) This case has two estimated variances for each gene i: $v_i^\star$ and $w_i^\star$

$$score^{(1)} = \sum_i^P \left((v_i^\star - \sigma_1^2)^2 + (w_i^\star - \sigma_2^2)^2\right)$$
$$\widetilde{score}^{(1)} = \sum_i^P \left(\frac{v_i^\star + w_i^\star}{2} - \sigma^2\right)^2$$

(2) This case has also two estimated variances for each gene i: $v_i^\star$ and $w_i^\star$

$$score^{(2)} = \sum_i^P \left((v_i^\star - \sigma_1^2)^2 + (w_i^\star - \sigma_2^2)^2\right)$$
$$\widetilde{score}^{(2)} = \sum_i^P \left(\frac{v_i^\star + w_i^\star}{2} - \sigma^2\right)^2$$

(3) This case has merely one estimated variance $u_i^\star$ for each gene i. Therefore the first score function could not be defined.

$$\widetilde{score}^{(3)} = \sum_i^P \left(u_i^\star - \sigma^2\right)^2$$

(4) This case is similar to 1. and 3.case, has also two estimated variances for each gene i: $v_i^\star$ and $w_i^\star$

$$score^{(4)} = \sum_i^P \left((v_i^\star - \sigma_1^2)^2 + (w_i^\star - \sigma_2^2)^2\right)$$
$$\widetilde{score}^{(4)} = \sum_i^P \left(\frac{v_i^\star + w_i^\star}{2} - \sigma^2\right)^2$$

(5) This case has again merely one estimated variance $z_i^\star$ for each gene i.

$$\widetilde{score}^{(5)} = \sum_i^P \left(z_i^\star - \sigma^2\right)^2$$

5.4. Result

6. Discussion

ACKNOWLEDGMENT

I would like to give many thanks to Yvonne Poeschl, Markus Boenn for many helpful discussions, to Ioana Lemmian for reading the drafting script. My special thank to Prof. Dr. Ivo Grosse, who supervised this project.

A. Fundamental statistic

This section provides a few fundamental concept in the statistic, which are crucial to understanding any proof given in this paper.

A.1. Expected value and variance

Some important properties of expected value and variance are related to this paper will be shown here.
Suppose X, Y are independent random variables.

$$E(\alpha X \pm \beta Y) = \alpha E(X) \pm \beta E(Y) \tag{37}$$
$$Var(\alpha X \pm \beta Y) = \alpha^2 Var(X) + \beta^2 Var(Y) \tag{38}$$

A.2. Sum of independent random variables

- Suppose X, Y are independent normal distributed random variables.

$$X \sim \mathcal{N}(\mu_X, \sigma_X^2)\,,\, Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$$
$$\Rightarrow Z = X \pm Y \sim \mathcal{N}(\mu_X \pm \mu_Y, \sigma_X^2 + \sigma_Y^2) \tag{39}$$

- Suppose X, Y are independent chi-squared distributed random variables.

$$X \sim \chi^2_{\nu_X}\,,\, Y \sim \chi^2_{\nu_Y}$$
$$\Rightarrow Z = X + Y \sim \chi^2_{\nu_X + \nu_Y} \tag{40}$$

A.3. Student's t-distribution

Some statistical characterizations which are related to Student's t-distribution will be mentioned.
Student's t-distribution is the probability distribution of the ratio

$$t = \frac{Z}{\sqrt{V/\nu}} \tag{41}$$

where

- Z $\sim \mathcal{N}(0,1)$
- V $\sim \chi^2_\nu$
- Z and V are statistical independence.

then the above ratio has a t-distribution with ν degrees of freedom.

A.4. Fisher's F-distribution

The F-distribution is the ratio of two chi-squared variates

$$F = \frac{S_1/d_1}{S_2/d_2} \tag{42}$$

where

- S_1 and S_2 have chi-squared distributions with d_1 and d_2 degrees of freedom respectively, and
- S_1 and S_2 are statistical independence.

B. Cochran's theorem

This section proves the Cochran's theorem in statistic, which is crucial to this paper. Cochran had proved in 1934 the central case of the theorem, the most important case in practice. Madow (1940) expanded the proof for the non-central case.
This theorem is very helpful for quadratic forms of random variables.

Cochran's theorem. Suppose X_n (n = 1 ... N) are independent normal distributed random variables and Q_k (k = 1 ... K) have quadratic form in $\mathbb{R}^N$ with:

$$(a) \sum_{n=1}^{N} X_n^2 = \sum_{k=1}^{K} Q_k$$

and

$$(b)\, n_k \,\text{is the rank of}\, Q_k$$

Then we have:
i) If $\sum_{k=1}^{K} n_k = N$, then the Q_k are independent, and each Q_k has a chi-square distribution with n_k degrees of freedom.
ii) If the Q_k are independent, and each Q_k has a chi-square distribution with n_k degrees of freedom, then $\sum_{j=1}^{K} n_j = N$

Proof. The proof of this theorem will not be shown here, but rather refer the reader to another literature (Schach, 1978).

To understand the conditions (a) and (b) of the theorem, let have a simple example:
Suppose we have N independent normal distributed random variables X_n (n = 1 ... N). Then with $Q_1 = \sum_{n=1}^{N-1} X_n^2$ and $Q_2 = X_N^2$ we'll have:

$$\sum_{n=1}^{N} X_n^2 = Q_1 + Q_2$$

Cochran's theorem states that Q_1 and Q_2 are independent, chi-square distributed with $N-1$ and 1 degrees of freedom respectively.

□

C. Fisher's theorem

It's maybe helpful for readers who never heard of Cochran's theorem, we'll also introduce the Fisher's theorem, which is known as the converse of Cochran's theorem.

Theorem C.1. (Fisher's Theorem) *Let A be a sum of squares of N independent normal standardized variates X_i, and suppose $A = Q_1 + Q_2$ where Q_1 is a quadratic form in the X_i, distributed as chi-squared with H degrees of freedom. Then Q_2 is distributed as χ^2 with $(N-H)$ degrees of freedom and is independent of Q_1.*

D. Distribution of sample mean $\overline{\mu}$ and sample variance $\hat{\sigma}^2$

Suppose X_n (n = 1,...,N) are independent normally distributed random variables with mean μ and standard deviation σ.

Sample mean and sample variance of X_n (n = 1 ... N) are defined by following:

$$\overline{\mu} = \frac{1}{N}\sum_{n=1}^{N} X_n \tag{43}$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}(X_n - \overline{\mu})^2 \tag{44}$$

It's simple to show that: $\sum_{n=1}^{N} X_n \sim \mathcal{N}(N\mu, N\sigma^2)$

$$\Rightarrow \overline{\mu} \sim \mathcal{N}(\mu, \frac{\sigma^2}{N}) \tag{45}$$

$$\Rightarrow \frac{(\overline{\mu}-\mu)\sqrt{N}}{\sigma} \sim \mathcal{N}(0,1) \tag{46}$$

We have:

$U_n = \frac{X_n-\mu}{\sigma} \sim \mathcal{N}(0,1)\,\forall n.$

It is possible to write:

$$\sum_{n=1}^{N} U_n^2 = \sum_{n=1}^{N} \left(\frac{X_n - \mu}{\sigma} \right)^2 = \sum_{n=1}^{N} \left(\frac{X_n - \overline{\mu} + \overline{\mu} - \mu}{\sigma} \right)^2$$
$$= \sum_{n=1}^{N} \left(\frac{X_n - \overline{\mu}}{\sigma} \right)^2 + \sum_{n=1}^{N} \left(\frac{\overline{\mu} - \mu}{\sigma} \right)^2 + \sum_{n=1}^{N} 2 \left(\frac{(X_n - \overline{\mu})(\overline{\mu} - \mu)}{\sigma^2} \right)$$

The second term $= N(\frac{\overline{\mu}-\mu}{\sigma})^2$
The third term

$$= 2 \sum_{n=1}^{N} \left(\frac{X_n \overline{\mu} - X_n \mu - \overline{\mu}^2 + \overline{\mu}\mu}{\sigma^2} \right) = 2 \left(\frac{\sum X_n \overline{\mu} - \sum X_n \mu - N\overline{\mu}^2 + N\overline{\mu}\mu}{\sigma^2} \right)$$
$$= 2 \left(\frac{N\overline{\mu}^2 - N\overline{\mu}\mu - N\overline{\mu}^2 + N\overline{\mu}\mu}{\sigma^2} \right) = 0$$

so

$$\sum_{n=1}^{N} U_n^2 = \sum_{n=1}^{N} (\frac{X_n - \overline{\mu}}{\sigma})^2 + N(\frac{\overline{\mu} - \mu}{\sigma})^2 = Q_1 + Q_2 \tag{47}$$

Cochran's theorem states that Q_1 and Q_2 are independent, chi-square distributed with $N-1$ and 1 degrees of freedom respectively.

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{n=1}^{N} (X_n - \overline{\mu})^2 = Q_1 \frac{\sigma^2}{N}$$
$$\Rightarrow N \frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{N-1} \tag{48}$$

The unbiased estimator of sample variance:

$$\hat{\sigma}^2_{N-1} = \frac{1}{N-1} \sum_{n=1}^{N} (X_n - \overline{\mu})^2 = Q_1 \frac{\sigma^2}{N-1}$$
$$\Rightarrow (N-1) \frac{\hat{\sigma}^2_{N-1}}{\sigma^2} \sim \chi^2_{N-1} \tag{49}$$

This shows that the sample mean and sample variance are independent; and this property characterizes the normal distribution. No other distribution have the sample mean and sample variance are independent.

References

Boldrick, J. C., Alizadeh, A. A., Diehn, M., Dudoit, S., Liu, C. L., Belcher, C. E., Botstein, D., Staudt, L. M., Brown, P. O., and Relman, D. A. (2002). Stereotyped and specific gene expression programs in human innate immune responses to bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 99(2):972–7.

Dudoit, S., Shaffer, J., and Boldrick, J. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18(1):71–103.

Efron, B., Morris, C., Efron, B., and Morris, C. (1973). Stein's estimation rule and its competitors–an empirical bayes approach. *Journal of the American Statistical Association*.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–7.

Hausser, J. and Strimmer, K. (2009). Entropy inference and the james-stein estimator, with application to nonlinear gene association networks. *J Mach Learn Res*, 10:1469–1484.

Kadota, K., Nakai, Y., and Shimizu, K. (2008). A weighted average difference method for detecting differentially expressed genes from microarray data. *Algorithms for molecular biology : AMB*, 3:8.

Ledoit, O. and Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10(5):603–621.

Lindley, D. V., Smith, A. F. M., Lindley, D. V., and Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society. Series B (Methodological)*.

Madow, W. G. (1940). Limiting distributions of quadratic and bilinear forms. *Ann. Math. Statist.*, 11(2):125–146.

Opgen-Rhein, R. and Strimmer, K. (2007). Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. *Stat Appl Genet Mol Biol*, 6:Article9.

Schach, S. (1978). *Regressions- und Varianzanalyse, Eine Einfuehrung*. Springer.

Schafer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications .... *Statistical applications in genetics and* ....

Stein, C. (1955). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley symposium on mathematical statistics and probability*, 1:197–206.

Thompson, J. R. and Thompson, J. R. (1968). Some shrinkage techniques for estimating the mean. *Journal of the American Statistical Association*.

Tibshirani, R. (2007). A comparison of fold-change and the t-statistic for microarray data analysis. *stanford.edu*.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, 98(9):5116.

Wikipedia (2010a). Affymetrix. `http://en.wikipedia.org/wiki/Affymetrix`.

Wikipedia (2010b). Dna microarray. `http://en.wikipedia.org/wiki/DNA_microarray`.

Zhang, A. (2007). *Advanced analysis of gene expression microarray data*. World Scientific.