Shrinkage Approach for Gene Expression Data Analysis

Jiří Haman[1,2], **Zdeněk Valenta**[1]

[1]Institute of Computer Science AS CR, v. v. i., Department of Medical Informatics and Biostatistics, Prague, Czech Republic

[2]First Faculty of Medicine, Charles University in Prague, Czech Republic

Abstract

Background: Microarray technologies are used to measure the simultaneous expression of a certain set of thousands of genes based on ribonucleic acid (RNA) obtained from a biological sample. We are interested in several statistical analyses such as 1) finding differentially expressed genes between or among several experimental groups, 2) finding a small number of genes allowing for the correct classification of a sample in a certain group, and 3) finding relations among genes.
Objectives: Gene expression data are high dimensional, and this fact complicates their analysis because we are able to perform only a few samples (e.g. the peripheral blood from a limited number of patients) for a certain set of thousands of genes. The main purpose of this paper is to present the shrinkage estimator and show its application in different statistical analyses.
Methods: The shrinkage approach relates to the shift of a certain value of a classic estimator towards a certain value of a specified target estimator. More precisely, the shrinkage estimator is the weighted average of the classic estimator and the target estimator.
Results: The benefit of the shrinkage estimator is that it improves the mean squared error (MSE) as compared to a classic estimator. The MSE combines the measure of an estimator's bias away from its true unknown value and the measure of the estimator's variability. The shrinkage estimator is a biased estimator but has a lower variability.
Conclusions: The shrinkage estimator can be considered as a promising estimator for analyzing high dimensional gene expression data.

Keywords

Microarray technology, high dimensional data, mean squared error, James-Stein shrinkage estimator, mutual information

Correspondence to:

Jiří Haman
Institute of Computer Science AS CR, v. v. i.
Address: Pod Vodárenskou věží 271/2, Prague 8, Czech Republic
E–mail: j.haman@seznam.cz, haman@cs.cas.cz

EJBI 2013; 9(3):2–8
received: August 1, 2013
accepted: October 31, 2013
published: November 20, 2013

1 Introduction

In this section we describe microarray technology, and mention several types of hypotheses for microarrays together with some problems which arise with microarray data.

1.1 Microarray Technology

Microarray technology is an important element in gene expression assessment. Basically, we can describe microarray as a solid wafer which contains from a few thousand to millions of one-stranded segments of deoxyribonucleic acid (DNA) nucleotides in precisely set positions. The positions are often called spots or probes. These spots correspond to individual genes. The mode of microarray production is described e.g. in [1].

Gene expression measurement is conducted by extracting ribonucleic acid (RNA) from a biological sample, e.g. the peripheral blood of a patient or a tuberculous tissue. The extracted RNA sample is marked with a fluorescent dye and spread on a microarray chip. During a process called hybridization the RNA sample binds with the microarray thanks to the complementarity of nucleotides.

The RNA concentration on a microarray is determined using fluorescence. The microarray is scanned using a special scanner and the output has the form of a two-dimensional scanned image. The scanned image is formed by pixels with certain intensities. A group of neighbouring pixels making up the spot corresponding to the fluorescent activity of a specific gene. The spots must be targeted using a specialized image analysis software.

After performing complicated preprocessing techniques, we obtain summarized values of gene expression for each spot. For more information about the techniques we refer the reader to [2] or [3].

Because we have usually several comparable biological samples, the final data of gene expressions are in the form of matrix. The matrix's rows typically correspond to individual genes and the columns correspond to individual samples (e.g. patients) of gene expression profiles from several microarrays.

1.2 Types of Analysis for Microarrays

Several types of statistical inferences are possible in the context of information extracted from microarrays:

(a) detection of differentially expressed genes between/among several experimental sample groups (e.g. comparing gene expression of patients with a tumour and without a tumour),

(b) classification of an unknown sample in a specific group based on the gene expression profile in the sample (e.g. classification of a patient with cancer, i.e. confirmation of the onset of cancer),

(c) discovery of new/unknown groups of genes (e.g. new tumour subtype),

(d) identification of genes which are important for a given group (e.g. for myocardial infarction),

(e) creation of gene networks (description of relations among/between genes in evolution of cancer).

A broad overview of statistical analysis for gene expression data analysis can be found in [3] or [4].

1.3 Problem of High Dimension

The problem with certain kinds of microarray technology, e.g. whole-genome microarrays, is high dimensionality due to the fact that there is a greater number of variables (genes) than observations (samples). If we consider p to be the number of variables and n to be the number of observations this means $n \ll p$. Problems associated with high dimensionality include the following, for example:

- the selection of relevant differentially expressed genes, i.e. the problem of multiple testing (see [5]),
- tractability of linear discriminant analysis (LDA) used for classification purposes. LDA requires an estimator of the inverse covariance matrix, which, however, does not exist for the data which we have in this case (see [6]).

2 Shrinkage Approach

In this section we present a shrinkage approach which can help deal with the high dimensionality of gene expression data. The approach results in an estimator based on reducing the mean squared error (MSE) compared to a classic estimator.

Firstly, we present the origin of the shrinkage approach, i.e. James-Stein estimator (JSE) which is motivated by the estimation of an unknown mean vector for multivariate normal distribution based on 1 observation from this distribution. Next, we mention the optimal value of the JSE's parameter which leads to the lowest value of the MSE. This optimal value also gives the estimator the shrinkage attribute. Finally, a generalization for both the JSE and its optimal value is mentioned.

2.1 Mean Squared Error (MSE)

Let us suppose we have a random sample from a nondegenerated distribution with an unknown one-dimensional parameter $\theta \in \mathbb{R}$. We calculate an estimator $\hat{\theta} \in \mathbb{R}$ on the basis of our data. Then the MSE (or quadratic risk) for the parameter $\hat{\theta}$ is defined as

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{E}(\hat{\theta} - \theta)^2. \tag{1}$$

Expression (1) can be rewritten in a different way. We expand $(\hat{\theta} - \theta)^2$ in (1) and "add-subtract" $(\mathrm{E}\hat{\theta})^2$. After using $\mathrm{var}(\hat{\theta}) = \mathrm{E}(\hat{\theta})^2 - (\mathrm{E}\hat{\theta})^2$ and $\theta = \mathrm{E}\theta$ and $\mathrm{E}\theta^2 = (\mathrm{E}\theta)^2$ we get an equivalent form of (1)

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{var}(\hat{\theta}) + (\mathrm{Bias}(\hat{\theta}, \theta))^2. \tag{2}$$

In (2) the term $\mathrm{Bias}(\hat{\theta}, \theta) = \mathrm{E}(\hat{\theta} - \theta)$ represents the accuracy of estimator $\hat{\theta}$ with respect to the unknown parameter θ.

We can see from (2) that (1) for MSE of estimator $\hat{\theta}$ is the sum of its variance and the square of its bias. If $\mathrm{Bias}(\hat{\theta}, \theta) = 0$ we say that estimator $\hat{\theta}$ is the unbiased estimator of parameter θ. If $\mathrm{Bias}(\hat{\theta}, \theta) \neq 0$ we say that estimator $\hat{\theta}$ is the biased estimator of parameter θ.

The idea for introducing the MSE is that allowing a small bias for estimator $\hat{\theta}$ of parameter θ can substantially improve its variability $\mathrm{var}(\hat{\theta})$. For more details we refer the reader to [7].

2.2 James-Stein Estimator (JSE)

The JSE, also known as the Stein estimator or shrinkage estimator, is an estimator which first appeared in [8]. The motivation for this estimator is the following.

Let us assume that $p \in \mathbb{N}$ and we have realization of random vector $X = (X_1, \ldots, X_p) \in \mathbb{R}^p$. Let vector X follow p-dimensional normal distribution with unit covariance matrix I and unknown mean vector $\mu = (\mu_1, \ldots, \mu_p) \in \mathbb{R}^p$. The task is to estimate unknown vector μ.

One estimator of μ is called the maximum likelihood estimator (MLE). We can say it is a classic estimator because it means that we estimate μ by realization x of X. This estimator is in the form

$$\hat{\mu}^{\mathrm{MLE}} = X$$

Notation is in the scalar form, i.e. each component μ_i of vector μ is estimated using the value of the respective component X_i of vector X, $i = 1, \ldots, p$.

Another μ estimator in which we are particularly interested is called the James-Stein estimator (JSE). This estimator is in the form

$$\hat{\mu}^{\mathrm{JSE}} = (1 - \theta_p(X))X$$

where

$$\theta_p(X) = \frac{p-2}{\|X\|^2}, \qquad \|X\|^2 = \sum_{i=1}^{p} X_i^2.$$

For $p > 2$ we can see that the shrinkage factor $\theta_p(X)$ is positive because the numerator and denominator are always positive.

The reason JSE is called the shrinkage estimator is that we implicitly suppose that $\theta_p(X) \in (0, 1)$. In this case each component X_i of vector X from MLE is proportionally shrunken by the same constant $1 - \theta_p(X)$ closer to respective zero value component of vector X.

It can also be observed that in MLE we use "pure" information represented by the individual component while in JSE we borrow information from the individual component together with information contained in all the components.

2.3 MSE of JSE

Introducing the shrinkage factor $\theta_p(X)$ results in a lower MSE. More precisely, we have some kind of norm for MSE because we are dealing with estimation of an unknown p-dimensional parameter μ. In [8] Stein uses

$$\mathrm{MSE}_p(\hat{\mu}) = \mathrm{E}(\hat{\mu} - \mu)^T(\hat{\mu} - \mu) \tag{3}$$

where μ is the true value of the mean and $\hat{\mu}$ is the estimator of μ based on observed data. Expression (3) for MSE_p can be further rewritten as

$$\mathrm{MSE}_p(\hat{\mu}) = \mathrm{E}\left(\sum_{i=1}^{p} (\hat{\mu}_i - \mu_i)^2\right). \tag{4}$$

Thus, MSE in case of estimation for p-dimensional parameter μ is the sum of MSEs for individual components of μ. According to the expression (2) we can rewrite (4) as

$$\mathrm{MSE}_p(\hat{\mu}) = \sum_{i=1}^{p} \left\{ \mathrm{var}(\hat{\mu}_i) + (\mathrm{Bias}(\hat{\mu}_i, \mu_i))^2 \right\}. \tag{5}$$

Thus, the MSE of estimator $\hat{\mu}$ equals the sum of variances and squared biases of its individual components. Let's demonstrate the computation of the MSE for MLE and JSE.

1) If $\hat{\mu}$ is MLE then

$$\mathrm{Bias}(\hat{\mu}_i^{\mathrm{MLE}}, \mu_i) = \mathrm{E}(X_i - \mu_i) = 0$$

where we derive benefit from $\mathrm{E}X_i = \mu_i$. Thus, the MLE is unbiased estimator. Similarly, we have

$$\mathrm{var}(\hat{\mu}_i^{\mathrm{MLE}}) = \mathrm{var}(X_i) = 1.$$

This implies according to (5) that $\mathrm{MSE}_p(\hat{\mu}^{\mathrm{MLE}}) = p$.

2) If $\hat{\mu}$ is JSE then

$$\begin{aligned} \mathrm{Bias}(\hat{\mu}_i^{\mathrm{JSE}}, \mu_i) &= \mathrm{E}((1 - \theta_p(X))X_i - \mu_i) \\ &= -\mathrm{E}(\theta_p(X)X_i) \neq 0 \end{aligned}$$

where we use $\mathrm{E}X_i = \mu_i$. Thus, JSE is biased estimator in contrast to the unbiased MLE.

Computation of the MSE for JSE is not as straightforward as for the MLE. So we restrict ourselves to stating that

$$\mathrm{MSE}_p(\hat{\mu}^{\mathrm{JSE}}) = p - (p-2)^2 \mathrm{E}\left(\frac{1}{\|X\|^2}\right).$$

We can see that for $p > 2$

$$\mathrm{MSE}_p(\hat{\mu}^{\mathrm{JSE}}) < \mathrm{MSE}_p(\hat{\mu}^{\mathrm{MLE}})$$

because $\|X\|^2 > 0$. Especially, for large p we can achieve a great improvement of MSE for JSE in contrast to MSE for MLE. However, the "price" we pay for lower MSE of JSE is bias.

2.4 The Optimal Shrinkage Factor

We can also compute the optimal shrinkage factor which guarantees the lowest possible MSE. This is given by maximizing each summand in $\mathrm{MSE}_p(\hat{\mu}^{\mathrm{JSE}})$. If the shrinkage factor $\theta_p(X)$ in $\hat{\mu}^{\mathrm{JSE}}$ is denoted as φ then according to (5) we have

$$\sum_{i=1}^{p} \left\{ (1-\varphi)^2 \mathrm{var}(X_i) + (\mathrm{E}((1-\varphi)X_i - \mu_i))^2 \right\}$$

which is maximized by taking derivative with respect to shrinkage factor φ. This leads to the optimal shrinkage factor

$$\varphi^\star = \left(\sum_{i=1}^{p} \mathrm{var}(X_i)\right) \Big/ \left(\sum_{i=1}^{p} \mathrm{E}X_i^2\right)$$

where we use $\mathrm{var}(X_i) + (\mathrm{E}X_i)^2 = \mathrm{E}X_i^2$.

2.5 The General Form of JSE

We have introduced only the basic version of JSE. If we again suppose p-dimensional normality for random vector X as in the subsection 2.2, the general form of JSE can be written as

$$\hat{\theta}^{\text{shrink}} = (1-\lambda)\hat{\theta} + \lambda\hat{\theta}^{\text{target}},$$

i.e. the shrinkage estimator $\hat{\theta}^{\text{shrink}}$ is the weighted average of the classic estimator $\hat{\theta}$ and the target estimator $\hat{\theta}^{\text{target}}$ chosen by us. In case of basic JSE $\hat{\theta}$ is $\hat{\mu}^{\text{MLE}}$ and $\hat{\theta}^{\text{target}}$ is the vector of p zeros.

The constant $\lambda \in (0,1)$ is the shrinkage factor, i.e. the weight which is borrowed from the classic estimator and obtained by target estimator. In case of the basic JSE we have $\lambda = \theta_p(X)$.

The estimator $\hat{\theta}^{\text{shrink}}$ is useful because it results in a lower MSE previously represented by (5).

We can compute the optimal shrinkage factor which minimizes the MSE. In case of unbiased classic estimator and nonrandom target estimator we have the optimal shrinkage factor with respect to the MSE in the form

$$\lambda^{\star} = \left\{\sum_{i=1}^{p} \text{var}(\hat{\theta}_i)\right\} \Big/ \left\{\sum_{i=1}^{p} \text{E}(\hat{\theta}_i - \hat{\theta}_i^{\text{target}})^2\right\}. \qquad (6)$$

For the optimal shrinkage factor where the target estimator is random and the classic estimator is biased we refer the reader to [9].

2.6 Several Generalizations of JSE

In [10] James and Stein show that JSE has a lower MSE for an arbitrary constant $c \in (0, 2(p-2))$ which is used in the nominator of $\theta_p(X)$. In [11] Baranchik observes that the shrinkage parameter $1-\theta_p(X)$ can be negative and considers positive part of JSE for improving the MSE, i.e.

$$\hat{\mu}^{\text{JSE+}} = (1-\theta_p(X))^{+}X \qquad (7)$$

where $t^+ = t$ for $t \geq 0$ and $t^+ = 0$ in other cases.

In [12] Bock expands JSE of μ for a situation when vector X from p-dimensional normal distribution has an arbitrary known or unknown positive definite covariance matrix V. According to (7), general JSE can be written in the form

$$\hat{\mu}^{\text{JSE+},V} = \left(1 - \frac{\hat{p}-2}{X^T V^{-1} X}\right)^{+} X \qquad (8)$$

where V^{-1} is inverse of the covariance matrix V and $\hat{p}$ is effective dimension given by the trace of matrix V divided by the maximum eigenvalue of V. For $\hat{p} > 2$ the JSE has a lower MSE than the MLE. Bock also shows that substituting of $\hat{p}-2$ in the nominator of the fraction in (8) by an arbitrary constant $\hat{c} \in (0, 2(\hat{p}-2))$ leads to a lower MSE for the JSE than for MLE. For more information about the generalization of the JSE we refer the reader to [7].

3 Applications of the Shrinkage Approach

In the subsection 1.2 we described several types of hypotheses in the context of microarrays. In this section we present shrinkage approach as a solution for some of them.

We desribe shrinkage version of the clustering algorithm K-means method corresponding to the hypothesis (c) or (d), shrinkage version of the t-statistic corresponding to the hypothesis (a), shrinkage version of the mutual estimation corresponding to the hypothesis (e) and shrinkage version of the covariance matrix corresponding to the hypothesis (b) or (e).

3.1 Shrinkage K-Means Method

In [13] Gao and Hitchcock introduce a shrinkage version for K-means clustering algorithm as an improvement of this when $n \ll p$. The method is applied to Saccharomyces cerevisiae yeast gene expression data. The data contain 78 genes, where each gene is supposed to be differentially expressed in exactly one of 5 groups (5 cell cycle phases). The expression of each gene is measured 18 times at 7-minute intervals.

The shrinkage K-means algorithm proceeds as follows. We have n observations divided into K groups, where $K < n$. Each observation has p-dimensional normal distribution with mean vector μ_i and covariance matrix V_i, $i = 1, \dots, K$. We choose randomly K observations which serve as initial estimates for μ_i's, i.e. group centroids. We compute the overall centroid $\overline{X}$ as the overall mean from all group centroids $\overline{X}_i$. Each centroid $\overline{X}_i$ is then shrunken to the overall centroid $\overline{X}$ as

$$\overline{X}_i^{\text{JSE+},V} = \overline{X} + (1-\theta(\hat{p}, V_i))^{+}(\overline{X}_i - \overline{X})$$

where

$$\theta(\hat{p}, V_i) = \frac{\hat{p}-2}{(\overline{X}_i - \overline{X})^T V_i^{-1} (\overline{X}_i - \overline{X})}$$

and $\hat{p}$ is the effective dimension given similarly as in (8) as the trace of matrix V_i divided by the maximum eigenvalue of V_i.

In comparison to the classic K-means algorithm the shrinkage K-means algorithm has better accuracy as given by the Rand Index. The Rand Index measures concordance between the true underlying clustering structure and the result produced by a clustering algorithm. For more details we refer the reader to [13].

3.2 Shrinkage t-Statistic

In [14] Opgen-Rhein and Strimmer introduce a shrinkage version of t-statistics in case of $n \ll p$. Shrinkage is applied to empirical variances $\nu_1, \ldots, \nu_p$ from gene expressions for each of p genes. Then the median ν_{median} from all empirical variances is computed. The shrinkage estimator for ν_k is proposed in the form

$$\nu_k^\star = (1-\lambda)\nu_k + \lambda\nu_{\text{median}}, \tag{9}$$

$k = 1, \ldots, p$, which is the weighted average of the target estimator for variance (ν_{median}) and the classic estimator for variance (ν_k). The optimal shrinkage parameter $\hat{\lambda}^\star$ with respect to the MSE is in the form

$$\hat{\lambda}^\star = \min\left(1, \frac{\sum_{k=1}^p \widehat{\text{var}}(\nu_k)}{\sum_{k=1}^p (\nu_k - \nu_{\text{median}})^2}\right). \tag{10}$$

Estimator (10) differs from estimator (6). Here $\hat{\lambda}^\star$ is composed of a minimum of 1 and the sample estimator of (6). In (6), the numerator and denominator are estimated from data from its sample counterparts. Using the minimum in (10) prevents the shrinkage parameter from "overflowing", i.e. if the estimator of (6) present in (10) is larger than 1 then $\hat{\lambda}^\star = 1$.

Shrinkage t-statistic for comparison of two independent groups of samples is in the form

$$t_k^\star = \frac{\overline{x}_{k1} - \overline{x}_{k2}}{\sqrt{\nu_{k1}^\star/n_1 + \nu_{k2}^\star/n_2}} \tag{11}$$

where $\overline{x}_{k1}$ and $\overline{x}_{k2}$ represent group averages of gene expressions for the k-th gene, $\nu_{k1}^\star$ and $\nu_{k2}^\star$ represent the shrunken group variances of gene expressions for k-th gene and n_1 and n_2 represent the number of samples in each group for the k-th gene.

The shrinkage t-statistic (11) is a compromise between standard t-statistics ($\lambda = 0$ in (9)) and differences of means t-statistics ($\lambda = 1$ in (9)).

The shrinkage t-statistics (11) and several competing methods, such as moderated t-statistics, Efron t-statistics and Cui t-statistics (see [15]), are performed on three gene expression data with different "setups" for variabilities of individual genes (two Affymetrix spike-in studies and one HIV study). The aim is to find the "true discovery rate" of genes, i.e. genes which are known to be truly differentially expressed among all differentially expressed genes from a statistical point of view. The shrinkage *t*-statistic has the best performance. For more details we refer the reader to [14].

3.3 Shrinkage Mutual Information

In [16] Hausser and Strimmer introduce a shrinkage version of mutual information in case of $n \ll p$. The method is applied for constructing association network among genes for Escherichia coli gene expression data. The data consist of 102 known differentially expressed protein coding genes of human superoxid dismutase whose expression is measured at time 0, 8, 15, 22, 45, 68, 90, 150 and 180 minutes after induction by dosage of isopropyl-beta-D-thiogalactopyranoside.

The gene association network is constructed via Algorithm for Reconstruction of Accurate Cellular NEtworks (ARACNE). The algorithm is based on mutual information computed for each pair of genes and model selection is carried out via information processing inequality applied to all gene triplets (see [17]).

Based on whole gene expression data of Escherichia coli, gene expressions for each gene are discretized into K common distinct categories of expression. For each pair of discretized genes we obtain $K \times K$ contingency table.

The mutual information $MI(A, B)$ between discrete random variables A and B (i.e. between discretized expressions of genes) is defined as

$$MI(A,B) = \sum_{i=1}^K \sum_{j=1}^K \theta_{ij}\left(\ln(\theta_{ij}) - \ln(\theta_i\theta_j)\right) \tag{12}$$

where θ_{ij} is the joint relative frequency for the (i, j)-th combination of row category i for the random variable A and column category j for the random variable B in $K \times K$ contingency table. Relative frequencies θ_i and θ_j correspond to marginal relative frequency of i-th row category and j-th column category, respectively. The task is the estimation of θ_{ij}, θ_i and θ_j for (12).

We restrict ourselves to estimation of joint relative frequencies θ_{ij} and especially with respect to (12) represented by the joint Shannon entropy. The joint Shannon entropy is given by

$$H(A,B) = -\sum_{i=1}^K \sum_{j=1}^K \theta_{ij}\ln(\theta_{ij}) \tag{13}$$

and measures the uncertainty associated with the discretized random variables A and B. When $H(A, B)$ is higher, the uncertainty is also higher.

The classic estimator of θ_{ij} in (13) is MLE, i.e.

$$\hat{\theta}_{ij}^{\text{MLE}} = \frac{y_{ij}}{n}$$

where y_{ij}, $i = 1, \ldots, K$, $j = 1, \ldots, K$ is the absolute frequency of the (i, j)-th category in the $K \times K$ contingency table and n is the total sum of absolute frequencies from all cells in the contingency table. The form of $\hat{\theta}_{ij}^{\text{MLE}}$ is based on the assumption of multinomial distribution for cell counts in the $K \times K$ contingency table.

The disadvantage of $\hat{\theta}^{\text{MLE}}$ is that it underestimates (13), leading to a biased estimator of (12). The reason is that the $K \times K$ contingency table is sparse, i.e. the majority of cell frequencies are equal to zero. If a cell with zero frequency that represents zero summand in (13).

We can also estimate relative frequencies by shrinkage estimator in the form

$$\hat{\theta}_{ij}^{\text{shrink}} = (1-\lambda)\hat{\theta}_{ij}^{\text{MLE}} + \lambda t_{ij}$$

where $t_{ij} > 0$ is the (i,j)-th term from the target distribution $\sum_{i=1}^{K}\sum_{j=1}^{K} t_{ij} = 1$. The role of target distribution is to regularize the contingency table, i.e. zero cell counts are "converted" to nonzero counts and this decreases underestimation of (13). Typically, target distribution is chosen as uniform, i.e. $t_{ij} = 1/L$ where L is the number of cells in the $K \times K$ contingency table.

According to equation (6) the optimal shrinkage intensity $\hat{\lambda}^{\star}$ with respect to the MSE is given by

$$\hat{\lambda}^{\star} = \left(\sum_{i=1}^{K}\sum_{j=1}^{K} \widehat{\text{var}}(\hat{\theta}_{ij}^{\text{MLE}})\right) \Big/ \left(\sum_{i=1}^{K}\sum_{j=1}^{K} (t_{ij} - \hat{\theta}_{ij}^{\text{MLE}})^2\right)$$

where the nominator and denominator are estimated without bias from data.

The shrinkage estimator for estimation of entropy has a performance similar to the Nemenman-Shafee-Bialek (NSB) estimator for entropy (see [18]). However, in contrast to the NSB estimator the shrinkage estimator is computationally much faster and fully analytical. For more information we refer the reader to [16].

3.4 The Shrinkage Covariance Matrix

In [9] Schäfer and Strimmer introduce a shrinkage estimator of the population covariance matrix Σ in case of $n \ll p$. The method uses the same Escherichia coli data as in the previous part, related to shrinkage estimation of mutual information. In this case we want to establish the gene network among 102 preselected genes.

The construction of the gene association network is based on a $p \times p$ matrix of partial correlations for gene expression data. Partial correlation measures the strength of the relationship between genes which is free of the influence of other genes. If partial correlation is larger than a certain value (e.g. larger than 0.8) then we can suppose there is an association between the genes.

Values of partial correlations can be computed from values of the inverted covariance matrix. For computing of the inverted covariance matrix Σ^{-1} we need an estimator of the population covariance matrix Σ. Two classic estimators for Σ are MLE and unbiased estimator. In other words, the elements σ_{ij} of Σ are estimated by elements s_{ij} from the sample covariance matrix S where

$$s_{ij} = \frac{1}{\text{df}}\sum_{k=1}^{n}(x_{ik} - \overline{x}_i)(x_{jk} - \overline{x}_j).$$

Here x_{ik} is the expression of the i-th gene in the k-th sample, x_{jk} is the expression of the j-th gene in the k-th sample, $\overline{x}_i$ is the average expression of the i-th gene across all samples and $\overline{x}_j$ is the average expression of the j-th gene across all samples, $i = 1,\dots,p$, $j = 1,\dots,p$, $k = 1,\dots,n$. If $\text{df} = n-1$ we have an unbiased estimator S for Σ. In case of $\text{df} = n$ we have MLE for Σ.

We restrict our attention to the unbiased estimator S. The problem of the estimator S of the covariance matrix is that it becomes singular in case of $n \ll p$ as shown in [19]. Thus, it is not possible to make its sample inversion S^{-1}.

The singularity of matrix S can be eliminated by an estimator based on JSE. The elements σ_{ij} of Σ are estimated by sample covariance matrix with the elements

$$s_{ij}^{\star} = (1-\lambda)s_{ij} + \lambda t_{ij} \tag{14}$$

where $i = 1,\dots,p$, $j = 1,\dots,p$. In equation (14), s_{ij} is the unbiased estimator of σ_{ij}, t_{ij} is an element of target matrix T which is regular and of the same dimension as matrix S. The shrinkage constant λ is supposed to be derived from interval $(0,1)$. An advantage of introducing variant of JSE is not only that it results in a lower MSE but also that it leads to regularization of the unbiased estimator of sample covariance matrix S.

The optimal shrinkage intensity $\hat{\lambda}^{\star}$ for the nonrandom target matrix $T = (t_{ij})$ and the singular unbiased estimate of covariance matrix $S = (s_{ij})$ is

$$\hat{\lambda}^{\star} = \left(\sum_{i=1}^{p}\sum_{j=1}^{p} \widehat{\text{var}}(s_{ij})\right) \Big/ \left(\sum_{i=1}^{p}\sum_{j=1}^{p} \text{E}(s_{ij} - t_{ij})^2\right)$$

where the nominator and denominator are estimated without bias from the data. Schäfer and Strimmer also examine several types of shrinkage targets. They especially pay attention to diagonal covariance target matrix with unequal variances computed from estimate S, i.e. $t_{ij} = s_{ij}$ for $i = j$ otherwise $t_{ij} = 0$. This represents a compromise between simple and complicated estimates. For more information we refer the reader to [9].

In [6] Guo et al. use a shrinkage estimator of covariance matrix for LDA in case of $n \ll p$. The LDA regularized in this way is then combined with the nearest shrunken centroids method (see [20]). Performance of regularized discriminant analysis is tested on nine gene expression data and has for example similar performance as the support vector machines method (see [21]).

4 Conclusion

In this paper we present the shrinkage approach. This is a promising approach for improving gene expression data analysis where the number of genes is much higher than the number of samples.

The shrinkage approach leads to the shrinkage estimator, which combines information from the classic estimator and a specified target estimator through a weighted average of these. The advantage of the shrinkage estimator is a lower MSE than for a classic estimator. The shrinkage estimator is biased but has a substantially lower variability than the classic estimator which is unbiased. This is valid especially for high-dimensional problems.

The shrinkage approach is applied to K-means algorithm, two-sample t-test, estimation of mutual information and estimation of covariance matrix. We can see that the shrinkage estimator is reasonably simple and provides a certain type of regularity. Regularity is in the sense of remedy of covariance matrix (from singular matrix to regular matrix) or sparsity of contingency table (from a higher underestimated value of the true entropy to a less underestimated value of the true entropy).

Acknowledgements

I would like to thank my supervisor Zdeněk Valenta for his productive advice about writing this text. The paper has been supported by the SVV-2013-266517 project of Charles University in Prague.

References

[1] Pirrung MC. How to Make a DNA Chip. Angewandte Chemie International Edition. 2002 Apr; 41(8):1276-1289

[2] Quackenbush J. Microarray Data Normalization and Transformation. Nature Genetics Supplement. 2002 Dec; 32:496-501

[3] Smyth GK, Yang YH, Speed T. Statistical Issues in cDNA Microarray Data Analysis. Methods in Molecular Biology. 2003; 224:111-136

[4] Göhlmann H, Talloen W. Gene Expression Studies Using Affymetrix Microarrays. Boca Raton: Chapman & Hall / CRC; 2009

[5] Cui X, Churchill GA. Statistical Tests for Differential Expression in cDNA Microarray Experiments. Genome Biology 2003; 4(4):Article210

[6] Guo Y, Hastie T, Tibshirani T. Regularized Discriminant Analysis and Its Application in Microarray. Biostatistics. 2007 Jan; 8(1):86-100

[7] Lehmann EL, Casella G. Theory of Point Estimation, 2nd ed. New York: Springer-Verlag; 1998

[8] Stein C. Inadmissibility of the Usual Estimator for the Mean of Multivariate Normal Distribution. Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability. 1956; 1:197-206

[9] Schäfer J, Strimmer K. A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics. Statistical Applications in Genetics and Molecular Biology. 2005; 4(1):Article32

[10] James W, Stein C. Estimation with Quadratic Loss. Proceedings of the Fourts Berkeley Symposium on Mathematical Statistics and Probability. 1961; 1:361-379

[11] Baranchik AJ. Multiple Regression and Estimation of the Mean of a Multivariate Normal Distribution. Stanford (CA): Department of Statistics, Stanford University. 1964 May. 39 p. Report No.:51

[12] Bock M. Minimax Estimators of the Mean of a Multivariate Normal Distribution. Annals of Statistics. 1975; 3(1):209-218

[13] Gao J, Hitchcock DB. James-Stein Shrinkage to Improve k-Means Cluster Analysis. Computational Statistics and Data Analysis. 2010 Sep; 54(9):2113-2127

[14] Opgen-Rhein R, Strimmer K. Accurate Ranking of Differentially Expressed Genes by a Distribution-Free Shrinkage Approach. Statistical Applications in Genetics and Molecular Biology. 2007; 6(1):Article9

[15] Cui X, Hwang G, Qiu J, Blaxes NJ, Churchill GA. Improved Statistical Test for Differential Gene Expression by Shrinking Variance Components. Biostatistics. 2005 Jan; 6(1):59-75

[16] Hausser J, Strimmer K. Entropy Inference and the James-Stein Estimator, with Application to Nonlinear Gene Association Networks. Journal of Machine Learning Research. 2009 Jul; 10:1469-1484

[17] Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A. ARACNE: an Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. BMC Bioinformatics. 2006 Mar; 7(Supplement 1):S7

[18] Nemenman I, Shafee F, Bialek W. Entropy and Inference, Revisited. Advances in Neural Information Processing Systems, 14:471-478

[19] Kwan CCY. An Introduction to Shrinkage Estimation of the Covariance Matrix: A Pedagogic Illustration. Spreadsheets in Education [Internet]. 2011 [cited 2013 Aug 1]; 4(6):Article6. Available from: http://epublications.bond.edu.au/ejsie/vol4/iss3/6/

[20] Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression. Proceedings of the National Academy of Sciences of the United States of America. 2002; 99(10):6567-6572

[21] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer-Verlag; 2009