

**Variable Selection and Pattern Recognition with Gene Expression Data
Generated by the Microarray Technology**

A. Szabo^{1*}, K. Boucher¹, W.L. Carroll¹, L.B. Klebanov², A.D. Tsodikov¹, A.Y. Yakovlev¹

¹*Huntsman Cancer Institute, Department of Oncological Sciences, University of Utah,
2000 Circle of Hope, Salt Lake City, Utah 84112*

²*Department of Mathematics, Idaho State University, Pocatello, ID 83209-8085*

Abstract

Lack of adequate statistical methods for the analysis of microarray data remains the most critical deterrent to uncovering the true potential of these promising techniques in basic and translational biological studies. The popular practice of drawing important biological conclusions from just one replicate (slide) should be discouraged. In this paper, we discuss some modern trends in statistical analysis of microarray data with a special focus on statistical classification (pattern recognition) and variable selection. In addressing these issues we consider the utility of some distances between random vectors and their nonparametric estimates obtained from gene expression data. Performance of the proposed distances is tested by computer simulations and analysis of gene expression data on two different types of human leukemia. In experimental settings, the error rate is estimated by cross-validation, while a control sample is generated in computer simulation experiments aimed at testing the proposed gene selection procedures and associated classification rules.

Keywords: statistical inference, nonparametric methods, pattern recognition, data adjustment, probability distance

1 Introduction

Microarray technology offers an exciting opportunity to simultaneously monitor the expression pattern of thousands of distinct genes. Researchers can track the effects of interventions or natural processes on gene expression levels thus identifying the functions of genes and the biochemical pathways they participate in. To attack this biologically intriguing problem one needs to address a wide spectrum of issues of general methodological interest in statistical analysis of gene expression data; these include (i) the reduction of experimental noise in the data, (ii) the identification of subsets of “marker” genes (target subsets) to characterize specific signaling pathways (*data reduction, variable selection, unsupervised learning*), and (iii) the classification of gene expression patterns (pathways) into known classes (*supervised learning*).

There are two daunting difficulties that make the analysis of microarray data extraordinarily challenging: (1) A very high variability between slides (even for the same tissue sample) makes it difficult to detect biologically significant changes in gene expression. Much of this variability comes from the probe labeling process and instability of experimental conditions (humidity, temperature, hybridization pattern, etc.). (2) The set of microarray data on p distinct genes represents a random vector $\mathbf{X} = X_1, \dots, X_p$ with mutually dependent components. The dimension of \mathbf{X} is extremely high relative to the number of observations (replicates of experiments). This unusual feature of microarray data prevents maximization of the posterior Bayes risk based on the adaptive kernel estimation of the class-conditional probability densities and other conventional discriminant analysis techniques using \mathbf{X} as the feature vector. Therefore, the identification of a feature subvector of much lower dimension than that of \mathbf{X} becomes the most crucial step on the road to efficient decision rules of pattern recognition with microarray data.

*Author for correspondence: Huntsman Cancer Institute, Department of Oncological Sciences, University of Utah, 2000 Circle of Hope, Salt Lake City, UT 84112-5550, ph: (801)585-5182, fax: (801)585-5357, e-mail: aniko.szabo@hci.utah.edu

The general problem of statistical pattern recognition can be formulated within the framework of discriminant analysis [1]. In discriminant analysis, each pattern is considered as an entity which belongs to one of a number of predefined classes or groups of patterns (tissues) and can be represented by a vector, \mathbf{Y} , of features of the pattern. Suppose there is a finite number of distinct classes or groups, G_1, \dots, G_k , whose existence is known a priori. An entity (tissue sample) is assumed to belong to one and only one of these classes. Let \mathcal{I} be an indicator of the group membership, that is $\mathcal{I} = i$ implies that the given entity belongs to class G_i , $i = 1, \dots, k$. Let the d -dimensional vector $\mathbf{y} = (y_1, \dots, y_d)$ represent the measurements on d features of the entity. The vector \mathbf{y} is a sample realization of the corresponding random vector $\mathbf{Y} = Y_1, \dots, Y_d$ known as the feature vector.

In the predictive setting of discriminant analysis (*supervised learning*), the intent is to assign the unknown entity to one of the k classes on the basis of individual values of measurements. In other words, the problem is to estimate the function \mathcal{I} from the observed values of \mathbf{y} . In the descriptive setting (*unsupervised learning*), no allocation of the entity to one of the classes G_1, \dots, G_k is intended, but the prime goal is to provide insight into the structure of possible predictor variables. The predictive and descriptive settings of discriminant analysis are closely related; the two ways of analysis supplement and enrich each other when used in combination.

While the basic principles of statistical classification or pattern recognition are clear, the daunting question still remains: how to overcome the “curse of dimensionality” in the analysis of microarray data generated by a wide scan of the genome. The central problem is to find a feature vector, say $\mathbf{Y} = Y_1, \dots, Y_d$, $d \ll p$, that could be efficiently used in the construction of a classification (discriminant) rule from available training-sample data. This problem is of a fundamental rather than a technical nature. It is a well known fact that the performance of a given decision rule does not keep improving as the dimension, d , of the feature vector \mathbf{Y} is increased. Rather, a sort of peaking phenomenon is typically present, i.e. the overall unconditional error rate of a discriminant rule stops decreasing and starts to increase as d exceeds some threshold, depending on the specific data set under study [1]. Good use may be made of discriminant analysis once a feature vector of substantially lower dimension than that of the full vector \mathbf{X} has resulted from exploratory data analyses.

The problem of initial selection of feature variables \mathbf{Y} for use in recognition of cell types (tissues) is the main subject matter of this paper. Many methodological trends in microarray data analysis, briefly reviewed in the next section, have a direct bearing on the problem under discussion.

The outline of the paper is as follows: Section 2 contains a brief and necessarily incomplete review of current methods used in microarray analysis. Sections 3 and 4 describe initial adjustment procedures for raw and ranked microarray data respectively. Section 5 is the heart of the paper, in which we describe the construction a new class of metrics, and the use of metrics of this class for multivariate variable selection. Section 6 is the most theoretical part of the paper, and describes a multidimensional two-sample test for the hypothesis. Simulation results comparing the new methodology with some marginal selection procedures are presented in Section 8. Finally, in Section 9 we use the new methodology to classify two leukemia data sets.

2 Modern Trends in Methodology of Microarray Data Analysis

2.1 Finding co-regulated genes

One commonly accepted methodology uses clustering techniques to study time-dependent changes in gene expression initiated by biological stimuli or developmental processes. Carr et al. [2] studied the activity of genes involved in the formation of the rat spinal cord during gestation and postnatal development. The authors have shown that genes with similar known functions tend to cluster together. Eisen et al. [3] obtained similar results from experimental studies of growth responses of yeast (*S. cerevisiae*) and human cells. Cho et al. [4] studied variations of gene expression during the cell cycle in *S. cerevisiae* and identified the well known regulators as well as groups of genes that had not been noted by then for their association with the cell cycle. The authors used biological approaches to show that the newly detected genes are required for successful completion of cell division. Hughes et al. [5] developed a database containing expression profiles of 300 diverse mutations and treatments in *S. cerevisiae*. Using clustering the authors determined groups of genes that react similarly to different interventions and thus, presumably, are involved in a common function. The functions of several uncharacterized open reading frames in these groups were confirmed by additional experiments. These results convincingly demonstrate that gene expression profiles are a powerful tool for classifying genes. Most papers on gene clustering use hierarchical clustering methods such as the FITCH

method designed to construct phylogenetic trees [2,6], the average [3] or single [7] linkage clustering and the two-way clustering [8]. Although the hierarchical trees produced by these methods provide an appealing visual display [3], there are several problems with their ability to adequately describe the data. The decision to include a pair of genes in the same cluster is based only on a specific distance between them (introduced to quantitatively characterize the degree of co-regulation) and any such decision is final. The local nature of this decision rule often inhibits the algorithm's ability to find a global structure. Also, the hierarchical trees are more suited for the description of real hierarchical relationships (e.g. evolutionary processes), while there is no evidence for the existence of such relationships in biological functions of different genes. In addition, the resulting structure is very complex and there is no general agreement on how to choose the location for cutting the tree.

There are methods that construct a direct partitioning of genes into clusters, thus circumventing some of the problems inherent in hierarchical algorithms, but they rely heavily on prespecified values of many relevant parameters that may influence the net results of clustering. The k -means algorithm [9,10] iteratively finds the partition of the data into k clusters that minimize the sum of squared distances between each observation and the centroid of the cluster it belongs to. The method of self-organizing maps (SOM), that was applied to expression data by Tamayo et al. [11], assigns each gene to a vertex of a prespecified lattice. Clusters that are close to each other in the geometry of the lattice are more similar to each other than the far-away clusters. No rigorous statistical results are available to provide a theoretical underpinning for these methods. It should be noted that the above mentioned clustering methods provide only an exploratory tool, they do not have a statistical model in the background, thereby providing no measures of confidence or assurance of the correctness of findings.

More recently several researchers have attempted to develop clustering techniques specifically for genetic data that can also provide measures of quality and/or correctness. Under certain parametric assumptions it is possible to provide probability guarantees for reconstructing the correct clusters [12,13]. Heyer et al. [14] proposed a method for constructing clusters of prespecified diameters that are large (relative to those produced by other methods) in size but still quite compact. The CLICK algorithm [15] finds small tight kernels that are subsequently expanded into larger clusters. This graph-based algorithm appears to perform better than the self-organizing maps and k -means clustering in terms of cluster homogeneity and separation efficiency. Hastie et al. [16,17] proposed the so-called 'gene-shaving' method to organize a multi-step search for clusters of genes that satisfy certain criteria, including their significance as predictors in post-treatment cancer survival. The 'gap statistic' introduced by these authors for the purpose of choosing the 'most significant' cluster deserves further exploration in the context of other methodological approaches.

The choice of an appropriate quantitative characteristic of co-regulation or co-expression of genes is of crucial importance for the analysis of microarray data. Use of similarity measures or probability distances is a standard practice in the field. However, it should be kept in mind that the choice of a similarity measure may have a strong effect on the outcome of a given clustering algorithm. Unfortunately, this issue has received little attention in the literature. In the literature on gene expression, the most commonly used measure of similarity (or rather co-monotonicity) of time-dependent variations in expression of two different genes is Pearson's correlation coefficient [3]. As a surrogate for time some authors suggest appending the differences between consecutive observations to the response vector [2,6]. The Pearson correlation coefficient is known to be sensitive to outliers. As a remedy for this problem Heyer et al. [14] proposed using a jackknife estimated correlation coefficient defined as the smallest (in absolute value) sample correlation obtainable with any single observation deleted from the sample.

Another similarity measure, the so-called mutual information distance, was introduced in the analysis of gene expression data by Michaels et al. [6] and also used by Butte et al. [7]. Computing the mutual information distance requires binning of the observed expression levels, resulting in certain losses of information. However, it holds much promise for developing robust similarity measures.

2.2 Differential Expression of Genes

While clustering techniques are useful in providing insights into interactions between different genes or finding genetically similar samples, these techniques do not answer the question many researchers are interested in: which genes are expressed differently in the tissues under comparison? Surprisingly little research has been done to find statistically sound methods for identifying differentially expressed genes. Most authors use the

logarithm of the ratio of the observed fluorescence signals from the two channels and then consider any gene with the ratio above a fixed value (usually 2 or 3) to be differentially expressed [18–20]. Often some kind of adjustment or normalization is carried out before and/or after computing the ratio, but there is no sufficiently general and well defined procedure. It is common practice to subtract the background intensity and normalize the expression level of each gene so that the average (over genes) ratio becomes equal to unity. Sometimes, the data are normalized to a fixed level of a housekeeping gene.

When selecting the initial feature vector for pattern recognition one can interpret it as a target subset of genes that merit special attention. There are simple selection rules that have received certain attention in the literature. For example, van der Laan and Bryan [13] suggest the following relatively simple rule: (1) select those genes which are at least m -fold differentially expressed (m being input by the user) with respect to the marginal mean level of expression in the two states (tissues) under comparison; (2) estimate a correlation-distance matrix for these differentially expressed genes; (3) apply a clustering algorithm to (some function of) this distance-matrix; and possibly (4) only include those genes in the target subset that are closest to the cluster centers. Newton et al. [21] obtained an empirical Bayes estimator of the fold change in gene expression that turned out to be different from the simple ratio. Kerr et al. [22] proposed to use ANOVA for the log-intensities (not ratios) to combine the adjustment step with the identification of differential expression. The authors also raise the issue of experimental design [22, 23] for microarray studies and point out confounding effects in common experimental setups. The authors state that a search for better ways of measuring differences in the expression of a given gene in two tissues is far from complete exploration.

Another important issue regarding differential expression of genes has to do with the sample size. From the statistical point of view, the popular practice of drawing important biological conclusions from just one replicate (slide) should be discouraged. Several papers [24, 25] have demonstrated a very high variability of microarray slides and a tangible improvement achieved through increasing the number of replications. However, efficient ways of combining information from several replicates have yet to be developed. General guidelines for sample size determination in microarray experiments still remain unclear, although a very interesting theoretical treatment of the problem is presented in [13].

The large variability in the gene expression measurements, as well as the common occurrence of observations that are completely corrupted by impurity of some spots on the slide, motivate the development of robust methods for data analysis. Tsodikov et al. [26, 27] proposed methods for testing differential expression of individual genes based on ranks and data categorization. These methods appear to be quite robust to the experimental noise present in microarray data; the corresponding procedures are discussed at length in Section 3.

2.3 Statistical pattern recognition

Golub et al. [28] proposed selecting genes that are individually highly correlated with the known classification and then using a voting procedure for classification of new samples. The authors successfully applied their technique for distinguishing acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) samples. Hastie et al. [16] extended this approach to find genes that are predictive of an endpoint (e.g. survival time) with gene expression treated as a covariate. The authors proceed from a selection criterion which, in some sense, is opposite to that by Golub et al. [28]: they find a linear combination of gene expression signals that explains the most of the variance of the covariate and then eliminate those genes that are least correlated with the combination (instead of choosing the ones that are most correlated).

Dudoit et al. [29] performed a preliminary selection of genes on the basis of the ratio of their between-group and within-group sum of squares and then compare predictors based on the d genes with the largest ratios. In particular, they preset d at values ranging from 30 to 50 most differentially expressed genes for different data sets. Ben-Dor et al. [30, 31] propose two non-parametric scores, termed the threshold number of misclassification and InfoScore, for marginally evaluating the “relevance” of a gene to the classification task. Dudoit et al. [29] compared various discrimination methods for the classification of tumors based on gene expression data including nearest neighbor classifiers, linear discriminant analysis, classification trees, and some more recent machine learning approaches such as bagging and boosting. They found the traditional linear classifiers and nearest neighbors perform remarkably well compared to more sophisticated methods such as aggregated classification trees. The authors note, however, that more advanced methods might work

better if a larger number of samples were available for training.

A common feature of the classification methods mentioned above is the univariate nature of the decision to include a particular gene in the initial feature set. The complex interaction pattern of gene functions makes it unlikely that the contribution of a gene to the between-tissue difference can be evaluated marginally. Even with just two correlated variables it can happen that the knowledge of their marginal distributions is not enough to predict class membership while jointly they may make the prediction very good. Methods that use multivariate information at every step are needed to utilize the information hidden in gene interactions and hence to increase the power of classification rules. Because of a strong “method by data type” interaction (particular methods will be best for particular types of data), no universal solution is available.

One of the most exciting uses of statistical classification as applied to gene expression data is the identification of subtypes of well known diseases. This includes both improving the existing classification into known classes and the discovery of new/unknown subgroups that are clinically significant. The genetic profile of a tissue determines its properties, which is why different tissues are expected to have different gene expression patterns. As demonstrated by various clustering methods of gene expression vectors with a measure of similarity formally coinciding with the Pearson correlation [32], tissues of the same histological origin tend to cluster together. Ross et al. [33] came to a similar conclusion when clustering 60 NCI cell lines. Alizadeh et al. [34] used hierarchical clustering to demonstrate the existence of two previously unknown genetically different subtypes of B-cell lymphoma that carry significantly different prognosis in terms of patients’ survival.

The above review shows that at least some practically important issues of microarray data analysis progressively become approachable. On the other hand, the search for better methods remains to be empirical in nature, and we probably are still a long way from a satisfactory solution to the problem. With its potential to quantitatively determine expression levels of a large number of genes in parallel, microarray technology holds the promise of becoming an extremely valuable tool in basic biological sciences and clinical diagnostics, but its ultimate usefulness will depend critically on whether or not the search for efficient statistical methods meets with success.

3 Differential Expression of Individual Genes and Data Adjustment

Currently there are (at least) two competing technologies for gene expression analysis. The spotted cDNA microarray technology developed at Stanford University [19] is geared toward a comparative measurement of gene expression, while high density oligonucleotide chips developed by *Affymetrix* aim at measuring the absolute gene expression levels. While the second approach is probably more appropriate for classification purposes, two-color spotted cDNA microarrays can be adapted as well either by using a reference tissue design or through adjustment procedures.

First we introduce some notation to describe experimental data generated in the form of two-color cDNA microarrays. Let n denote the number of slides, and p be the total number of genes. Let $A_i, B_i, i = 1, \dots, p$ be a pair of random variables representing gene expression measurements for the two tissues in an ideal reproducible experiment.

Remark 1. In a reference tissue design we assume that the red channel corresponding to A_i contains the tissue of interest. Associated with each slide, indexed by $j = 1, \dots, n$, is a pair of dependent random variables X_{ij}, Y_{ij} representing paired (two channel) measurements of fluorescent intensity for gene $i = 1, \dots, p$, where X refers to the red channel, and Y to the green channel.

The notation is simpler for one-channel technologies (oligonucleotide chips, radio-labeled arrays) where A_i denotes the random variable for the ‘ideal’ measurement and X_{ij} the variable for the actual observation. While X_{ij} and Y_{ij} are random variables, in many experimental settings genes appear on a given slide only once, so for these variables only one observation, denoted by x_{ij} and y_{ij} respectively, is available for each slide.

Due to the errors in the measurement process, the distribution of the observed intensities X_{ij}, Y_{ij} is different from the distribution of the ‘ideal’ measurements A_i, B_i . Hence in order to make inference about A_i, B_i , the relationship between the two sets of random variables needs to be explored. Informally, the model

behind a microarray experiment can be written as

$$\varphi_1(A_{ij}, \epsilon_{ij}) = X_{ij} \quad \varphi_2(B_{ij}, \delta_{ij}, \cdot) = Y_{ij}, \quad (1)$$

where ϵ, δ are measurement errors, and φ_1, φ_2 are some non-random functions. The observed gene expression levels need to be adjusted with the aim to restore the ideal sample a_{ij}, b_{ij} (drawn from A_i, B_i) or its surrogate by transforming the observed sample x_{ij}, y_{ij} .

It is difficult to specify the form of $\varphi_k, k = 1, 2$ on mechanistic grounds. There are many sources of the observed experimental noise so that some of the errors are likely to be additive (background), some others are multiplicative (dye incorporation, fluorescence efficiency, spot size), while saturation effects have a non-linear form, and many of them may vary within one slide as well. By far the most frequently used assumption is that the measurement error has a simple multiplicative structure. More specifically,

$$X_i = \alpha A_i \quad Y_i = \beta B_i, \quad (2)$$

where α and β are scalar random variables taking on positive values. Suppose two random samples from $X = X_1, \dots, X_p$ and $Y = Y_1, \dots, Y_p$ are available, i.e.,

$$X_{ij} = \alpha_j A_{ij}, \quad Y_{ij} = \beta_j B_{ij},$$

$i = 1, \dots, p, j = 1, \dots, n$. Under this model, it is assumed that the multiplicative measurement error is slide-specific and is shared by genes on the same slide. The systematic part of α and β accounts for the difference in intensity associated with the type of fluorescent dye used with a specific channel.

Generally, A_{ij}, B_{ij} are not recoverable from X_{ij}, Y_{ij} without additional assumptions. In what follows, we will be interested in testing the hypothesis: $\mathbf{H}_0 : A_i \stackrel{d}{=} B_i$ (A_i and B_i are identically distributed) for the i th gene rather than restoring exact sample values of these two random variables. As the first step, it makes sense to find an adjustment that would reduce the problem to testing the hypothesis: $A_i \stackrel{d}{=} \sigma B_i$, where σ is some nonrandom constant. One commonly used adjustment procedure is to divide each expression signal, say x_{ij} (j is fixed), by the arithmetic mean taken over all the expression signals recorded on the same slide. The rationale for such a procedure is as follows. If the measurement error does not depend on i , all genes on the same slide (or half-slide) share the same slide-specific random effect. Introduce the notation

$$A_{.j} = \frac{1}{p} \sum_{i=1}^p A_{ij}, \quad X_{.j} = \frac{1}{p} \sum_{i=1}^p X_{ij}.$$

Then one can eliminate the noise by generating the following adjusted observations

$$\tilde{A}_{ij} = \frac{X_{ij}}{X_{.j}} = \frac{A_{ij}}{A_{.j}}, \quad \tilde{B}_{ij} = \frac{Y_{ij}}{Y_{.j}} = \frac{B_{ij}}{B_{.j}}. \quad (3)$$

In like manner, this adjustment can be applied to the log-transformed multiplicative model resulting in

$$A_{ij}^* = \frac{X_{ij}}{\sqrt[p]{\prod_{i=1}^p X_{ij}}} = \frac{A_{ij}}{\sqrt[p]{\prod_{i=1}^p A_{ij}}}, \\ B_{ij}^* = \frac{Y_{ij}}{\sqrt[p]{\prod_{i=1}^p Y_{ij}}} = \frac{B_{ij}}{\sqrt[p]{\prod_{i=1}^p B_{ij}}}.$$

It is clear that the equality $A_i \stackrel{d}{=} B_i$ does not follow from $\tilde{A}_i \stackrel{d}{=} \tilde{B}_i$ (or $A_i^* \stackrel{d}{=} B_i^*$), and additional assumptions are in order here. Suppose that the law of large numbers (for dependent random variables) is valid for the sequences of A_i and $B_i, i = 1, \dots$, or their logarithms. For this condition to be met it is sufficient to require that the variance, given it exists, of the arithmetic mean $\frac{1}{p} \sum_{i=1}^p A_i$ tends to zero as $p \rightarrow \infty$, the same being valid for the sequence B_i . Then for sufficiently large finite p we can contend that the equality $\tilde{A}_i \stackrel{d}{=} \tilde{B}_i$ implies $A_i \stackrel{d}{=} \sigma B_i, \sigma > 0$.

Remark 2. Another assumption leading to the same result is that the expression levels of at least three out of p genes are independent random variables and all marginal distributions of gene expression levels have

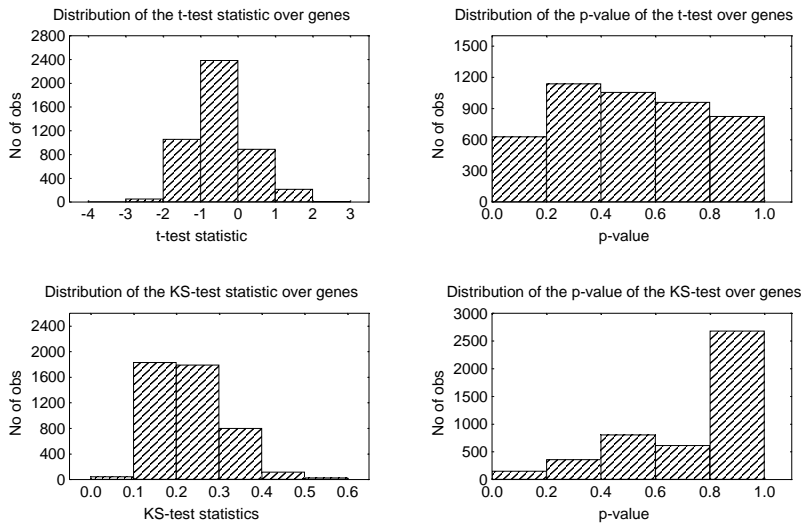


Figure 1: Distributions over genes of the test statistics and p-values for t-test and the Kolmogorov-Smirnov (KS) test when applied to test for differential expression in the two subsets of ALL data, presumably under the null hypothesis.

finite first moment. The assertion can be proven using a method similar to that for the Prokhorov theorem on the maximal invariant [35], but it does not by itself (without invoking the law of large numbers) suggest a constructive way of data adjustment.

It is clear that even if sufficient conditions for the law of large numbers are satisfied, we still are unable to test the hypothesis: $A_i \stackrel{d}{=} B_i$. However, if we assume in addition that

$$\frac{1}{p} \sum_{i=1}^p EA_i = \frac{1}{p} \sum_{i=1}^p EB_i,$$

where E is the symbol of expectation, then from $A_i \stackrel{d}{=} \sigma B_i$ it follows that $\sigma = 1$ and $A_i \stackrel{d}{=} B_i$. The rationale for the latter assumption is that the majority of genes are believed to perform housekeeping functions so that only a small proportion of genes is expected to change their expression when comparing two different tissues with the average over all genes remaining unaltered. In a similar manner, more complex multiplicative-additive slide-specific error effects can be considered, but substantiating the corresponding adjustment procedure would require even stronger assumptions.

Both assumptions formulated above are untestable with experimental data on gene expression. The best we can do is to check whether or not they are consistent with real data. To this end we applied the multiplicative adjustment procedure (3) to the data on acute lymphoblastic leukemia (ALL data, Data Set 2, see Section 9). The data set was split randomly into two different subsets of equal size. By the design of experiment, no difference is expected in gene expressions between the two subsets. The t -test and the Kolmogorov-Smirnov (KS) test were used to test the hypothesis: $\tilde{A}_i \stackrel{d}{=} \tilde{B}_i$ for each $i = 1, \dots, p$, both at the significance level of 0.05. Histograms of the test statistics and the corresponding p-values are shown in Figure 1. The t -test classified 1.09% genes as significantly differentially expressed, while with the KS test this number was as low as 0.6%. It is worth noting that the distribution of p-values for the t -test appears to be quite uniform while this apparently is not the case for the KS test. It is possible to test the equality of the distributions of the slide-specific effects α and β in the two tissues by applying a two-sample test to compare $X_{.j}$ and $Y_{.j}$. We used the t -test to compare these distributions for the ALL versus AML data (Data Set 2). No difference was found ($p=0.49$).

This analysis does not reject the multiplicative error model but the evidence in favor of the model is rather weak. More analyses of various data sets are necessary before the model can be adopted as a common tool for testing differential expression of genes.

4 Categorical Adjustments of Microarray Data

However appealing properties of the proposed adjustment procedures based on specific models of gene expression signals may be, the most fundamental question still remains: what is a sufficiently realistic model for the functions (transformations) φ_k in Eq. (1). A reasonably general model follows from the assumption that φ_k are monotone transformations that preserve the ordering of an entire set of gene expression levels arranged in order of magnitude. The idea motivated Tsodikov et al. [26] to suggest replacing the actual observations with their fractional rank (that is the rank divided by the total number of genes) within the slide:

$$X_{ij}^{(r)} = (\text{rank}_j X_{ij})/p, \quad Y_{ij}^{(r)} = (\text{rank}_j Y_{ij})/p,$$

where $\text{rank}_j u_{ij}$ is the place (counted from the left) of u_{kj} in the sequence u_{ij} , $i = 1, \dots, p$ arranged in decreasing order for each j .

In many practical situations, this adjustment restores the correct ordering of observations in the presence of experimental noise of a fairly general structure. Another obvious advantage of this adjustment is its stability to outliers. However, the price to be paid for these advantages is a substantial loss of information which may be especially tangible when the sample size is small. In particular, the expression of a given gene may change significantly with its rank remaining unchanged. Conversely, the rank of a given gene may change (because of changes in expression of other genes) while there is no change in its own expression level. More generally, identical distribution of ranks in two tissues does not imply identical distribution of the corresponding vectors of expression signals. Furthermore, if the components of some subvector of gene expression signals behave like independent and identically distributed random variables, then the ranks of all the genes included in this subvector are equally likely. Therefore, it would require very large sample sizes to make statistical inference from ranked observations on such a subvector. However, we believe that the situations described above are rather the exception than the rule for microarray data analysis, and all the caveats do not outweigh the usefulness of the robust inference based on ranks.

Tsodikov et al. [27] conducted computer simulations assuming independent and log-normally distributed expression of individual genes with the error structure specified by a model that includes both the slide-specific multiplicative and the slide-specific additive noise components. The authors used the Kolmogorov-Smirnov statistic (not the test!) and the t -statistic to produce a list of top differentially expressed genes. These statistics were employed just to order genes, but no conclusions in terms of statistical hypotheses testing were made. Then this list was compared with a list of those genes whose mean expression levels had been changed in computer experiments. Using simulated data, the ordering of differentially expressed genes suggested by the method under evaluation can be verified against the “true” ordering. Newton et al. [21] proposed a simple plot to compare the performance of different methods in restoring the “true” orderings. Consider the N “truly” most differentially expressed genes that are preset in a simulation study. Let $M(N)$ be the number of these genes ranked in the top N by a particular method. Ideally, if the method restores the true (error free) order, one will have $M(N) = N$ for any N . If errors are present one should expect $M(N) \leq N$. Obviously, $M(N)$ is an increasing function such that $M(0) = 0$ and $M(p) = p$. The more the curve $M(N)$ arches towards the lower right corner $(0, p)$ on the plane $\{M\} \times \{N\}$, the worse the method’s performance. The rank-based adjustment was shown to perform well (in terms of correct selection of differentially expressed genes) under the simplistic model used in the simulation study by Tsodikov et al.

When comparing expression levels of a given gene in two different states of a tissue or two different tissues, use is frequently made of the so-called reference design that places one sample from the tissue of interest and, whenever another channel is available, a reference tissue (one of no interest, usually a related cell line) on each slide. The study by Tsodikov et al. [26] suggests that even when readings from both channels are available, simply ignoring the observations of the reference tissue and using the robust adjustment based on ranks for the tissue of interest leads to reliable results.

Another idea is to use a scatter plot of expression measurements from a particular slide for data categorization. Measurements of fluorescent intensity in two channels (x =Green and y =Red) gives a point (x, y) on the plane. A set of all such points for the genes associated with a given slide forms a scatter plot. Ideally, non-differentially expressed genes would preserve a constant Green/Red ratio of 1, the corresponding (x, y) points building a line on the plane. A differentially expressed gene would ideally show a different ratio, the corresponding points being away from the line.

However, for a number of reasons the picture is more complex:

- Additive background effect provides for a non-zero intercept of the line;
- Due to measurement errors and random nature of gene expression, the points corresponding to non-differentially expressed genes are scattered considerably around the line;
- A strong slide-specific effect makes the scale and the scatter plot pattern variable from slide to slide.

The purpose of data adjustment is to transform the measurements of gene expression so that they be on the same scale. Statistical tests can then be applied to the transformed sample, a surrogate of ideal measurements. Generally speaking, the sample of x and y values is drawn from a system (vector) of dependent random variables with an unknown dependency structure. The set of values $\{(x_i, y_i)\}_{i=1}^p$ contains an unknown fraction of “outliers” that are not expected to follow the line. Also, both x and y are subject to measurement error. In a situation where both x and y are measured with error, a linear structural relationship is nonidentifiable without additional constraints. Even in the simplest case of independent measurements, a least squares line for the model

$$X_i = U_i + \delta_i \quad Y_i = V_i + \epsilon_i, \quad (4)$$

where δ and ϵ are measurement errors, and $V = a + bU$, underestimates the slope b of the latent structural relationship [36].

For the reason explained above, we resort to an ad hoc method to define a reference line for the scatter plot. Having explored a number of robust procedures for linear regression using real and simulated expression data we came up with a simple and computationally fast method based on the one developed by Bartlett [37]. Once the reference line is determined, it is rotated rigidly to coincide with the x -axis and all p points of the scatter plot are projected on the line by the closest point projection. The coordinate system is changed from (x, y) to (t, d) , where d is a signed (directed) distance from the point (x, y) to its projection, and t is a similar distance from the projection to the minimal projection on the reference line. The signed distance d quantifies an instance of differential expression for a particular gene on the slide. Points above the line bear a positive d indicating potential overexpression, while negative d is a sign of potential underexpression.

The distribution of d for genes in some small interval $[t - \Delta, t + \Delta]$ appears to be a function of t , indicating that genes with different order of absolute expression cannot be measured on the same d -scale. The above observation prevents us from directly using d as a surrogate of differential expression. A summary measure of differential expression can be constructed by ranking genes with respect to the directional distance d adjusted for the surrogate of absolute expression signal t . To categorize differential expression, define a cross section layer $W_t^+ = \{0 < d < \infty, t - \Delta(t) < t < t + \Delta(t)\}$, where $\Delta(t)$ is a bandwidth. Similarly, $W_t^- = \{-\infty < d < 0, t - \Delta(t) < t < t + \Delta(t)\}$. Define a set of cutpoints α_j , $j = 0, \dots, k + 1$ that break the interval of total probability $[0, 1]$ down into $k + 1$ subintervals. By definition $\alpha_0 = 0$, $\alpha_{k+1} = 1$, $\alpha_{j-1} < \alpha_j$. A gene with coordinates (t_i, d_i) above the reference line is assigned a category of differential expression C_j^+ if $C_{\alpha_j}^+ < d_i \leq C_{\alpha_{j+1}}^+$, where C_{α}^+ is the empirical α -percentile of the distribution of d for genes in the layer W_t^+ . All genes in W_t^- under the line are categorized in a similar manner. In fact, as W_t depends on t , C_{α_j} is a function of t representing a moving-average estimator of the α_j -percentile of the distribution of d given t . The step-functions $C_{\alpha_j}(t)$ cut the plane into $2k + 1$ percentile bands $\mathcal{B}_j^+ = \{0 \leq t < \infty, C_{\alpha_j}^+ < d \leq C_{\alpha_{j+1}}^+\}$ and $\mathcal{B}_j^- = \{0 \leq t < \infty, C_{\alpha_{j+1}}^- < d \leq C_{\alpha_j}^-\}$ (the bands \mathcal{B}_0^+ and \mathcal{B}_0^- are combined into a single one).

To keep the estimation accuracy to a constant, Δ is treated as data-adaptive and such that for any t the layer W_t contains approximately the same number of points. A constraint can be also imposed on the maximal bandwidth.

With $k = 1$, the observed gene expression falls into one of the following three categories: “Overexpressed” (the point is in the upper band \mathcal{B}_1^+), “Not differentially expressed” (the point is in the middle band \mathcal{B}_0) and “Underexpressed” (the point is in the lower band \mathcal{B}_1^-). With $k > 1$ overexpression and underexpression are

represented as a multiple of categories

$$(X_{ij}, Y_{ij}) \rightarrow \left\{ \begin{array}{ll} \text{Overexpressed } k & \text{if } (t_{ij}, d_{ij}) \in \mathcal{B}_{kj}^+ \\ \dots & \dots \\ \text{Overexpressed } 1 & \text{if } (t_{ij}, d_{ij}) \in \mathcal{B}_{1j}^+ \\ \text{Not diff. expressed} & \text{if } (t_{ij}, d_{ij}) \in \mathcal{B}_{0j} \\ \text{Underexpressed } 1 & \text{if } (t_{ij}, d_{ij}) \in \mathcal{B}_{1j}^- \\ \dots & \dots \\ \text{Underexpressed } k & \text{if } (t_{ij}, d_{ij}) \in \mathcal{B}_{kj}^- \end{array} \right. \quad (5)$$

An important feature of the proposed categorical summary measure of differential expression is that any rank preserving transformation (possibly dependent on the absolute expression level t) of ideal expression data will be adequately adjusted for.

Under the null hypothesis of no differential expression, genes are expected to show overexpression approximately as often as they show underexpression. In other words, the distribution of a categorical measure of differential expression over a set of slides is symmetric under the null hypothesis.

For a given gene i , introduce the notation: n_i^+ = the number of slides where the gene happened to be in category \mathcal{C}_i^+ ; n_i^- = the number of slides where the gene happened to be in category \mathcal{C}_i^- ; n_0 = the number of slides where the gene happened to be in category \mathcal{C}_0 ; p_i^+ = the probability for the gene of being in category \mathcal{C}_i^+ ; p_i^- = the probability for the gene of being in category \mathcal{C}_i^- ; p_i^0 = the probability for the gene of being in category \mathcal{C}_i^0 . The total number of slides $n = \sum_{i=1}^k (n_i^+ + n_i^-) + n_0$.

The null hypothesis that the gene is not differentially expressed can be formulated as $p_i^+ = p_i^- = p_i$, $i = 1, \dots, k$. Under the null hypothesis $\hat{p}_i = (n_i^+ + n_i^-)/(2n)$, $\hat{p}_0 = n_0/n$. Under a saturated model, $\hat{p}_i^+ = n_i^+/m$, $\hat{p}_i^- = n_i^-/n$, $\hat{p}_0 = n_0/n$.

The likelihood ratio statistics can be used to summarize and quantify differential expression over a series of experiments: $LR = 2 \sum_{i=1}^k (n_i^- \log(n_i^-) + n_i^+ \log(n_i^+) - (n_i^- + n_i^+) \log((n_i^- + n_i^+)/2))$. Under the null hypothesis LR is asymptotically χ^2 -distributed with k degrees of freedom. The power of the symmetry-test for differential expression with categorical data can be increased by noting that under the null hypothesis of no difference large over/underexpression should occur less often than a less pronounced deviation. That is the distribution of the categorical measure of differential expression is not only symmetric and unimodal but it also has monotonically decreasing tails. This suggests an isotonic version of the test for symmetry in order to account for the above mentioned constraint on the corresponding test statistic: $p_1^+ = p_1^- \geq p_2^+ = p_2^- \geq \dots \geq p_k^+ = p_k^-$. The maximum likelihood estimates under the ordering restriction can be found using the method of isotonic estimation [38]. The asymptotic distribution of the likelihood ratio test statistic is no longer expected to be χ_k^2 , but rather a mixture of χ^2 variables with different degrees of freedom. The likelihood ratio statistic computed for each gene can be used to order genes according to their differential expression. Computer simulations conducted by Tsodikov et al. [26] show that at least under some models of microarray data this method may out-perform the one based on ranks.

5 Searching for the Initial Feature Vector

For n independent observations of gene expression in a given state of the biological system under study, we expect the same genes to be expressed at certain levels subject to random variation in expression. This set of observations forms an observation matrix of dimension $n \times p$, where p is the total number of genes. The first step on the road to multidimensional classification is to reduce the full feature vector represented by the data on expression of all genes. Most of the cDNA's spotted on the array represent genes that are not involved in the processes that distinguish the two samples under comparison. As described in Section 2, current methods for determining differentially expressed genes are based on univariate choices like those mentioned in Section 2.2. This approach ignores the correlation information contained in the data and thus may limit the power of classification rules. Another concern is that the selection of the feature set is not closely related to the classification of unknown entities. Thus while the gene selection process might select 'significant' genes in the sense of marginal differential expression, they might not be the best choice as a feature set for the classification method.

It looks like a good idea to search for a subset of genes that in some sense differs the ‘most’ between two tissues and then develop a classification rule based on the same notion of difference. To attain this goal we need a pertinent probability distance between two subsets (clusters) of genes. This distance must satisfy the following requirements: (1) it has to be a probability distance (metric) [39] so that its empirical counterpart can combine information from different slides; (2) it should accommodate ranks and categorical data (thus should not necessarily assume normality); (3) the computation of the distance should not be too time consuming. One such distance is proposed below and will be discussed at length in this paper. By calculating the distance based on an entire cluster instead of separately for each gene, one more fully utilizes the multidimensional information on gene expression. Since clusters of size one are also considered, this generalization can only improve the univariate procedure of variable selection.

5.1 Differential expression of subsets of genes

Hastie et al. [16,17] have attempted to use clusters in prediction, but their method of averaging the readings over the genes included in a given cluster seems to disregard a substantial part of the multidimensional information contained in sample observations. A high correlation between expression levels for individual genes and large cluster sizes compared to the number of independent replicates hampers the use of two-sample statistical tests. In discriminant analysis settings, the Mahalanobis distance has become the standard measure of distance between two groups when the feature variables are continuous. The distance is defined as follows: if the feature vector \mathbf{Y} is drawn from a two-variate distribution with means \mathbf{m}_1 and \mathbf{m}_2 , and common covariance matrix \mathbf{S} , then $R_{Mah}^2 = (\mathbf{m}_1 - \mathbf{m}_2)' \mathbf{S}^{-1} (\mathbf{m}_1 - \mathbf{m}_2)$.

To ensure the nonsingularity of the matrix \mathbf{S} estimated from sample observations one should impose the constraint: $n > d$, where n is the sample size, $d \leq p$ is the number of genes in the target subset. This important practical constraint should be kept in mind when working with microarray data, especially at the stage of the initial selection of feature variables. The same can be said about the Chernoff distance in the multivariate normal case. In addition, empirical counterparts of these distances, as well as those based on kernel estimates of multivariate distributions, are not robust enough to the experimental noise which is inherent in the microarray technological process and is difficult to eliminate completely by adjustment procedures. Although robustified versions of the Mahalanobis distance are available (they can be obtained from some functions of trimmed or Winsorized variances, see e.g. [40]), their practical use can become prohibitively expensive even with high-speed computers.

We propose a new distance and its nonparametric estimate to measure differential expression between subsets of genes. Let μ and ν be two probability measures defined on the Euclidean space \mathbb{R}^d . Let $L(\mathbf{x}, \mathbf{y})$ be a strictly negative definite kernel, that is $\sum_{i,j=1}^s L(\mathbf{x}_i, \mathbf{x}_j) h_i h_j \leq 0$ for any $\mathbf{x}_1, \dots, \mathbf{x}_s$ and h_1, \dots, h_s , $\sum_{i=1}^s h_i = 0$ with equality if and only if all $h_i = 0$. Introduce the following expression

$$N(\mu, \nu) = 2 \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} L(\mathbf{x}, \mathbf{y}) d\mu(\mathbf{x}) d\nu(\mathbf{y}) - \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} L(\mathbf{x}, \mathbf{y}) d\mu(\mathbf{x}) d\mu(\mathbf{y}) - \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} L(\mathbf{x}, \mathbf{y}) d\nu(\mathbf{x}) d\nu(\mathbf{y}).$$

It can be shown [41] that $\sqrt{N(\mu, \nu)}$ is a metric in the space of all probability measures on \mathbb{R}^d .

Consider two independent samples, consisting of n_1 and n_2 observations respectively, represented by the d -dimensional vectors $\mathbf{x}_1, \dots, \mathbf{x}_{n_1}$ and $\mathbf{y}_1, \dots, \mathbf{y}_{n_2}$, and introduce an empirical counterpart of $N(\mu, \nu)$ as follows

$$\hat{N} = N(\hat{\mu}_{n_1}, \hat{\nu}_{n_2}) = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} 2L(x_i, y_j) - \frac{1}{n_1^2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} L(x_i, x_j) - \frac{1}{n_2^2} \sum_{i=1}^{n_2} \sum_{j=1}^{n_2} L(y_i, y_j).$$

5.2 Constructing Negative Definite Kernels for Gene Selection

When using the distance $\sqrt{N(\mu, \nu)}$ one needs to choose a pertinent function L . A natural choice is a monotone function of the Euclidean distance between ranks. We describe a simple alternative class of functions which can be used to measure pairwise gene interaction.

Let \mathbf{x} and \mathbf{y} denote observations in two samples on a gene set \mathcal{S} and \mathbf{x}^r and \mathbf{y}^r denote the corresponding rank-adjusted observations. We consider either of these observations to be points in Euclidean space \mathbb{R}^d . Let \mathbf{S} be a measurable subset of \mathbb{R}^d . Define $L_{\mathbf{S}}$ by the rule $L_{\mathbf{S}}(\mathbf{x}, \mathbf{y}) = 0$ if both $\mathbf{x} \in \mathbf{S}$ and $\mathbf{y} \in \mathbf{S}$ and

$L_{\mathbf{S}}(\mathbf{x}, \mathbf{y}) = 1$ otherwise. It is easy to see that $L_{\mathbf{S}}$ is a negative definite kernel. In fact, suppose, for simplicity, that $\mathbf{x}_i \in \mathbf{S}, 1 \leq i \leq r$, and $\mathbf{x}_i \notin \mathbf{S}, r+1 \leq i \leq s$. Then $\sum_{i,j=1}^s (1 - L_{\mathbf{S}}(\mathbf{x}_i, \mathbf{x}_j))h_i h_j = \sum_{i,j=1}^r h_i h_j = (\sum_{i=1}^r h_i)^2 \geq 0$. Thus $(1 - L_{\mathbf{S}})$ is a positive definite kernel, and $L_{\mathbf{S}}$ is negative definite.

More generally, let $f(\mathbf{x})$ be a function from a space \mathbb{R}^d to the interval $[0, 1]$, and define $L_f(\mathbf{x}, \mathbf{y}) = \max(f(\mathbf{x}), f(\mathbf{y}))$. Then L_f is a negative definite kernel. In fact, if we define $g_a(\mathbf{x}, \mathbf{y}) = 0$ provided both $f(\mathbf{x}) > a$ and $f(\mathbf{y}) > a$ and $g_a(\mathbf{x}, \mathbf{y}) = 1$ otherwise, then, from the previous paragraph, g_a is a negative definite kernel. It follows from the equality $L_f(\mathbf{x}, \mathbf{y}) = \int_0^1 g_a(\mathbf{x}, \mathbf{y}) da$ that L_f is negative definite. Since a negative definite kernel is unaffected by an arbitrary additive shift, it is clear that $L_f(\mathbf{x}, \mathbf{y}) = \max(f(\mathbf{x}), f(\mathbf{y}))$ will be a negative definite kernel for any bounded function f .

If w_i are positive weights and $f_i, 1 \leq i \leq d$, are functions from \mathbb{R}^d to $[0, 1]$, then $L = \sum_{i=1}^d w_i L_{f_i}$ is also a negative definite kernel. It is clear from the above argument that if $\{f_i\}$ separates points, in the sense that $f_i(\mathbf{x}) = f_i(\mathbf{y})$ for all i implies $\mathbf{x} = \mathbf{y}$, then L is strictly negative definite.

Negative definite kernels of the type described above may be combined with the usual Euclidean distance to form composite kernel functions. For example, define a region function $R_q(u, v) = q \lfloor qu \rfloor + \lfloor qv \rfloor$ (here $\lfloor \cdot \rfloor$ denotes the floor function, its value is the largest integer not exceeding the argument and $q \geq 2$ is an integer parameter). This function is constant on each of the q^2 ‘little squares’ obtained by dividing the sides of the $(0, 1)^2$ square into q equal segments. Then we may define the following kernels on the ranked data:

$$L_1(\mathbf{x}^r, \mathbf{y}^r) = \sqrt{\sum_{g \in \mathcal{S}} (x_g^r - y_g^r)^2},$$

$$L_2(\mathbf{x}^r, \mathbf{y}^r) = w_1 L_1(\mathbf{x}^r, \mathbf{y}^r) + w_2 \sum_{(g_1, g_2) \in \mathcal{S}^2} \left(1 - \mathcal{I}\{R_q(x_{g_1}^r, x_{g_2}^r) = R_q(y_{g_1}^r, y_{g_2}^r)\} \right),$$

where \mathcal{I} is the indicator function. Then L_1 is the standard Euclidean distance and L_2 falls into the class described above. We choose the weights w_1 and w_2 to balance the two components of L_2 with respect to their maximum values: $w_1 = 1/d_{max}$ and $w_2 = 1/\binom{d_{max}}{2}$, where d_{max} is the maximum subset dimension under consideration. The second component of the kernel will be insensitive to perturbation, yet pick up sets of genes which have similar expression levels across samples in one tissue and different expression patterns in the two tissues.

Another choice for a function L_f is based on the correlation coefficient. Let \mathbf{x}^n and \mathbf{y}^n denote data has been normalized so that the tissue-specific sample mean and variance are zero and one respectively. For each pair of genes g_1 and g_2 it makes sense to consider the function $f_{g_1, g_2}(\mathbf{x}^n) = x_{g_1}^n x_{g_2}^n$. The corresponding negative definite kernel L_{g_1, g_2} will detect differences in correlation between the two tissues. For example, if the expressions of g_1 and g_2 have correlation coefficient ρ in one tissue and are uncorrelated in the other, it follows from $2 \max(\rho, 0) - \max(\rho, \rho) - \max(0, 0) = |\rho|$ that the corresponding distance between the tissues will be approximately equal to $|\rho|$.

We may form the negative definite kernel

$$L_3(\mathbf{x}, \mathbf{y}) = w_1 L_1(\mathbf{x}, \mathbf{y}) + w_2 \sum_{(g_1, g_2) \in \mathcal{S}^2} L_{g_1, g_2}(\mathbf{x}, \mathbf{y}).$$

The weights w_1 and w_2 should be chosen to balance the contribution of the two components. A distance based on L_3 will tend to pick up sets of genes with separated means and differences in correlation in the two samples.

5.3 Random search for differentially expressed subsets of genes

As we mentioned in Section 1, selecting too many feature variables can deteriorate the performance of a discriminant rule. It is therefore natural to attempt at finding the ‘‘best subset’’ in accordance with some selection criterion. In discriminant analysis, the rate of misallocation of unclassified entities is the most widely used criterion for the choice of feature variables. Several useful error-based procedures have been proposed under the assumption of the homoscedastic normal model [1]. These procedures are formulated in the form of a statistical test with an adjustment for multiple testing. With stepwise selection procedures,

as noted by McKay and Campbell [42, 43], the tests are not independent and it is difficult to design a theoretically sound adjustment to control the simultaneous significance level for the sequence of tests.

The class of distances introduced in Section 5.2 is worth considering for its usefulness in selecting a reduced feature vector and testing for differentially expressed subsets of genes. The intention to examine all possible subsets in order to find the one for which the distance between two groups of entities is maximal meets with serious difficulties in practical settings. Whenever the size of a target subset is small, the feasibility of the branch-and-bound algorithm [44] merits evaluation. The algorithm guarantees finding a maximum and yet it is generally more efficient than the straightforward checking of all possibilities. The branch-and-bound method works best when the initial vector is close to the optimal, and when the intrinsic dimension of the feature space is small [44]. Fukunaga [44] provides empirical evidence that the method works well on uniformly distributed data when the intrinsic dimension is two and poorly when the intrinsic dimension is eight. The intrinsic dimension dim can be estimated based on the following relationship for the average distance to the ℓ^{th} nearest neighbor denoted by $E(d_\ell)$: $E(d_\ell)/E(d_{\ell+1}) = 1 + 1/(\ell \cdot dim)$. We have used Euclidean distance for several sets of rank-adjusted biological data and surprisingly, given the high external dimension, our initial estimates place the intrinsic dimension for the feature space at between four and six.

Since the number of possible subsets exponentially increases with the total number of genes, stepwise procedures seem to be an indispensable aid to variable selection. For relatively large subsets of genes, the issue of computational complexity can be resolved by applying random search methodology [45]. Random search methods can be easily implemented and they are rather insensitive to irregularities of the underlying optimization problem and to the presence of noise in the objective function; these properties make random search approach especially attractive for our purposes. Random search can be designed in a number of various ways. For example, simulated annealing [46] can be used for this purpose. Below we describe the basic structure of a simple random search algorithm for finding a subset (cluster) of size k with the largest distance between two classes (tissues):

1. Randomly select k genes to form the initial approximation; calculate the distance between the two classes for this cluster.
2. Replace at random one gene from the current cluster by a gene from outside the cluster; calculate the distance for this new cluster.
3. If the distance for the new cluster is larger than for the original cluster (improvement), keep the change, otherwise revert to the previous cluster.
4. Repeat steps 2 and 3 until convergence.

A modification of this algorithm with the aim of reducing selection bias is described in the next section.

5.4 Reduction of selection bias

An ever-present problem of variable selection is the danger of overfitting, that is finding overly specific patterns that do not extend to new samples. Cross-validation techniques provide a powerful tool to eliminate or at least largely reduce the effect of overfitting. Whenever a small number of variables is selected from a large set, one should expect a selection bias associated with choosing the optimal of a large number of subsets, regardless of the criterion used. To reduce this selection bias, Ganeshanandam and Krzanowski [47] suggested that cross-validation should precede the variable selection itself. Resorting to this idea and the well-known principles of the v -fold cross-validation (see, for example [48]), we have developed a ‘cross-validated search’ procedure that checks for reproducibility of its results. The basic structure of the algorithm is as follows:

1. Randomly divide the data into v groups of nearly equal size.
2. Drop one of the parts and find the optimal (in accordance with the chosen criterion) subset of genes using only the data from $v - 1$ groups.
3. Repeat step 2 in succession for each of the groups, obtaining ‘ v -optimal’ sets.
4. Combine these sets by selecting the genes with the highest frequencies of occurrence.

An alternative method for reducing the effect of overfitting is discussed in the paper by A.Chilingaryan et al. [49] published in this issue.

6 Multidimensional Two-sample Tests

The distance $\sqrt{N(\mu, \nu)}$ can be used to derive an asymptotic two-sample statistical test for the hypothesis $\mu = \nu$ (or $N(\mu, \nu) = 0$). Consider the case: $n_1 = n_2 = 2n$ and introduce the statistic

$$N_n = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n [2L(\mathbf{x}_i, \mathbf{y}_j) - L(\mathbf{x}_i, \mathbf{x}_{j+n}) - L(\mathbf{y}_i, \mathbf{y}_{j+n})], \quad (6)$$

which need not be a metric. Suppose the kernel L is chosen so that $0 \leq L \leq 1$ for all (\mathbf{x}, \mathbf{y}) . A relevant example is given by $L(\mathbf{x}, \mathbf{y}) = 1 - \exp(-\|\mathbf{x} - \mathbf{y}\|^2)$, where $\|\cdot\|$ is the Euclidean norm. Under the null hypothesis: $\mu = \nu$, the statistic N_n converges to 0 with probability 1 as $n \rightarrow \infty$. Using the central limit theorem it is easy to show that nN_n is asymptotically normal with a zero mean and variance $\sigma^2 \leq 8$. The latter inequality suggests a conservative asymptotic test for the hypothesis: $\mu = \nu$. While this reasoning is theoretically flawless the snag is that the particular half-sample chosen may be found to affect the results of practical application of the test to a substantial extent [50].

The alternative is to develop the corresponding statistical test using resampling techniques. The following line of reasoning leads to a statistic that seems to be suitable for parametric bootstrap. Let X_1, \dots, X_n, \dots be a sequence of independent and identically distributed (i.i.d.) d -dimensional random vectors with common distribution function $F(x)$ and characteristic function (ch.f.) $f(t) = \int_{R^d} \exp\{i\langle t, x \rangle\} dF(x) = \text{Re}f(t) + i\text{Im}f(t)$. Let $F_n(x)$ be the empirical distribution function of X_1, \dots, X_n . Denote by $f_n(t)$ the corresponding d -dimensional empirical ch.f.:

$$f_n(t) = \frac{1}{n} \sum_{j=1}^n \exp\{i\langle t, x \rangle\} = \int_{R^d} \exp\{i\langle t, x \rangle\} dF_n(x), \quad t \in R^d,$$

and define a d -variate empirical process

$$T_n(t) = \sqrt{n} (f_n(t) - f(t)) = \int_{R^d} \exp\{i\langle t, x \rangle\} d\beta_n(x), \quad (7)$$

where

$$\beta_n(x) = \sqrt{n} (F_n(x) - F(x)).$$

Consider a complex valued d -variate Gaussian random field $T_F(t) = U(t) + iV(t)$ with $ET_F(t) = 0$ and having the same cross-covariance matrix as T_n (for each n), i.e.,

$$E \begin{pmatrix} U(t)U(s) & U(t)V(s) \\ V(t)U(s) & V(t)V(s) \end{pmatrix} = \begin{pmatrix} \frac{\text{Re}f(t-s) + \text{Re}f(t+s)}{2} - \text{Re}f(t)\text{Re}f(s) & \frac{-\text{Im}f(t-s) + \text{Im}f(t+s)}{2} - \text{Re}f(t)\text{Im}f(s) \\ \frac{\text{Im}f(t-s) + \text{Im}f(t+s)}{2} - \text{Re}f(s)\text{Im}f(t) & \frac{\text{Re}f(t-s) - \text{Re}f(t+s)}{2} - \text{Im}f(t)\text{Im}f(s) \end{pmatrix},$$

and specifically $ET_F(t)\overline{T_F(s)} = f(t-s) - f(t)f(-s)$. The process $T_F(t)$ has the following stochastic integral representation (see, e.g. [51])

$$T_F(t) = \int_{R^d} \exp\{i\langle t, x \rangle\} dB_F(x),$$

where $B_F(x)$ is a d -variate Brownian bridge process associated with the distribution function F , i.e., B_F is a d -variate Gaussian process with the following properties:

$$EB_F(x) = 0, \quad EB_F(x)B_F(y) = F(x \wedge y) - F(x)F(y),$$

$$\lim_{x_j \rightarrow -\infty} B_F(x_1, \dots, x_d) = 0, \quad j = 1, \dots, d,$$

$$\lim_{(x_1, \dots, x_d) \rightarrow (\infty, \dots, \infty)} B_F(x_1, \dots, x_d) = 0,$$

where $x \wedge y = (\min(x_1, y_1), \dots, \min(x_d, y_d))$.

Denote

$$\phi(t) = (1 - \operatorname{Re}f(t))^{1/2}$$

and let λ be d -dimensional Lebesgue measure. Define

$$m(y) = \lambda\{t: \|t\| \leq \frac{1}{2}, \phi(t) < y\}, \quad 0 \leq y \leq 1.$$

Denote the inverse of $m(y)$ by

$$\tilde{\phi}(h) = \sup\{y: m(y) < h\},$$

and let K be a compact subset of R^d . Denote by $\mathcal{C}(K)$ the Banach space of all continuous complex valued functions on K with the usual sup-norm. Csörgő [51] showed that

T_n converges weakly to T_F in $\mathcal{C}(K)$ if and only if

$$\int_0^1 \frac{\tilde{\phi}(h)}{h(\log \frac{1}{h})^{1/2}} dh < \infty. \quad (8)$$

Now, we can consider the corresponding statistical problem. Let X, X_1, \dots, X_n be i.i.d. d -dimensional random vectors with common distribution function $F(x)$ and ch.f. $f(t)$. Suppose that Y, Y_1, \dots, Y_n are i.i.d. d -dimensional random vectors with common distribution function $G(x)$ and ch.f. $g(t)$. Denote by $\sqrt{M}(X, Y)$ the following special case of the metric \sqrt{N} :

$$\begin{aligned} M(X, Y) &= 2E\|X - Y\|^r - E\|X - X'\|^r - E\|Y - Y'\|^r \\ &= \frac{1}{2}c_{d,r} \int_{R^d} |f(t) - g(t)|^2 \frac{dt}{\|t\|^{r+d}}, \end{aligned}$$

where $c_{d,r}$ is some (known) constant. We wish to test the hypothesis H_0 that the distribution of the vector X is identical to the distribution of Y . In terms of the distance \sqrt{M} the hypothesis assumes the form

$$2E\|X - Y\|^r - E\|X - X'\|^r - E\|Y - Y'\|^r = 0, \quad (9)$$

where X' is an i.i.d. copy of X , and Y' is an i.i.d. copy of Y .

The choice of r is driven by a particular application. For example, a reasonable choice would be $r = 1$, since it is usually believed that the first moment of the underlying probability distribution is finite.

An empirical analog of the left hand side of (9) has the form

$$S_n = \frac{1}{n^2} \sum_{j=1}^n \sum_{k=1}^n (2\|X_j - Y_k\|^r - \|X_j - X_k\|^r - \|Y_j - Y_k\|^r). \quad (10)$$

It follows now that

$$S_n = \frac{1}{2}c_{d,r} \int_{R^d} |f_n(t) - g_n(t)|^2 \frac{dt}{\|t\|^{r+d}} \geq 0.$$

Clearly, $S_n \rightarrow 2E\|X - Y\|^r - E\|X - X'\|^r - E\|Y - Y'\|^r$ as $n \rightarrow \infty$. Under H_0 we have $N(X, Y) = 0$, so that $S_n - (E\|X - Y\|^r - E\|X - X'\|^r - E\|Y - Y'\|^r) = S_n$. The distribution of nS_n is identical with

$$\frac{1}{2}c_{d,r} \int_{R^d} |\sqrt{n}(f_n(t) - f(t)) - \sqrt{n}(g_n(t) - g(t))|^2 \frac{dt}{\|t\|^{r+d}}.$$

Suppose the condition (8) holds true. Then, from the result mentioned above (see, e.g. [51]), it follows that the limiting distribution of nS_n coincides with the distribution of the following functional,

$$S_F = \frac{1}{2}c_{d,r} \int_{R^d} |T_F(t) - T_G(t)|^2 \frac{dt}{\|t\|^{r+d}}, \quad (11)$$

and percentiles of the corresponding sample distribution can be obtained by an appropriate computer-intensive method.

7 Classification Using a Reduced Feature Vector

Once a sufficiently low dimension (compared to the number of samples) feature vector is chosen, a classification rule can be developed using one of many discriminant analysis approaches. However it makes sense to take advantage of the optimal character of the reduced feature vector: the multivariate distance between two tissues calculated for this subset of genes is maximal. For this we need to define a probability distance between a single point in d -space and a set of points in the same space. Such a distance, say $\hat{N}(\mathbf{x}, \nu)$, derives from the metric $\sqrt{N(\mu, \nu)}$, defined by formula (5.1), if either μ or ν is a delta-measure. Then the classification rule will assign an ‘unknown’ vector to the group to which it is closest in terms of the metric.

8 Simulation Studies

Our model for microarray data simulates the underlying stochastic mechanism by superimposing biological sample-to-sample variability and technology-related errors on the ‘true’ gene expression levels. The goal is to generate samples from “normal” and “pathological” tissues where a known number of genes shows differential expression with respect to both the marginal means and the correlation structure.

8.1 Ideal gene expression

A set of 100 genes was divided into groups of size 10; within each group the gene expressions are dependent (see below for details), however genes included in different groups are independent. To describe the gene-to-gene variability of expression levels, for each gene a “normal” mean expression level a_i was generated from a mixture of log-normal distribution with mean $\mu = 2.5$ and standard deviation (STD) $\sigma = 3$, and uniform distribution on $[0, 10]$. The proportion of the uniform distribution in the mixture was taken to be $\pi = 0.15$. One of the gene clusters was selected to be differentially expressed. Within this cluster we modified the mean values b_i for the “pathological” tissue by multiplying each of them by a gene-specific ratio d_i generated from a log-normal distribution. The mean and variability of this ratio was changed according to the requirements of the simulation, mean 1 and variance 0 implying no marginal difference.

8.2 Biological variability

Within-cluster correlation and sample-to-sample variability were introduced as described below. Gene expression levels A_{ij} and B_{ij} for each sample were generated from a log-normal distribution with the ideal mean a_i , b_i determined at the previous step and coefficient of variation $v = 0.4$. Dependence between random variables was introduced by generating for each cluster a set of exchangeable standard normal variables with correlation coefficient $\rho = 0.9$ for each sample. The required marginal distributions were obtained by a suitable linear transformation followed by exponentiation. The log-normal distribution with a constant coefficient of variation is consistent with our and other researchers’ experience with microarray data. The chosen value of $v = 0.4$ mimics the high variability inherent in biological samples.

To introduce changes in correlation structure between the two tissues, the expression levels in the differentially expressed clusters of the “pathological” tissue were generated as independent random variables. This setup simulates loss of regulation of a pathway.

8.3 Simulation results

To evaluate the performance of multidimensional variable selection and the role of correlation structure of gene expression data, we set up a simulation study using the model described in the previous section. Technology related experimental errors were not introduced as we did not want the choice of adjustment to confound the results.

First we designed a scenario where marginal methods could not possibly work: out of 100 simulated genes one cluster of size 10 was chosen to be “differentially expressed” and these genes were set to have no marginal difference (that is $d_i \equiv 1$ was chosen), but their joint distribution was changed from highly correlated to uncorrelated. The rest of the genes had the same joint distributions in both tissues. In each of the 50 simulation runs we generated 24 samples for both tissues and selected k , $k = 1, \dots, 10$ most differentially

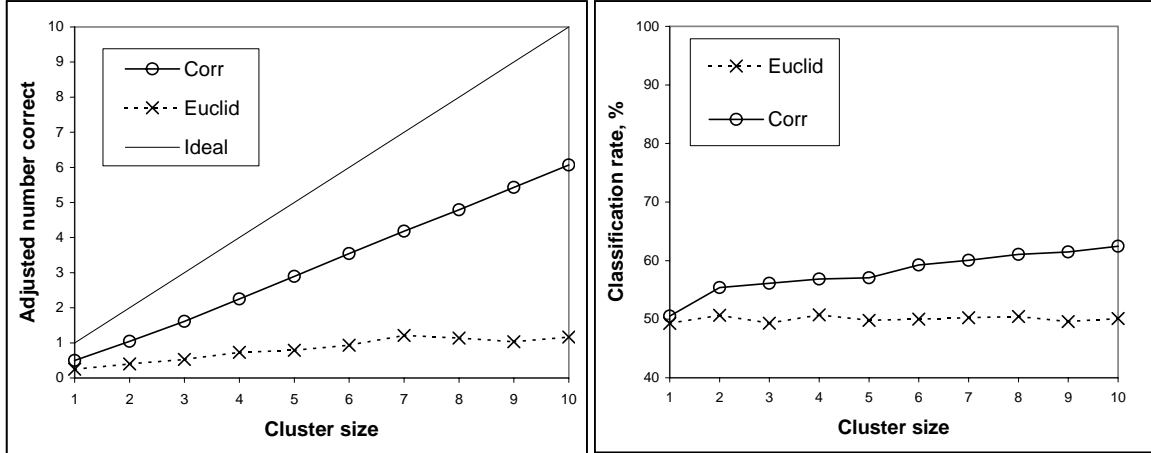


Figure 2: Comparison of two kernels for the search procedure when there is only correlation structure difference. The search used cluster size 10. The x -axis shows the number of genes chosen for inclusion in the final list. Left panel: the number of genes from the differentially expressed cluster that were included in the final list adjusted for chance. Right panel: test set classification rates.

expressed genes by a marginal t -statistic and by cross-validated random search using the estimated distance $\sqrt{N(\mu, \nu)}$ with the kernel L_3 sensitive to correlation changes (with L_1 being the Euclidean distance). For each selection we estimated the correct classification rate on a test set of size 100 (50 samples of each tissue) and calculated the proportion p_{obs} of the genes from the differentially expressed cluster in the finally selected set. Because there is *no* marginal difference in gene expression one might expect that the average proportion of “differentially expressed” genes from marginal selection would be around 10% and the test set classification rate would be equivalent to a random choice, that is around 50%. In our simulation experiments, however, the marginal selection picked up the differential cluster quite well: in more than 25% of our simulations the highest t -statistic was associated with a gene from that cluster! The reason has to do with the fact that the t -statistics generated in this setting are not independent. In other words, while the marginal distributions of the t -statistics are identical, the simulated gene expressions are correlated and so are the test statistics. More importantly, this correlation is different in the cluster where only one of the tissues has correlated genes. Since variability was the highest in the “differentially expressed” cluster a maximum is more likely to occur in this rather than in another cluster. To account for this effect the results of the multivariate search were rescaled using a transformation analogous to the Kappa statistic:

$$n_{adj} = k \frac{p_{obs} - p_{baseline}}{1 - p_{baseline}}. \quad (12)$$

where k is the cluster size and $p_{baseline}$ is the proportion of differentially expressed genes among the genes with the top k (absolute) values of the t -statistic. The results are presented in Figure 2. The straight diagonal line represents the ideal case of selecting only genes from the differentially expressed cluster; the value of 0 corresponds to random chance. The multivariate method shows a remarkable success in finding the changed cluster and even provides some improvements in classification. Classification is clearly difficult under this setup, as correlation structure cannot be estimated based on one classifiable observation.

Next we were interested in determining whether the multivariate approach would continue providing advantage over the marginal selection when marginal differences were introduced. In these simulations the ratio, d_i , controlling differential expression in the preselected cluster was generated from a log-normal distribution with mean 1 and STD 0.3. Figure 3 compares the results of selection by the marginal t -statistic and using the multidimensional selection with Euclidean distance kernel and kernel L_3 . The values have been adjusted according to Eq. (12) using the baseline estimated in the previous simulations. The marginal method finds the genes that happen to have a large value of d_i , while the joint selection is capable of picking up genes that have less pronounced differential expression but belong to the cluster whose correlation structure has been changed.

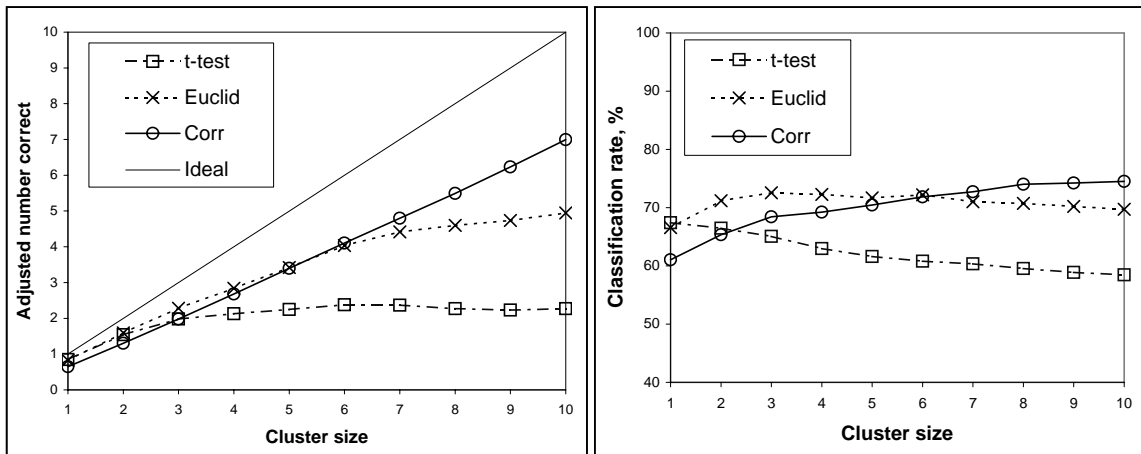
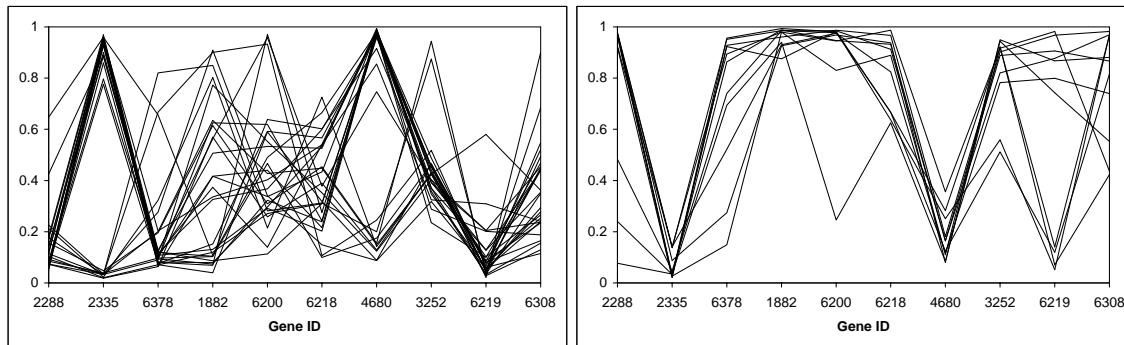


Figure 3: Comparison of two kernels for the search procedure and a marginal t -statistic when there is both marginal mean and correlation structure difference. The search used cluster size 10. The x -axis shows the number of genes chosen for inclusion in the final list. Left panel: the number of genes from the differentially expressed cluster that were included in the final list adjusted for chance. Right panel: test set classification rates.

9 AML versus ALL classification

We have applied our methodology to two leukemia datasets. The first set of data (Data Set 1) was the same as that analyzed by Golub et al. [28]. It includes 27 ALL (acute lymphoblastic leukemia) and 11 AML (acute myeloid leukemia) samples processed using Affymetrix oligonucleotide microarrays. A test set of 34 samples is also available. Golub et al. have shown that these two classes could be separated quite well using 10 or more genes as predictors. We employed a 10-fold cross-validated search for the best subset of genes maximizing the estimated distance $D = \sqrt{N(\mu, \nu)}$ with the Euclidean distance kernel; the search was repeated 50 times with 10,000 iterations each to find the most differentially expressed cluster of size 10. The procedure was applied to rank-adjusted data. The list of the selected genes together with a line plot of the corresponding expression levels is given in Figure 4. Three of these genes (marked with a star) were also included in the group of 50 predictors by Golub et al. This set of genes provides a 95% cross-validated correct classification rate and the prediction on the test set is perfect with the exception of two samples where a decision is not made due to an extremely low prediction strength (the same is true for genes selected by Golub et al.). The prediction strength was calculated as $PS = |D_1 - D_2| / \max(D_1, D_2)$, where D_1 and D_2 are distances between the sample to be classified and each of the two classes. It measures how confident one can be when classifying the sample into one of the groups. The same classification performance was achieved using only two top genes. In this case, inclusion of additional genes led to a decrease in the prediction strength for the difficult to classify samples. A striking feature of the plot in Figure 4 is that the ALL samples appear to be divided into two groups. A closer look led us to the conclusion that these groups correspond to the T-cell/B-cell division of the ALL samples. Our analysis suggests two genes (# 2335, 4680) for discrimination between the groups; these genes are well known as markers for T-cell leukemia. Note that a marginal search would never turn up these genes, since taken individually they misclassify B-cell ALL samples, however their sensitivity to T-cell leukemia samples makes them valuable predictors in multivariate classification.

The second set of data (Data Set 2) that we analyze has been collected at the Primary Children’s Medical Center in Utah from children newly diagnosed with leukemia. A detailed description of the data and microarray procedures can be found in [52]. Probes of mRNA extracted from bone marrow samples were hybridized to a microarray spotted with a minimally redundant set of 4608 cDNA clones. Only a fraction of the clones have been sequence verified, so the identification of the genes is not completely reliable. The patient samples were all hybridized on the red channel, the green channel contained samples from the HL-60 cell line. Since the data contained only a few T-cell ALL samples, our focus was on B-cell ALL. As we only had 10 AML, but 25 B-cell ALL samples, the training set was formed to contain 10 samples of each kind, and the test set was represented by the remaining 15 B-cell ALL samples.



2288* D component of complement (adipsin)
 2335 Immunoglobulin-associated beta (B29)
 6378 NF-IL6-beta protein mRNA
 1882* Cystatin C
 6200* Interleukin 8 (IL8) gene
 6218 Elastase 2, neutrophil
 4680 TCL1 gene (T cell leukemia)
 3252 Glutathione S-transferase
 6219 Neutrophil elastase gene, exon 5
 6308 GRO2 oncogene

Figure 4: Rank adjusted expression of genes in Data Set 1. The left panel shows the ALL samples, the right panel the AML samples. The genes are listed in order of decreasing frequency of occurrence in the selected subset.

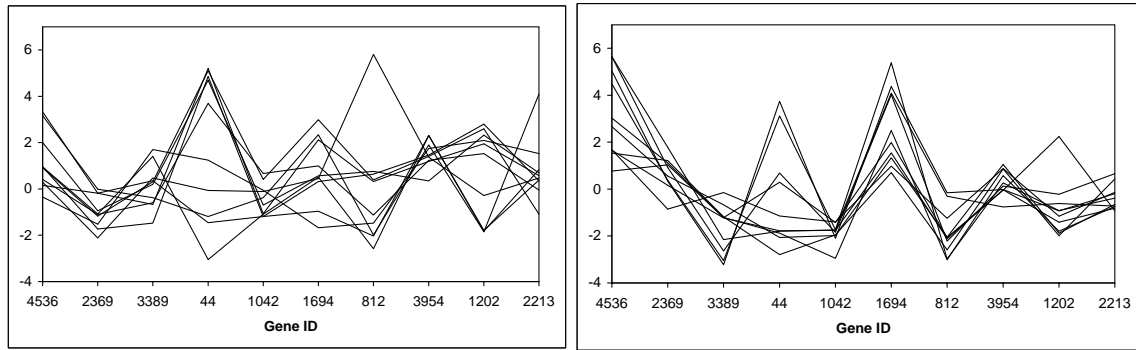
We used two adjustments for this dataset, namely the adjustment for a multiplicative random effect and the rank-based adjustment. Since we found considerable fluctuations among the pins used to deposit the mRNA, both adjustments were performed separately for each of the 12 pins (384 spots each). Figure 5 shows the results based on the random effect adjustment and Figure 6 the results based on ranks. As previously, we use stars to mark the genes that are present in the list by Golub et al. As this is an independent data set obtained from a significantly smaller clone set, we did not expect to find many of the same genes. In addition to the previously found common genes one more gene (*Glutathione S-transferase*) appeared on the list of top 10 genes for both data sets. We found two genes (marked with ⁺) selected from Data Set 2 with both adjustments. This overlap tends to increase with increasing the length of the lists. Another encouraging feature of the lists is the inclusion of two different copies of the same gene in one of them and of two strongly related genes in the other one.

As the plots in Figures 5 and 6 show, with Data Set 2 the two types of leukemia are not separated as well as with Data Set 1. This is reflected in the classification performance as well. With the multiplicative effect adjustment the prediction levels out after the first four genes at 85% cross-validation and 66.7% test set classification rate. The rank adjustment performs better: based on the first eight genes the cross-validated estimate of the error rate is 100% with 73.3% of the test set samples being correctly classified as B-cell ALL. Setting the sought-for subset size at either 5 or 15 genes did not change significantly our findings resulted from the search for the best subset of 10 genes.

The computations were performed on a IBM-compatible computer with two 1000 MHz Pentium III processors. A typical search ($50 \times 10,000$ iterations with cluster size 10) was carried out in 20 minutes.

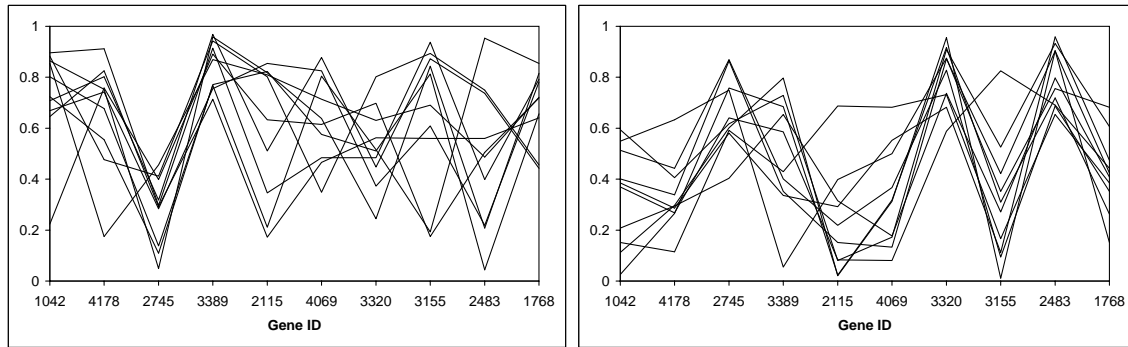
Acknowledgment

This research was supported by Grant AB2-2005 awarded by the U.S. Civilian Research and Development Foundation (CRDF), NCI Cancer Center Support Grant 5P30 CA42014, and the Huntsman Cancer Foundation. The research of Dr. Klebanov was supported in part by Grant MSM 113200008 from the Czech Republic. We would like to thank Dr. A. Chilingaryan for fruitful discussions.



4536* Hematopoetic proteoglycan core prot.
 2369 ESTs
 3389+ Mad homolog Smad1
 44 ESTs
 1042+ Homo Sapiens P5-1
 1694* Hematopoetic proteoglycan core prot.
 812 Homo sapiens CASK
 3954* Topoisomerase (DNA) II beta
 1202 ESTs
 2213 CD53 antigen

Figure 5: Expression levels of genes in Data Set 2 adjusted for multiplicative random effect. The left panel shows the (B-cell) ALL samples, the right panel the AML samples. The genes are listed in order of decreasing frequency of occurrence in the selected subset.



1042+ Homo Sapiens P5-1
 4178 Membrane metallo-endopeptidase
 2745 H.sapiens p63
 3389+ Mad homolog Smad1
 2115 ESTs
 4069 Myocyte specific enhancer factor 2
 3320 Glutathione S-transferase
 3155 ESTs
 2483 ESTs
 1768 ESTs

Figure 6: Rank adjusted expression levels of genes in Data Set 2. The left panel shows the (B-cell) ALL samples, the right panel the AML samples. The genes are listed in order of decreasing frequency of occurrence in the selected subset.

References

- [1] G. J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*, Wiley, New York, 1992.
- [2] D. B. Carr, R. Somogyi, G. Michaels, Templates for looking at gene expression clustering, *Statistical Computing & Statistical Graphics Newsletter* 8 (1) (1997) 20–29.
- [3] M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci. USA* 95 (1998) 14863–14868.
- [4] R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart, R. W. Davis, A genome-wide transcriptional analysis of the mitotic cell cycle, *Molecular Cell* 2 (1998) 65–73.
- [5] T. R. Hughes, M. J. Marton, A. R. Jones, C. J. R. and Roland Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Dai, Y. D. He, M. J. Kidd, A. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepaniants, D. D. Shoemaker, D. Gachotte, K. Chakraborty, J. Simon, M. Bard, S. H. Friend, Functional discovery via a compendium of expression profiles, *Cell* 102 (2000) 109–126.
- [6] G. S. Michaels, D. B. Carr, M. Askenazi, S. Fuhrman, X. Wen, R. Somogyi, Cluster analysis and data visualization of large-scale gene expression data, *Pacific Symposium on Biocomputing* 3 (1998) 42–52.
- [7] A. J. Butte, I. S. Kohane, Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements, in: *Proceedings of the Pacific Symposium on Biocomputing*, World Scientific, Singapore, 2000, pp. 418–429.
- [8] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, A. J. Levine, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *Proc. Natl. Acad. Sci. USA* 96 (12) (1999) 6745–6750.
- [9] S. Tavazoie, J. Hughes, M. Campbell, R. Cho, G. Church, Systematic determination of genetic network architecture, *Nature Genetics* 22 (3) (1999) 281–285.
- [10] H. Herzel, D. Beule, S. Kielbasa, J. Korbel, C. Sers, A. Malik, H. Eickhoff, H. Lehrach, J. Schuchhardt, Extracting information from cDNA arrays, *Chaos* 11 (1) (2001) 98–107.
- [11] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, T. R. Golub, Interpreting patterns of gene expression with self-organizing maps: methods and application to hemopoietic differentiation, *Proc. Natl. Acad. Sci. USA* 96 (1999) 2907–2912.
- [12] A. Ben-Dor, R. Shamir, Z. Yakhini, Clustering gene expression patterns, *Journal of Computational Biology* 6 (3–4) (1999) 281–97.
- [13] M. J. van der Laan, J. F. Bryan, Gene expression analysis with the parametric bootstrap, *Public Health series* 86, University of California, Berkeley (June 2000).
- [14] L. J. Heyer, S. Kruglyak, S. Yoosheph, Exploring expression data: identification and analysis of co-expressed genes, *Genome Research* 9 (1999) 1106–1115.
- [15] R. Sharan, R. Shamir, CLICK: A clustering algorithm with applications to gene expression analysis, in: *Proc. ISMB*, AAAI Press, Menlo Park, CA, 2000, pp. 307–316.
- [16] T. Hastie, R. Tibshirani, M. B. Eisen, A. Alizadeh, R. Levy, L. Staudt, W. C. Chan, D. Botstein, P. Brown, 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns, *Genome Biology* 1 (2) (2000) 0002.1–0003.21.
- [17] T. Hastie, R. Tibshirani, D. Botstein, P. Brown, Supervised harvesting of expression trees, Tech. rep., Stanford University, <http://www-stat.stanford.edu/~tibs> (August 2000).

- [18] J. DeRisi, L. Penland, P. O. Brown, M. L. Bittner, P. S. Meltzer, M. Ray, Y. Chen, Y. A. Su, J. M. Trent, Use of a cDNA microarray to analyse gene expression patterns in human cancer, *Nature Genetics* 14 (4) (1996) 457–460.
- [19] M. Schena, D. Shalon, R. Davis, P. Brown, Quantitative monitoring of gene expression patterns with a complementary dna microarray., *Science* 270 (5235) (1995) 467–470.
- [20] M. Schena, D. Shalon, R. Heller, A. Chai, P. Brown, R. Davis, Parallel human genome analysis: microarray-based expression monitoring of 1000 genes, *Proc Natl Acad Sci USA* 93 (20) (1996) 10614–10619.
- [21] M. A. Newton, C. M. Kendziorski, C. S. Richmond, F. R. Blattner, K. W. Tsui, On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data, *Journal of Computational Biology* 8 (1) (2000) 37–52.
- [22] M. K. Kerr, M. Martin, G. A. Churchill, Analysis of variance for gene expression microarray data, *Journal of Computational Biology* 7 (6) (2000) 819–837.
- [23] M. K. Kerr, G. A. Churchill, Experimental design for gene expression microarrays, Tech. rep., The Jackson Laboratory, <http://www.jax.org/research/churchill/pubs/index.html> (August 2000).
- [24] C. H. Rhee, K. Hess, J. Jabbur, M. Ruiz, Y. Yang, S. Chen, A. Chenchik, G. N. Fuller, W. Zhang, cDNA expression array reveals heterogeneous gene expression profiles in three glioblastoma cell lines, *Oncogene* 18 (1999) 2711–2717.
- [25] M.-L. T. Lee, F. C. Kuo, G. A. Whitmore, J. Sklar, Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations, *Proc. Natl. Acad. Sci. USA* 97 (18) (2000) 9834–9839.
- [26] A. Tsodikov, A. Szabo, D. Jones, Adjustments and tests for differential expression, ENAR conference, Charlotte NC (March 2001).
- [27] A. Tsodikov, A. Szabo, D. Jones, Adjustments and measures of differential expression for microarray data, *Bioinformatics* 18 (2002).
- [28] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, E. S. Lander, Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, *Science* 286 (1999) 531–537.
- [29] S. Dudoit, J. Fridlyand, T. P. Speed, Comparison of discrimination methods for the classification of tumors using gene expression data, Tech. Rep. 576, University of California, Berkeley (June 2000).
- [30] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, Z. Yakhini, Tissue classification with gene expression profiles, *Journal of Computational Biology* 7 (2000) 559–584.
- [31] A. Ben-Dor, N. Friedman, Z. Yakhini, Scoring genes for relevance .
- [32] J. Khan, R. Simon, M. Bittner, Y. Chen, S. B. Leighton, T. Pohida, P. D. Smith, Y. Jiang, G. C. Gooden, J. M. Trent, P. S. Meltzer, Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays, *Cancer Research* 58 (1998) 5009–5013.
- [33] D. T. Ross, U. Scherf, M. B. Eisen, C. M. Perou, C. Rees, P. Spellman, V. Iyer, S. S. Jeffrey, M. Van de Rijn, M. Waltham, A. Pergamenschikov, J. C. F. Lee, D. Lashkari, D. Shalon, T. G. Myers, J. N. Weinstein, D. Botstein, P. O. Brown, Systematic variation in gene expression patterns in human cancer cell lines, *Nature Genetics* 24 (2000) 227–235.
- [34] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, J. J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, L. M. Staudt, Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling, *Nature* 403 (6769) (2000) 503–511.

- [35] A. Kagan, Y. Linnik, C. Rao, *Characterization Problems in Mathematical Statistics*, Nauka, Moscow, 1972.
- [36] G. A. F. Seber, *Linear Regression Analysis*, John Wiley and Sons, New York, 1977.
- [37] M. S. Bartlett, Fitting a straight line when both variables are subject to error, *Biometrics* 5 (1949) 207–213.
- [38] T. Robertson, F. T. Wright, R. L. Dykstra, *Order Restricted Statistical Inference*, Wiley, London, 1988.
- [39] S. T. Rachev, *Probability Metrics and the Stability of Stochastic Models*, Wiley, Chichester, 1991.
- [40] R. Gnanadesikan, *Methods of Statistical Data Analysis of Multivariate Observations*, 2nd Edition, Wiley, New York, 1997.
- [41] A. A. Zinger, L. B. Klebanov, A. V. Kakosyan, *Stability Problems for Stochastic Models*, VNIISI, Moscow, 1989, Ch. Characterization of distributions by mean values of statistics in connection with some probability metrics, pp. 47–55.
- [42] R. McKay, N. Campbell, Variable selection techniques in discriminant analysis I. Description., *Br. J. Math. Statist. Psychol.* 35 (1982) 1–29.
- [43] R. McKay, N. Campbell, Variable selection techniques in discriminant analysis II. Allocation., *Br. J. Math. Statist. Psychol.* 35 (1982) 30–41.
- [44] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd Edition, Academic Press, London, 1990.
- [45] A. A. Zhigljavsky, *Theory of global random search*, Vol. 65 of *Mathematics and its Applications (Soviet Series)*, Kluwer Academic Publishers Group, Dordrecht, 1991.
- [46] D. T. Pham, D. Karaboga, *Intelligent optimisation techniques : genetic algorithms, tabu search, simulated annealing and neural networks*, Springer, London, New York, 2000.
- [47] S. Ganeshanandam, W. Krzanowski, On selecting variables and assessing their performance in linear discriminant analysis, *Austral. J. Statist.* 32 (1989) 443–447.
- [48] L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, *Classification and regression trees*, Wadworth and Brooks/Cole Advanced Books and Software, Monterey, CA, 1984.
- [49] A. Chilingaryan, N. Gevorgyan, A. Vardanyan, D. Jones, A. Szabo, Multivariate approach for selecting sets of differentially expressed genes, *Mathematical Biosciences* (2002) in press.
- [50] J. Durbin, *Distribution Theory for Tests Based on the Sample Distribution Function*, Society for Industrial and Applied Mathematics, Philadelphia, 1973.
- [51] S. Csörgő, Multivariate empirical characteristic functions, *Z. Wahrscheinlichkeitstheorie verw. Gebiete* 35 (1981) 203–229.
- [52] E. A. R. Philip J. Moos, M. A. Carlson, A. Szabo, F. E. Smith, S. P. Hunger, Q. Wei, C. Willman, W. L. Carroll, Identifiavation of gene expression profiles that segregate patients with childhood leukemia, submitted .