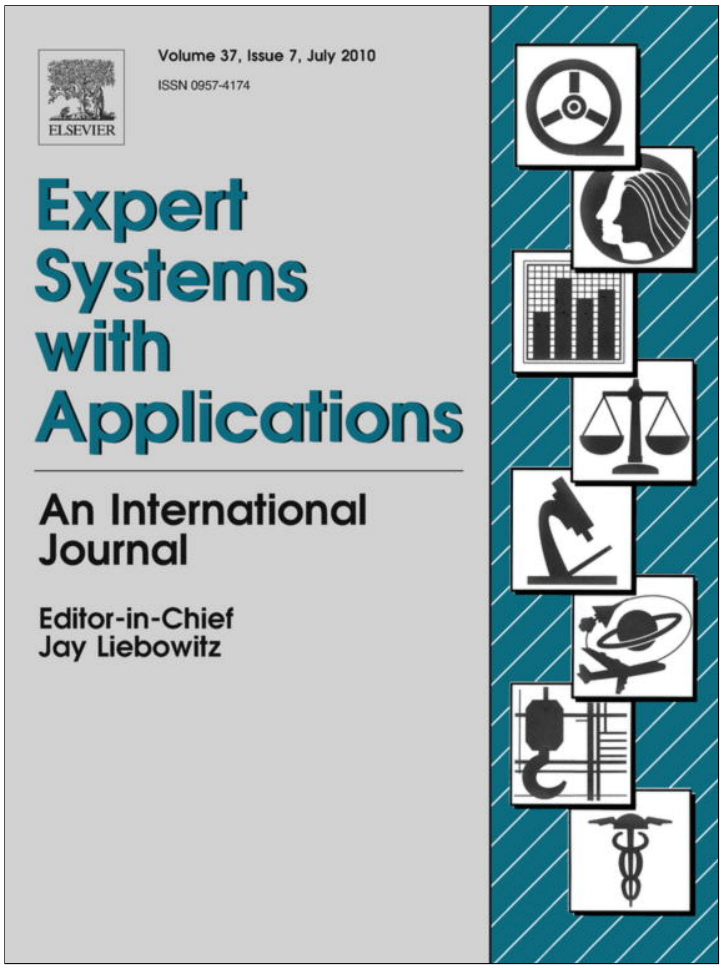

Expert Systems with Applications 37 (2010) 5178–5185

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Rasch-based high-dimensionality data reduction and class prediction with applications to microarray gene expression data

Andrej Kastrin *, Borut Peterlin *

Institute of Medical Genetics, University Medical Centre Ljubljana, Šlajmerjeva 3, SI-1000 Ljubljana, Slovenia

ARTICLE INFO

Keywords:
High-dimensional data
Feature extraction
Gene expression
Class prediction
Rasch model

ABSTRACT

Class prediction is an important application of microarray gene expression data analysis. The high-dimensionality of microarray data, where number of genes (variables) is very large compared to the number of samples (observations), makes the application of many prediction techniques (e.g., logistic regression, discriminant analysis) difficult. An efficient way to solve this problem is by using dimension reduction statistical techniques. Increasingly used in psychology-related applications, Rasch model (RM) provides an appealing framework for handling high-dimensional microarray data. In this paper, we study the potential of RM-based modeling in dimensionality reduction with binarized microarray gene expression data and investigate its prediction accuracy in the context of class prediction using linear discriminant analysis. Two different publicly available microarray data sets are used to illustrate a general framework of the approach. Performance of the proposed method is assessed by re-randomization scheme using principal component analysis (PCA) as a benchmark method. Our results show that RM-based dimension reduction is as effective as PCA-based dimension reduction. The method is general and can be applied to the other high-dimensional data problems.

1. Introduction

With decoding of the human genome and other eukaryotic organisms molecular biology has entered into a new era. High-throughput technologies, such as genomic microarrays can be used to measure the expression levels of essentially all the genes within an entire genome scale simultaneously in a single experiment and can provide information on gene functions and transcriptional networks (Cordero, Botta, & Calogero, 2007). The major challenge in microarray data analysis is due to their size, where the number of genes or variables (p) far exceeds the number of samples or observations (n), commonly known as the "large p, small n" problem. This takes it difficult or even impossible to apply class prediction methods (e.g., logistic regression, discriminant analysis) to the microarray data.

Class prediction is a crucial aspect of microarray studies and plays important role in the biological interpretation and clinical application of microarray data (Chen, 2007; Larrañaga et al., 2008). For the last few years, microarray-based class prediction has been a major topic in applied statistics (Slawski, Daumer, & Boulesteix, 2008). In a class prediction study, the task is to induce a class predictor (classifier) using available learning samples (i.e., gene expression profiles) from different diagnostic classes. Given the learning samples representing different classes, first the classifier is learned and then the classifier is used to predict the class membership (i.e., diagnostic class) of unseen samples (Asyali, Colak, Demirkaya, & Inan, 2006).

Generally, the performance of a classifier depends on three factors: the sample size, number of variables, and classifier complexity (Jain, Duin, & Mao, 2000; Raudys, 2006). It was shown that for the fixed sample size, the prediction error of a designed classifier decreases and then increases as the number of variables grows. This paradox is referred to as the peaking phenomenon (Hughes, 1968). Moreover, some well known classifiers are even inapplicable in the setting of high-dimensional data. For example, the pooled within-class sample covariance in linear discriminant analysis (LDA) is singular if number of variables exceeds the number of samples. Similarly, in logistic regression the Hessian matrix will not have full rank and statistical packages will fail to produce reliable regression estimates (Zhang, Fu, Jiang, & Yu, 2007). Therefore, the number of samples must be larger than the number of variables for good prediction performance and appropriate use of classifiers. This naturally calls for the reduction of the ratio of sample size to dimensionality.

There are two major ways to handle high-dimensional microarray data in the class prediction framework. The first approach is to eliminate redundant or irrelevant genes (a.k.a. feature

* Corresponding authors. Tel.: +386 1 522 60 57; fax: +386 1 540 11 37 (A. Kastrin); tel./fax: +386 1 540 11 37 (B. Peterlin).
E-mail addresses: andrej.kastrin@guest.arnes.si (A. Kastrin), borut.peterlin@guest.arnes.si (B. Peterlin).

0957-4174/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2009.12.074

selection). The idea is to find genes with maximal discrimination performance and induce a classifier using those genes only (Asyali et al., 2006). The most commonly used procedures of feature selection are based on simple statistical tests (e.g., fold change, *t*-test, ANOVA, etc.), which are calculated for all genes individually, and genes with the best scores are selected for classifier construction (Dupuy & Simon, 2007; Jeffery, Higgins, & Culhane, 2006). The advantages of this approach are its simplicity, low computational cost, and interpretability. An alternative approach to overcome the problem of high-dimensionality is application of dimension reduction techniques (a.k.a. feature extraction). Generally, the aim of dimension reduction procedures is to summarize the original p-dimensional gene space in a form of a lower K-dimensional gene components space ($K < n$) that account for most of the variation in the original data (Jain et al., 2000). Most commonly used methods for feature extraction with microarray gene expression data are principal component analysis (PCA) (Alter, Brown, & Botstein, 2000; Chiaromonte & Martinelli, 2002; Holter et al., 2000), partial least squares (PLS) (Boulesteix, 2004; Boulesteix & Strimmer, 2007; Nguyen & Rocke, 2002, 2004), and sliced inverse regression (SIR) (Antoniadis, Lambert-Lacroix, & Leblanc, 2003; Bura & Pfeiffer, 2003; Chiaromonte & Martinelli, 2002). Although statistical analysis dealing with microarray data has been one of the most investigated areas in the last decade, there are only a few papers addressing the development and experimental validation of new methods and techniques for microarray dimension reduction. As Fan and Li (2006) claimed, the high-dimensional data analysis will be one of the most important research topics in statistics in the nearest future. Here, we fill this gap by proposing a latent variable approach for handling high-dimensional microarray data and show its promising potential for class prediction.

2. Background

The conceptual framework on latent variable modeling originates from psychometrics, starting at the beginning of the 20th century (Fischer & Molenaar, 1995). Utility of these models in biomedical research has only quite recently been recognized (Li & Hong, 2001; Rabe-Hesketh & Skrondal, 2008). By latent variable model we mean any statistical model that relates a set of observed variables to set of latent variables (De Boeck & Wilson, 2004). A latent variable is a variable that is not directly observable but does have a measurable impact on observable variables. Latent variables describe features that underlie the data. For example, a child's intelligence (i.e., latent variable) is typically assessed by measuring their answers to solving problems or items (i.e., observed variables) on intelligence test. The more items we ask of the child and the wider the breadth of items is included in the assessment, the more our understanding of that child's intellectual ability will be accurate.

The Rasch model (RM), originally proposed by Rasch (1966), is the simplest latent variable model. The idea behind the RM is that the probability of getting an item correct is a function of a latent trait or ability. For example, a child with higher intellectual ability would be more likely to correctly respond to a given item on an intelligence test. In psychological applications, data are usually given in a matrix, with rows being participants and columns being responses to a set of items. Microarray gene expression data can be represented in a similar way: columns are used to represent genes and rows are used to represent expression levels in biological samples. The RM can therefore be used to explain the observed gene expression patterns over different samples.

We assume that gene expression levels vary under the influence of K latent gene factors. The idea behind our approach is to partition p genes into K functional subgroups and that covariations between genes with similar expression (i.e., genes in the same partition) could be described with a single gene factor. The factors in the model are latent, unobserved variables that account for the covariation among genes. It is assumed that genes with similar expression patterns might share biological function or might be under common regulatory control (Do & Choi, 2008). Moreover, it is particularly interesting to model a large set of genes as functions of fewer gene factors, because biologist believe the changes of mRNA levels are due to some regulatory factors (Orlando et al., 2008). Specifically, regulatory factors are proteins that bind certain DNA elements to regulate gene transcription to mRNA. Therefore, the gene factors obtained from latent modeling of gene expression could be interpreted as latent measurements of common regulatory factors of related genes. The number of gene factors is considered to be a meta-parameter and must be estimated or directly supplied based on researcher's prior knowledge. RM can then be used to estimate the magnitude of gene factors. Class prediction using standard prediction methods can then be carried out in the reduced space by using constructed gene factors as predictor variables.

The main objective of this paper is to evaluate the potential of RM-based dimensionality reduction with microarray gene expression data and investigate its prediction accuracy in the context of class prediction using LDA. To validate the proposed approach we apply a parallel PCA-based dimension reduction.

3. Methods

We propose a framework for dimension reduction and class prediction with application to gene expression data as illustrated in Fig. 1. Our procedure consists of two basic steps: the first step is dimension reduction, in which data are reduced from high p-dimensional gene space to a lower K-dimensional gene factor space; the second step is class prediction, in which response classes are predicted using a class prediction method on the extracted gene factors.

3.1. Data sets and preprocessing

We apply our algorithms to two publicly available microarray data sets which have been considered before by several authors. The Leukemia data set (Golub et al., 1999) contains $n = 72$ tissue samples with $p = 7129$ genes: 47 samples of acute lymphoblastic leukemia (ALL) and 25 samples of acute myeloid leukemia (AML).

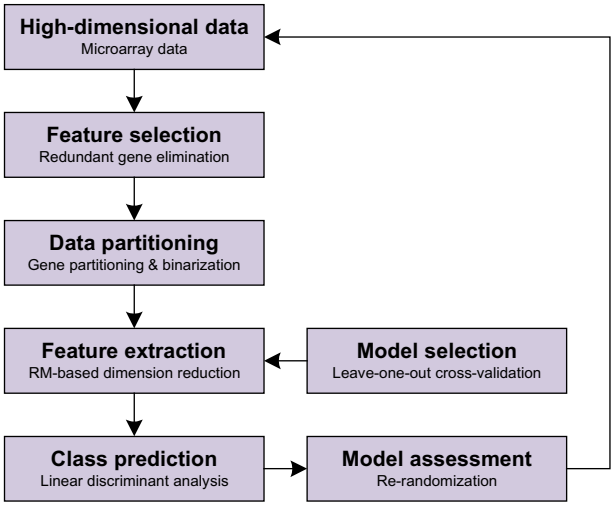

Fig. 1. The framework of dimension reduction.

The Prostate data set (Singh et al., 2002) contains $n = 102$ tissue samples with $p = 12600$ genes: 52 prostate tumor samples and 50 non-tumor prostate samples. Both data sets are from Affymetrix high-density oligonucleotide microarrays and are publicly available (Dettling, 2004).

For both data sets, the preprocessing steps are applied as follows (Dudoit, Fridlyand, & Speed, 2002): (a) thresholding, floor of 100 and ceiling of 16,000; (b) filtering, exclusion of genes with $\max / \min \leqslant 5$ and $(\max - \min) \leqslant 500$, where max and min refer to the maximum and minimum intensities for a particular gene across all samples; and (c) $\log_{10}$ transformation and standardization to zero mean and unit variance. The data were then summarized by a matrix $\mathbf{X} = (x_{ij})$, where x_{ij} denotes the expression level for gene j in sample i. The data for each sample consist of a gene expression profile $\mathbf{x}_i = (x_{1i}, \ldots, x_{pi})^T$ and a class label y_i. After data preprocessing the dimension of the matrix $\mathbf{X}$ was 72×3571 and 102×6033 for Leukemia and Prostate data set, respectively.

3.2. Feature selection

Although the procedure described here can handle a large number (thousands) of genes, the number of genes may still be too large for practical use. The model assessment procedure is very CPU-expensive and therefore time-consuming process, because it requires fitting the data many times due to cross-validation and re-randomization. Furthermore, a considerable percentage of the genes do not show differential expression across groups and only a subset of genes is of interest.

We used two different methods for feature selection in this study. First we performed an unsupervised random subset selection, consisting of p^* $(p^* < p)$ genes from the set of all genes, as described by Dai, Lieu, and Rocke (2006). We selected p^* genes with $p^* = \{50, 100, 200\}$ from both experimental data sets.

As the supervised alternative, we applied Welch's t-test (Jeffery et al., 2006) to embed the class information in feature selection process and thus improved prediction accuracy. Welch's t-test provides the measure of the statistical significance of changes in gene expression between classes. Welch's t-test defines the statistic t by

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}},$$

where $\bar{x}_g, s_g^2$, and n_g are the mean, sample variance, and sample size of the class g $(g = 1, 2)$ for each gene, respectively. Feature selection was carried out based on absolute value of the t-statistic and the top p^* genes with $p^* = \{50, 100, 200\}$ were used for further processing.

3.3. Feature extraction

Here we describe the core of feature extraction based on the RMs and then briefly outline the PCA method, which we used as a benchmark.

3.3.1. Rasch model

In this subsection we first give a short overview of the RM theory in its original form, and then present its application to gene expression data.

The RM is a simple latent factor model, primarily used for analyzing data from assessments to measure psychological constructs such as personality traits, abilities, and attitudes (Fischer & Molenaar, 1995). Assume that we have I persons and J items. Let y_{ij} be the response of person i to the item j, where the y_{ij} is '1' if person i answered item j correctly and '0' otherwise. In the RM, the probability of the outcome $y_{ij} = 1$ is given by

$$P(y_{ij} = 1 | \eta_i) = \frac{\exp(\beta_j + \eta_i)}{1 + \exp(\beta_j + \eta_i)}$$

for $i = 1, \ldots, I$ and $j = 1, \ldots, J$. η_i is the person parameter that denotes the latent factor of the ith person that is measured by the item j and β_j is the item parameter, which denotes the difficulty of the item j. Difficulty of the item describes the region of the latent trait distribution where the probability of producing a specific response changes from low to high. Probability of the response is monotonous in both person and item parameters. Fig. 2 plots the Rasch probabilities as a function of the value of the latent factor (η) for three different items. It can be seen that for a given item, persons with larger η value tend to have greater probability of expressing high scores on the latent factor, and for a given person, the response probabilities are different for items with different β values.

The final step of the RM-based modeling is to derive latent scores from the item responses. Latent score is the total score of person i over J items. The most common approach to calculate latent scores is to use the expectation of the posterior distribution of η_i given $\mathbf{y}_i$ with parameter estimates plugged in Rabe-Hesketh and Skrondal (2008). Details on the calculation can be found in Fischer and Molenaar (1995).

3.3.1.1. Application of Rasch model to gene expression data. In terminology of RM, we denote each gene as an "item" and each sample as a "person". The expression level x_{ij} of gene j in sample i is the response of a given sample to a given gene. Our main assumption is, that one-dimensional latent model may not hold for the complete set of p^* genes selected in the gene filtering step (see Section 3.2). Based on the assumption that expressional similarity implies functional similarity of the genes (and vice versa), we assume that genes with similar expression patterns determine one latent factor (Do & Choi, 2008).

To identify coexpressed genes we used k-means clustering (Gan, Ma, & Wu, 2007) to partition p^* genes into K partitions $(k = 1, \ldots, K)$ based on their gene expression profiles over n samples. K-means clustering is a simple and widely used partitioning algorithm. Its helpfulness in discovering groups of coexpressed genes has been demonstrated (Richards, Holmans, O'Donovan, Owen, & Jones, 2008). The optimal number of K is estimated following the procedure described in Section 3.5.

To apply the RM to the gene expression data, we need to discretize the gene expression data matrix $\mathbf{X}$ into binary form. We

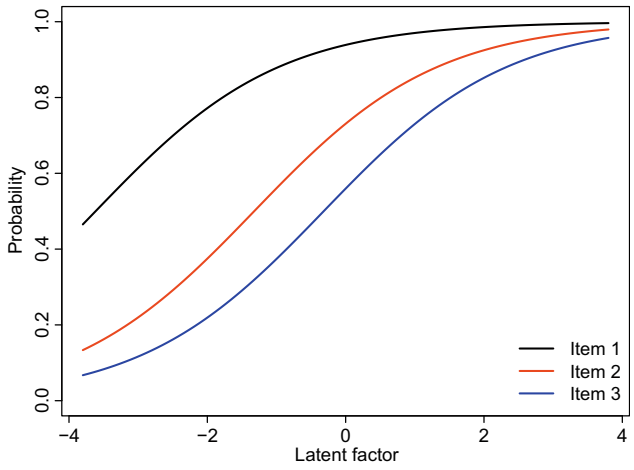

Fig. 2. Rasch probabilities as a function of the value of the latent factor for three different items. The person is likely to respond correctly to the Item 1 and unlikely to respond correctly to Item 3. Item parameters are $\beta_1 = -3.62, \beta_2 = -1.32$, and $\beta_3 = -0.32$, respectively.

use the median as a cut-off point for discretization. The intensity of every gene expression value is compared with the median gene expression data of the **X** and assigned a '1' if it is above and '0' otherwise. Note that gene partitioning is done before discretization step.

We fit a RM for genes in each of the K partitions respectively and calculate gene factor scores. Specifically, we construct K latent gene factors on which each gene in the k partition is located. The measure for each sample for each gene factor is then estimated. To fit the RM for genes in the kth partition, let i be the sample index, and j be the gene index, for $i = 1, \ldots, n$, and $j = 1, \ldots, p_k$, and let $\eta_i = \eta_{ik}$ be the latent gene factor for the ith sample which is determined by the genes in the kth partition, and β_j be the gene specific parameter for the jth gene. Class prediction is then carried out in the reduced space by using the gene factors.

3.3.2. Principal component analysis

PCA is the most commonly used technique for dimension reduction in microarray data analysis (Alter et al., 2000). The main idea behind the PCA is to reduce the dimensionality of a data set, while retaining as much as possible the variation in the original variables (Hastie, Tibshirani, & Friedman, 2001). This is achieved by transforming the p^* original variables $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_p]$ to a new set of K predictor variables $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_K]$, which are linear combinations of the original variables. More formally, PCA sequentially maximizes the variance of a linear combination of the original variables,

$$\mathbf{a}_K = \underset{\mathbf{a}^T\mathbf{a}=1}{\arg\max}\, \mathrm{Var}\ (\mathbf{Xa})$$

subject to the constraint $\mathbf{a}_i^T \mathbf{S}_\mathbf{X} \mathbf{a}_j$, for all $1 \leqslant i < j$, where $\mathbf{S}_\mathbf{X}$ is covariance matrix of the original data. The orthogonal constraint ensures that the linear combinations are uncorrelated. Linear combinations $\mathbf{t}_i = \mathbf{X}\mathbf{a}_i$ are known as the principal components. These linear combinations represent the selection of a new coordinate system obtained by rotating the original system. The new axes represent the directions with maximum variability and are ordered in terms of the amount of variation of the original data they account for. The first principal component accounts for as much of the variability in the original data as possible, and each succeeding component accounts for as much of the remaining variability as possible. Computations of the weighting vectors **a** involves the calculation of the eigenvalue decomposition of a data covariance matrix $\mathbf{S}_\mathbf{X}$,

$$\mathbf{S}_\mathbf{X}\mathbf{a}_i = \lambda_i \mathbf{a}_i,$$

where λ_i is the ith eigenvalue in the descending order for $i = 1, \ldots, K$ and $\mathbf{a}_i$ is the corresponding eigenvector. The eigenvalue λ_i measures the variance of the ith principal component and the eigenvector $\mathbf{a}_i$ provides the loadings for the linear transformation. The number of components K is specified on the basis of researcher's prior knowledge or determined using dedicated procedures (e.g., Kaiser–Guttman rule, Cattell's scree test, etc.). Class prediction using standard methods can then be carried out in the reduced space by using the constructed principal components.

3.4. Class prediction

After dimension reduction, the high dimension of p^* is now reduced to a lower dimension. The original data matrix is approximated by matrix of gene factors ($n \times K$, where $K < n$), constructed by RMs or PCA, as described in the previous section. To avoid confusion, we use the term "factor" to refer to both latent factor obtained from RM analysis and principal component derived from PCA. Once the K gene factors are constructed we consider prediction of the response classes.

To describe the class prediction problem formally, let we have a learning set $\mathscr{L}$ consisting of samples whose class is known and a test set $\mathscr{T}$ consisting of samples whose class has to be predicted. Denote the data matrix corresponding to $\mathscr{L}$ as the learning data matrix $\mathbf{X}_L$, and the data matrix corresponding to $\mathscr{T}$ as the test data matrix $\mathbf{X}_T$. The vector containing the classes of the samples from $\mathscr{L}$ is denoted as $\mathbf{y}_L$. The goal is to build a rule implementing the information from $\mathbf{X}_L$ and $\mathbf{y}_L$ in order to predict the class g of the ith sample from the test set given the gene expression profile $\mathbf{x}_{new,i}$:

$$\delta(., \mathbf{X}_L, \mathbf{y}_L) : \mathbb{R}^K \to \{1, \ldots, G\}$$
$$\mathbf{x}_{new,i} \mapsto \delta(\mathbf{x}_{new,i}, \mathbf{X}_L, \mathbf{y}_L).$$

Because our focus here is on dimension reduction, we fixed the class prediction step with LDA, although other methodologies can be used (Bellazzi & Zupan, 2008; Dudoit et al., 2002). A short description of the LDA method is given in the following (Boulesteix, 2004). Suppose we have K predictor variables. The random vector $\mathbf{x} = (X_1, \ldots, X_K)^T$ is assumed to a multivariate normal distribution within class $g = 1, \ldots, G$ (in our procedure $G = 2$) with mean $\boldsymbol{\mu}_g$ and covariance matrix $\boldsymbol{\Sigma}_g$. In LDA, $\boldsymbol{\Sigma}_g$ is assumed to be the same for all classes: for all $g, \boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$. Using estimates $\hat{\boldsymbol{\mu}}_g$ and $\widehat{\boldsymbol{\Sigma}}$ in place of $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}$, the discriminant rule assign the ith new observation $\mathbf{x}_{new,i}$ to the class

$$\delta(\mathbf{x}_{new,i}) = \arg\max_g (\mathbf{x}_{new,i} - \hat{\boldsymbol{\mu}}_g)\widehat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}_{new,i} - \hat{\boldsymbol{\mu}}_g)^T.$$

LDA has been well studied and widely used for class prediction problems. LDA relies on a hypothesis of multinormality, and assumes that the classes have the same covariance matrix. Although these hypotheses are rarely satisfied with real data sets, LDA generally gives good results. Studies have demonstrated favorable prediction performances of LDA models when compared with more complicated and computationally intensive algorithms such as neural networks and tree method (Dudoit et al., 2002; Tibshirani, Hastie, Narasimhan, & Chu, 2002). For the details of the calculation we refer the reader to Ripley (1996).

3.5. Model selection

The number of gene factors K is a meta-parameter in the procedure. We estimate K on the learning set $\mathscr{L}$ using leave-one-out cross-validation (LOOCV). LOOCV has been shown to give an almost unbiased estimator of the prediction error (Hastie et al., 2001), and therefore provide a sensible criterion for our purposes. In a nutshell, a subset of p^* genes is selected ($p^* < p$) from $\mathscr{L}$, and one of the samples is left-out. The feature extraction models are fitted to all but the left-out sample (see Section 3.3). The fitted models are then used to predict the class of the left-out sample (see Section 3.4). This is repeated for all samples in the learning data set n_L with K taking successively different values. The mean error rate (MER) over the n_L runs is computed for each value of K by

$$MER = \frac{1}{n_L}\sum_{i=1}^{n_L} I(\hat{y}_i \neq y_i),$$

where $\hat{y}_i$ is the predicted response class and y_i is the observed response class. I is the indicator function ($I(A) = 1$ if A is true, $I(A) = 0$ otherwise). The value of K minimizing the MER is selected and denoted as K^*. We select $K = \{1, 2, 3, 4, 5\}$ in our experiments.

3.6. Performance evaluation

To assess the performance of the RM- and PCA-based dimension reduction methods in the framework of class prediction we perform a re-randomization study. This evaluation approach was first used by Dudoit et al. (2002). The reader may refer to the review of Boulesteix, Strobl, Augustin, and Daumer (2008) on this subject. The procedure consists of the six steps as follows.

Step 1. For each data set, create $R = 100$ random partitions into learning data set $\mathscr{L}$ with n_L samples and a test set $\mathscr{T}$ with n_T samples ($n_L + n_T = n$). Denote $\mathbf{X}_L$ as the learning data matrix of size $n_L \times p$, and $\mathbf{X}_T$ as the test data matrix of size $n_T \times p$.

Step 2. Select a subset of p^* genes from the set of all genes p from matrix $\mathbf{X}_L$ using one of the gene selection methods, resulting in $\mathbf{X}_L^*$ matrix of size $n_L \times p^*$ and $\mathbf{X}_T^*$ matrix of size $n_T \times p^*$ (see Section 3.2).

Step 3. Use the learning data matrix $\mathbf{X}_L^*$ to determine the number of latent factors K^*, by LOOCV (see Section 3.5).

Step 4. Perform dimension reduction using RM- or PCA-based dimension reduction (see Section 3.3). Let $\mathbf{W}$ denote the matrix containing the factor loadings of size $p^* \times K^*$. Compute the matrix $\mathbf{Z}_L$ of gene factors for the learning data set ($\mathbf{Z}_L = \mathbf{X}_L^* \times \mathbf{W}$), and the matrix $\mathbf{Z}_T$ of gene factors for the test data set ($\mathbf{Z}_T = \mathbf{X}_T^* \times \mathbf{W}$).[1]

Step 5. Fit the class prediction model to the learning gene factors $\mathbf{Z}_L$. Predict the classes of samples in the test set using the fitted classifier and the test gene factors, $\mathbf{Z}_T$ (see Section 3.4).

Step 6. Repeat all above steps R times with re-randomization of the whole data set. The mean error rate (*MER*) for each method is given by

$$MER = \frac{1}{R}\sum_{r=1}^{R}\frac{1}{n_T}\sum_{i=1}^{n_T} I(\hat{y}_i \neq y_i),$$

where $\hat{y}_i$ is the predicted response class and y_i is the observed response class. I is the indicator function ($I(A) = 1$ if A is true, $I(A) = 0$ otherwise).

Although the *MER* is the most widely used metric for measuring the performance of the prediction systems, it considers all mispredictions as equally damaging (Boulesteix et al., 2008). It has been demonstrated (Huang & Ling, 2005) that, when the prior class probabilities are different, this measure is not appropriate because it does not consider misprediction cost, is strongly biased to favor the majority class, and is sensitive to class skewness. To overcome these problems the performance evaluation was also carried out via receiver operator characteristic (ROC) analysis. For comprehensive introduction to ROC analysis we refer the reader to Fawcett (2006). The ROC analysis was performed by plotting true positive rate (sensitivity) versus the false positive rate (1-specificity) at various threshold values, and the resulting curve was integrated to give an area under the curve (*AUC*) value. *AUC* is a measure of the discriminative power of the classes using the given features and classifier, and varies from $AUC = 0.5$ for non-distinguishable classes to $AUC = 1.0$ for perfectly distinguishable classes (Huang & Ling, 2005). The *AUC* can be interpreted as the probability that two random samples from the two classes will be predicted correctly, and is invariant to changes in class proportions (unlike *MER*). An $AUC \geqslant 0.7$ is generally considered acceptable, $AUC \geqslant 0.8$ as good, and $AUC \geqslant 0.9$ as excellent prediction performance.

3.7. Software

All computations were carried out in the R software environment for statistical computing and graphics (R Development Core Team, 2008). *K*-means clustering was performed using generic `kmeans` function. Binarization of continuous gene expression values was performed by `binarize` function in the `minet` package. The function `generate.split` of the `WilcoxCV` package was used to generate random splitting into learning and test data sets. RM analysis was performed using `ltm` package. PCA was conducted using generic `prcomp` function. LDA was carried out using `lda` function in the `MASS` package. ROC analysis was performed using the `caret` package. The procedures described here can be reproduced using the R scripts available from http://www2.arnes.si/akastr1/rasch/.

[1] Note that matrix $\mathbf{W}$ refers to component loadings, and matrix $\mathbf{Z}$ to gene components when PCA is performed.

4. Results

We illustrate the interest of RM-based dimension reduction by considering applications for the class prediction of microarray data. We compare the results from our procedure with the performance of the PCA-based approach. We will consider in turn the Leukemia and Prostate data sets, as described previously in Section 3.1.

4.1. Application to Leukemia data set

After data preprocessing, we applied the proposed performance evaluation procedure on the Leukemia data set. First, we consider $p^* = \{50, 100, 200\}$ randomly selected genes and used $R = 100$ random subsets. We randomly split each subset of genes into two data sets: a learning set with $n_L = 36$ samples and a test set with $n_T = 36$ samples. We used LOOCV procedure on the learning set to determine the number of gene factors (components in the case of PCA), and the test set for evaluating prediction performances. In total, 3600 class predictions were calculated using each of the dimension reduction method based on 100 randomization trials.

Table 1 gives the estimated mean error rates (*MER*), average values of the estimated meta-parameters (K^*), corresponding standard deviations, and areas under the ROC curves (*AUC*) with the considering methods, for different number of variables. A ROC curve analysis is depicted in Fig. 3(a). It can be seen from Table 1 that *MER* decreases with the increase of the size of gene subsets p^*. Inversely, the *AUC* increases when more predictor genes are included in model building. At any of the subsets size, the *MER*s of PCA-based class prediction are lower that of RM-based procedure. *AUC* scores suggest acceptable prediction performance for RM-based dimension reduction model and excellent performance for the PCA-based model.

Next, we present the results of performance evaluations using subsets of genes selected based on Welch's *t*-test. As described in Section 3.2, genes were ranked according to absolute value of the *t*-statistic and the top p^* genes with $p^* = \{50, 100, 200\}$ were used for extraction of gene factors. The analysis design was the same as for the randomly selected subset of genes described previously. The performance results on both methods are given in Table 2. Corresponding ROC curves are presented in Fig. 3(b). Comparing the results in Table 2 with the results given in Table 1, one can see that the accuracy of class prediction has been improved significantly.

Table 1
Prediction performances of the RM- and PCA-based prediction models using random gene selection on the Leukemia data set with 36/36 split of samples over 100 randomization trials.

p^*	RM–LDA			PCA–LDA		
	MER	K^*	*AUC*	*MER*	K^*	*AUC*
50	0.31 (0.10)	2.05 (1.23)	0.66	0.17 (0.09)	3.20 (1.29)	0.90
100	0.29 (0.10)	3.06 (1.42)	0.73	0.12 (0.07)	3.26 (1.14)	0.94
200	0.27 (0.10)	3.44 (1.44)	0.77	0.09 (0.06)	3.23 (1.17)	0.96

RM–LDA – RM-based class prediction; PCA–LDA – PCA-based class prediction; p^* – number of selected genes; *MER* – mean error rate; K^* – estimated number of gene factors (components); *AUC* – area under the ROC curve.

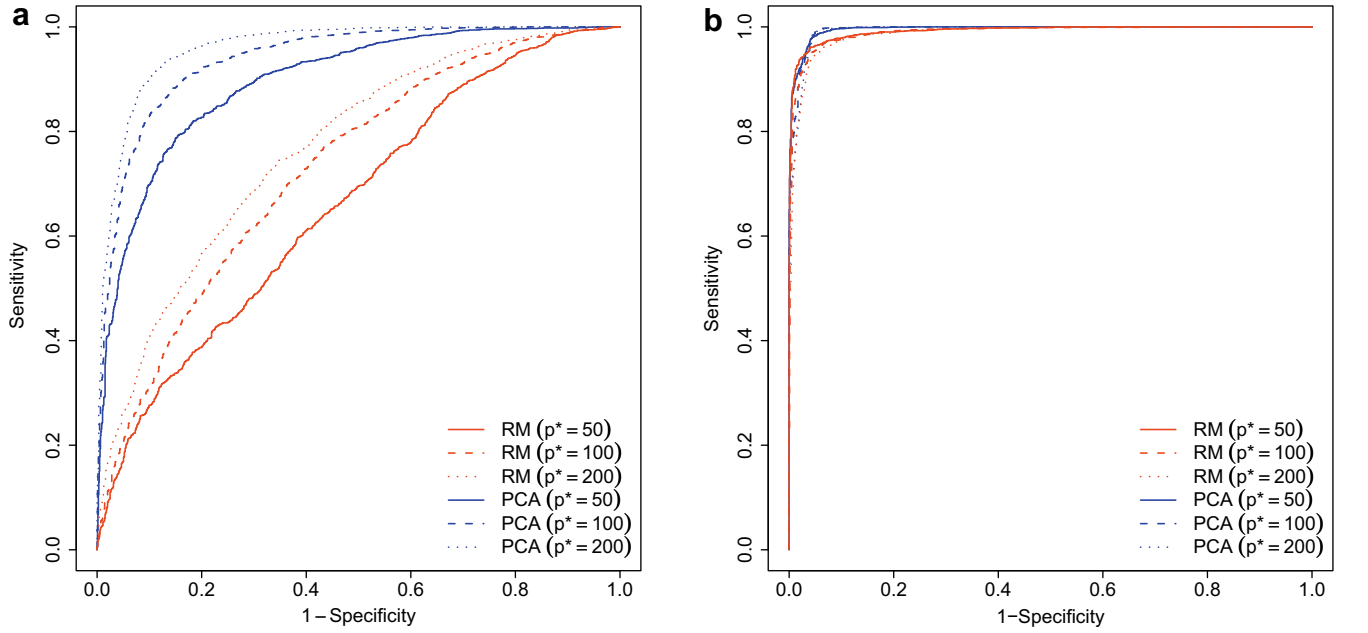

Fig. 3. Receiver operating characteristic curves (ROC) for RM- and PCA-based prediction models using random (a) and supervised (b) gene selection on the Leukemia data set with 36/36 split of samples over 100 randomization trials.

Table 2
Prediction performances of the RM- and PCA-based prediction models using supervised gene selection on the leukemia data set with 36/36 split of samples over 100 randomization trials.

p^*	RM–LDA			PCA–LDA		
	MER	K^*	*AUC*	*MER*	K^*	*AUC*
50	0.04 (0.03)	2.27 (0.65)	0.99	0.03 (0.02)	1.07 (0.33)	1.00
100	0.04 (0.03)	2.55 (0.83)	0.99	0.03 (0.02)	1.11 (0.63)	0.99
200	0.05 (0.04)	3.07 (0.96)	0.99	0.03 (0.02)	1.11 (0.40)	0.99

RM–LDA – RM-based class prediction; PCA–LDA – PCA-based class prediction; p^* – number of selected genes; *MER* – mean error rate; K^* – estimated number of gene factors (components); *AUC* – area under the ROC curve.

Table 3
Prediction performances of the RM- and PCA-based prediction models using random gene selection on the Prostate data set with 36/66 split of samples over 100 randomization trials.

p^*	RM–LDA			PCA–LDA		
	MER	K^*	*AUC*	*MER*	K^*	*AUC*
50	0.46 (0.08)	2.01 (1.19)	0.58	0.34 (0.10)	3.48 (1.46)	0.73
100	0.45 (0.08)	2.27 (1.29)	0.57	0.32 (0.10)	3.71 (1.23)	0.75
200	0.45 (0.08)	2.33 (1.26)	0.57	0.30 (0.11)	3.48 (1.37)	0.77

RM–LDA – RM-based class prediction; PCA–LDA – PCA-based class prediction; p^* – number of selected genes; *MER* – mean error rate; K^* – estimated number of gene factors (components); *AUC* – area under the ROC curve.

*MER*s of all two methods are reduced. Regarding *AUC* scores both models achieved excellent prediction performances. This is in agreement with the hypothesis that supervised gene selection should improve the classification accuracy. Moreover, the relative performance of the methods is basically the same: RM and PCA method have similar performances. The average value of meta-parameter (K^*) is lower when supervised gene selection is used.

4.2. Application to Prostate data set

The second data set used in this study is the Prostate tumor data. The experimental design was the same as for the Leukemia data set described previously (see Section 4.1). For both, random and supervised gene selection approaches, a learning set consists of $n_L = 36$ samples and a test set consists of $n_T = 66$ samples. In total 6600 class predictions were generated by each of the dimension reduction method based on 100 randomization trials.

Performance statistics of the RM- and PCA-based prediction models using random gene selection approach are summarized in Table 3. ROC analysis is visualized in Fig. 4(a). Although the pattern of performances between methods on the Prostate data set is similar to that on Leukemia data set, the results suggest that the classes are less well predicted. *AUC* scores indicate that RM-based prediction performances are close to random prediction, while PCA-based approach is generally acceptable.

Application of supervised gene selection using Welch's *t*-test yielded much better prediction performances (Table 4 and Fig. 4(b)). *MER*s and *AUC*s increase substantially. Moreover, the differences between both methods are minimal. Regarding ROC analysis both model performs excellent, although the *AUC* scores for the RM-based model are slightly lower.

5. Discussion

In this paper we explored the possibility of RMs to solve the course of dimensionality problem arising in the context of microarray gene expression data, and evaluated its performance in class prediction framework using LDA. To our knowledge, this is the first extensive validation study addressing RMs for microarray data analysis. In terms of using RM-based dimension reduction of microarray data, the evaluated approach appears to be as effective as widely used PCA-based dimension reduction.

Theoretically RMs can handle a large number of genes. However, as many other multivariate methods it is challenged by large computational time and danger of over-fitting. Therefore, we have used unsupervised random selection of small subset of genes and supervised Welch's *t*-test gene selection procedure. Applying random selection and using the *MER* and *AUC* scores as class prediction performances, overall average values of *MER* =

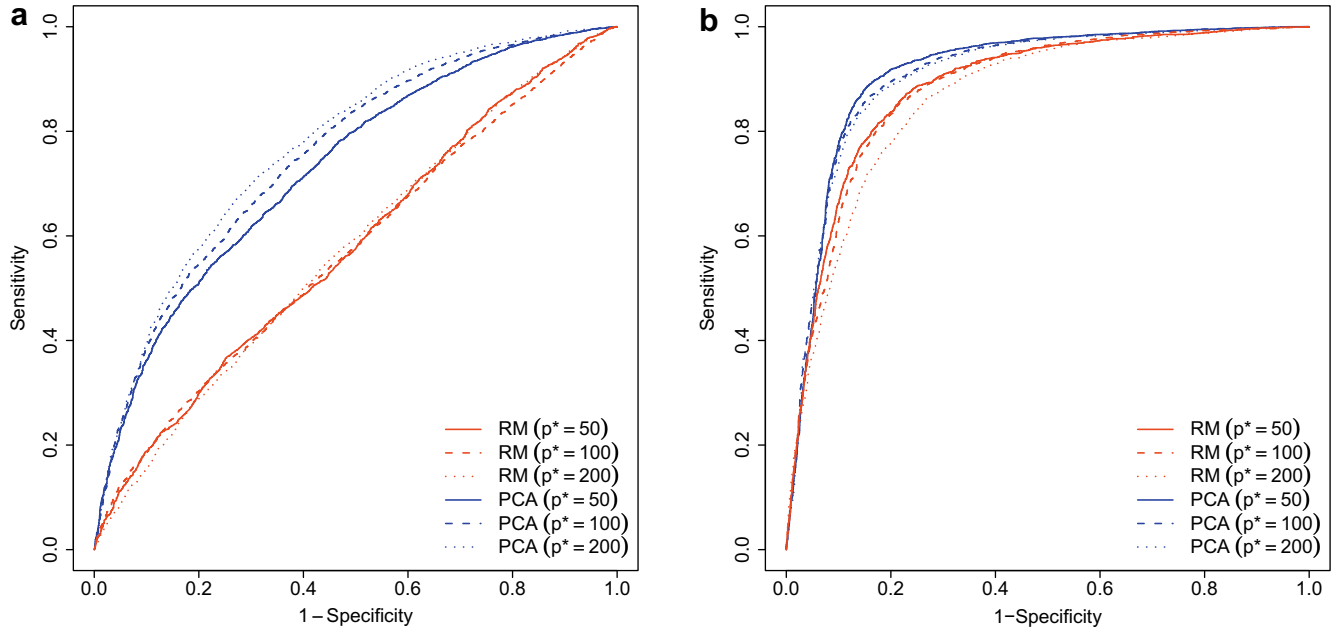

Fig. 4. Receiver operating characteristic curves (ROC) for RM- and PCA-based prediction models using random (a) and supervised (b) gene selection on the Prostate data set with 36/66 split of samples over 100 randomization trials.

Table 4
Prediction performances of the RM- and PCA-based prediction models using supervised gene selection on the Prostate data set with 36/66 split of samples over 100 randomization trials.

p^*	RM–LDA			PCA–LDA		
	MER	K^*	*AUC*	*MER*	K^*	*AUC*
50	0.18 (0.08)	2.91 (1.23)	0.88	0.14 (0.06)	2.06 (1.25)	0.91
100	0.19 (0.07)	3.35 (1.19)	0.88	0.15 (0.06)	2.41 (1.29)	0.91
200	0.21 (0.08)	3.55 (1.28)	0.86	0.15 (0.07)	2.72 (1.30)	0.91

RM–LDA – RM-based class prediction; PCA–LDA – PCA-based class prediction; p^* – number of selected genes; *MER* – mean error rate; K^* – estimated number of gene factors (components); *AUC* – area under the ROC curve.

0.29 ($AUC = 0.72$) and $MER = 0.45$ ($AUC = 0.57$) have been reached for Leukemia and Prostate data sets, respectively. We demonstrated that simple *t*-test improve the prediction performance significantly. Considering supervised gene selection procedure, overall average values of $MER = 0.04$ ($AUC = 0.99$) and $MER = 0.19$($AUC = 0.87$) have been reached for Leukemia and Prostate data sets, respectively. The patterns of performance measures between RM- and PCA-based procedures were similar, although the results suggested that RMs benefit more from preliminary gene selection. Compared to other studies aimed at class prediction, such as ones by Boulesteix (2004), Dai et al. (2006), or Nguyen and Rocke (2004), our performance values are comparable. Slightly better prediction performances in the case of Leukemia data set confirm the fact that the biological separation between the two classes is more pronounced in Leukemia data set (Antoniadis et al., 2003; De Smet et al., 2004).

We have developed our approach for discretized microarray data because RM scaling assumes a binary response of a gene expression level. Although our results indicate that the loss of information due to discretization step in our procedure is minimal, it is still the issue, if it is reasonable to consider gene expression discretely. Referring to the work of Sheng, Moreau, and De Moor (2003), who demonstrated the effectiveness of the Gibbs sampling to the biclustering of discretized microarray data, we argue that discretization may improve the robustness and generalizability of the prediction algorithm with regard to the high noise level in the microarray data. Following Hartemink (2001) the discretization of continuous gene expression levels is preferred for three reasons: (i) gene transcription occurs in one of a small number of states (low–high, off–low–high, low–medium–high, off–low–medium–high, etc.); (ii) the mechanisms of cellular regulatory networks can be reasonably approximated by primarily qualitative statements describing the relationships between states of genes; (iii) discretization, as a general rule, introduces a measure of robustness against error. It is a worthwhile future project to study the performance of our method on different levels of dicretization, using polytomous latent variable models (e.g., partial credit model) that could model more than two states of gene expression.

The major dilemma coupled with class prediction studies is the measurement of the performance of the classifier. The classifier cannot be evaluated accurately when sample size is very low. Moreover, feature selection and feature extraction steps should be an integral part of the classifier, and as such they must be a part of the evaluation procedure that is used to estimate the prediction performance (Asyali et al., 2006). Simon (2003) reported several studies published in high impact factor journals where this issue is overlooked, and biased prediction performances are reported. To address these issues we applied LOOCV scheme to estimate the appropriate number of gene factors and re-randomization experimental design to stabilize performance measures. It seems that LOOCV is appropriate, but a more sophisticated design (e.g., subsampling, bootstrap sampling, 0.632 estimator, etc.) to determine the number of gene factors could improve the prediction performance of the RM-based approach.

The vast amounts of gene expression data generated in the last decade have significantly reshaped statistical thinking and data analysis. The approach presented here can be also applied to many other problems in computational biology (e.g., single nucleotide polymorphism (SNP) array data analysis) and could be generalized to the other fields of sciences and humanities (e.g., health studies, risk management, financial engineering, etc.). Hence, innovative statistical methods like the one presented in this paper are of great relevance.

6. Conclusions

We have proposed a RM-based dimension reduction approach for the class prediction on microarray gene expression data. Our method is designed to address the course of dimensionality and overcome the problem of "large p, small n" so common in microarray data analysis. Experimental results showed that our procedure appears to be as effective as widely used PCA-based dimension reduction method. We demonstrated that binarization of continuous gene expression levels does not affect prediction performance of the classifier. We showed that appropriate gene selection is crucial before dimension reduction is performed. We restricted our approach to the binary prediction problem, but the methodology can be extended to cover multiclass prediction. The application of our method to other prediction problems (e.g., regression, survival analysis) is straightforward. We are currently working on extending this work on other latent variable models (e.g., partial credit model, Rasch–Andrich model, etc.).

Acknowledgement

The first author was supported by Junior Research Fellowship granted by Slovenian Research Agency.

References

Alter, O., Brown, P. O., & Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences of the United States of America, 97*(18), 10101–10106.

Antoniadis, A., Lambert-Lacroix, S., & Leblanc, F. (2003). Effective dimension reduction methods for tumor classification using gene expression data. *Bioinformatics, 19*(5), 563–570.

Asyali, M. H., Colak, D., Demirkaya, O., & Inan, M. S. (2006). Gene expression profile classification: A review. *Current Bioinformatics, 1*(1), 55–73.

Bellazzi, R., & Zupan, B. (2008). Predictive data mining in clinical medicine: Current issues and guidelines. *International Journal of Medical Informatics, 77*(2), 81–97.

Boulesteix, A.-L. (2004). PLS dimension reduction for classification with microarray data. *Statistical Application in Genetics and Molecular Biology, 3*(1). Retrieved from doi:10.2202/1544-6115.1075.

Boulesteix, A.-L., Strobl, C., Augustin, T., & Daumer, M. (2008). Evaluating microarray-based classifiers: An overview. *Cancer Informatics, 6*, 77–97. Retrieved from <http://www.la-press.com/evaluating-microarray-based-classifiers-an-overview-a577>.

Boulesteix, A.-L., & Strimmer, K. (2007). Partial least squares: A versatile tool for the analysis of high-dimensional genomic data. *Brief Bioinformatics, 8*(1), 32–44.

Bura, E., & Pfeiffer, R. M. (2003). Graphical methods for class prediction using dimension reduction techniques on DNA microarray data. *Bioinformatics, 19*(10), 1252–1258.

Chen, J. J. (2007). Key aspects of analyzing microarray gene-expression data. *Pharmacogenomics, 8*(5), 473–482.

Chiaromonte, F., & Martinelli, J. (2002). Dimension reduction strategies for analyzing global gene expression data with a response. *Mathematical Biosciences, 176*(1), 123–144.

Cordero, F., Botta, M., & Calogero, R. A. (2007). Microarray data analysis and mining approaches. *Briefings in Functional Genomics and Proteomics, 6*(4), 265–281.

Dai, J. J., Lieu, L., & Rocke, D. (2006). Dimension reduction for classification with gene expression microarray data. *Statistical Application in Genetics and Molecular Biology, 5*(1). Retrieved from doi:10.2202/1544-6115.1147.

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer.

De Smet, F., Moreau, Y., Engelen, K., Timmerman, D., Vergote, I., & De Moor, B. (2004). Balancing false positives and false negatives for the detection of differential expression in malignancies. *British Journal of Cancer, 91*(6), 1160–1165.

Dettling, M. (2004). BagBoosting for tumor classification with gene expression data. *Bioinformatics, 20*(18), 3583–3593.

Do, J. H., & Choi, D. K. (2008). Clustering approaches to identifying gene expression patterns from DNA microarray data. *Molecules and Cells, 25*(2), 279–288.

Dudoit, S., Fridlyand, J., & Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association, 97*(457), 77–87.

Dupuy, A., & Simon, R. M. (2007). Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *Journal of the National Cancer Institute, 99*(2), 147–157.

Fan, J., & Li, R. (2006). Statistical challenges with high dimensionality: Feature selection in knowledge discovery. In M. Sanz-Solé, J. Soria, J. L. Varona, & J. Verdera (Eds.), *Proceedings of the international congress of mathematicians* (pp. 595–622). Madrid: European Mathematical Society Publishing House.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874.

Fischer, G. H., & Molenaar, I. W. (Eds.). (1995). *Rasch models: Foundations, recent developments, and applications*. Berlin: Springer.

Gan, G., Ma, C., & Wu, J. (2007). *Data clustering: Theory, algorithms, and applications*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., et al. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science, 286*(5439), 531–537.

Hartemink, A. J. (2001). *Principled computational methods for the validation and discovery of genetic regulatory networks*. Unpublished doctoral dissertation, Massachusetts Institute of Technology, Boston.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). *The elements of statistical learning*. New York, NY: Springer.

Holter, N. S., Mitra, M., Maritan, A., Cieplak, M., Banavar, J. R., & Fedoroff, N. V. (2000). Fundamental patterns underlying gene expression profiles: Simplicity from complexity. *Proceedings of the National Academy of Sciences of the United States of America, 97*(15), 8409–8414.

Huang, J., & Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering, 17*(3), 299–310.

Hughes, G. (1968). On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory, 14*(1), 55–63.

Jain, A. K., Duin, R. P. W., & Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(1), 4–37.

Jeffery, I. B., Higgins, D. G., & Culhane, A. C. (2006). Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics, 7*, 359. Retrieved from doi:10.1186/1471-2105-7-359.

Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., et al. (2008). Machine learning in bioinformatics. *Brief Bioinformatics, 7*(1), 86–112.

Li, H., & Hong, F. (2001). Cluster–Rasch models for microarray gene expression data. *Genome Biology, 2*(8). Retrieved from doi:10.1186/gb-2001-2-8-research0031.

Nguyen, D. V., & Rocke, D. M. (2002). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics, 18*(1), 39–50.

Nguyen, D. V., & Rocke, D. M. (2004). On partial least squares dimension reduction for microarray-based classification: A simulation study. *Computational Statistics and Data Analysis, 46*(3), 407–425.

Orlando, D. A., Lin, C. Y., Bernard, A., Wang, J. Y., Socolar, J. E., Iversen, E. S., et al. (2008). Global control of cell-cycle transcription by coupled CDK and network oscillators. *Nature, 453*(7197), 944–947.

Rabe-Hesketh, S., & Skrondal, A. (2008). Classical latent variable models for medical research. *Statistical Methods in Medical Research, 17*(1), 5–32.

Rasch, G. (1966). An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology, 19*(1), 49–57.

Raudys, Š. (2006). Measures of data and classifier complexity and the training sample size. In M. Basu, & T. K. Ho (Eds.), *Data complexity in pattern recognition* (pp. 59–68). London: Springer.

R Development Core Team (2008). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Available from <http://www.r-project.org>.

Richards, A. L., Holmans, P., O'Donovan, M. C., Owen, M. J., & Jones, L. (2008). A comparison of four clustering methods for brain expression microarray data. *BMC Bioinformatics, 9*, 490. Retrieved from doi:10.1186/1471-2105-9-490.

Ripley, B. D. (1996). *Pattern recognition and neural networks*. Cambridge: Cambridge University Press.

Sheng, Q., Moreau, Y., & De Moor, B. (2003). Biclustering microarray data by Gibbs sampling. *Bioinformatics, 19*(Suppl. 2), 196–205.

Simon, R. (2003). Diagnostic and prognostic prediction using gene expression profiles in high-dimensional microarray data. *British Journal of Cancer, 89*(9), 1599–1604.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., et al. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell, 1*(2), 203–209.

Slawski, M., Daumer, M., & Boulesteix, A.-L. (2008). CMA – A comprehensive bioconductor package for supervised classification with high dimensional data. *BMC Bioinformatics, 9*(1), 439. Retrieved from doi:10.1186/1471-2105-9-439.

Tibshirani, R., Hastie, T., Narasimhan, B., & Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America, 99*(10), 6567–6572.

Zhang, C., Fu, H., Jiang, Y., & Yu, T. (2007). High-dimensional pseudo-logistic regression and classification with applications to gene expression data. *Computational Statistics and Data Analysis, 52*(1), 452–470.