Henryk MACIEJEWSKI

PREDICTIVE MODELLING IN HIGH-DIMENSIONAL DATA: PRIOR DOMAIN KNOWLEDGE-BASED APPROACHES

Oficyna Wydawnicza Politechniki Wrocławskiej
Wrocław 2013

Publication of this book was partly supported by
the Polish National Science Centre (NCN), grant N516 510239

Reviewers
Witold Jacak
Ewaryst Rafajłowicz

Cover design
Marcin Zawadzki

OFICYNA WYDAWNICZA POLITECHNIKI WROCŁAWSKIEJ
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław
http://www.oficyna.pwr.wroc.pl; e-mail: oficwyd@pwr.wroc.pl
zamawianie.ksiazek@pwr.wroc.pl

ISBN 978-83-7493-794-8

Drukarnia Oficyny Wydawniczej Politechniki Wrocławskiej. Order No. 741/2013.

Contents

Preface

Analysis of high-dimensional data is becoming increasingly important in many areas of contemporary science and technology due to the proliferation of massive throughput experimental techniques. DNA microarrays used in genomics are one of most prominent examples of such techniques, however high-dimensional data arise also in areas ranging from proteomics, spectroscopy, flow cytometry, magnetic resonance imaging, and satellite imaging to social science surveys or text mining. Analysis of such data has posed severe challenges, and has driven development of new, dedicated methods and algorithms in statistics, bioinformatics and machine learning.

This book is devoted to the problem of building predictive models from high-dimensional data, focusing mainly on feature selection and classification based on high-throughput genomic data, such as results from gene expression studies. The key characteristic of such datasets is the small number of samples available, which leads to difficulties with feature selection and generalization error of classifiers. Given such data, purely data driven methods of feature selection are virtually unable to provide stable, unique subsets of features which account for the differences between the groups of samples compared. Consequently, building robust predictive models from such data is a challenging task.

As a remedy to this, we propose in this work to use prior domain knowledge in the process of feature selection and classification. In the context of genomic high-throughput data, such domain knowledge may provide *a priori* information abouts sets of features (genes) which are likely to be functionally related, and is available e.g. in gene ontology or signalling pathway databases. Classification of samples relies then on activation of genes sets (pathways) rather than activation of individual genes (features). In this work we provide a comprehensive study as to how association of features sets with the target should, and should not, be quantified to produce statistically sound, interpretable results which stay in line with the actual organization of the high-throughput experiment. We propose

the algorithm of samples classification based on activation of feature sets and numerically evaluate this approach in terms of stability and generalization error.

This work builds on, and summarizes my research in the field of bioinformatics conducted over the last years. There are many colleagues who contributed to this work by helping me to get involved in the challenging but fascinating area of bioinformatics. First I want to mention Dr. Michał Jank who introduced me to the challenges of real life high-throughput studies in genomics. The difficult questions he kept asking pertaining to the analysis of data from genomic assays greatly inspired my research interests in these areas. Our collaboration has brought a number of joint papers where many of the techniques discussed in this worked were put into practice. I am also grateful to Dr. Ida Franiak-Pietryga with whom I have worked on a number of high-throughput studies concerning treatment of leukemia. These demanding but very exciting projects have not only resulted in a number of joint research papers but have also given me real satisfaction from working in the interdisciplinary team pursuing new leukemia therapies. I would also like to thank Prof. Beata Sobieszczańska and Dr. Robert Śmigiel for our common research in the interdisciplinary environment which has resulted in joint publications.

Finally, I am deeply indebted to Prof. Witold Jacak, who hosts me at his Studiengang Bioinformatik (Faculty of Bioinformatics) at the University of Applied Sciences in Hagenberg, Upper Austria. Numerous discussions with him concerning not only analysis of high-throughput data but also various problems in bioinformatics, machine learning and statistics have been an invaluable source of inspiration, motivation and encouragement to pursue my research in these areas.

Chapter 1

Introduction – class prediction based on high dimensional genomic data

In this chapter we want to provide an overview of the research problems addressed in this monograph. We present challenges related to analysis of high-throughput data, focusing on feature selection and predictive modelling. Although we discuss this in the context of high-throughput genomic assays, many of the issues raised are generic in nature and intrinsically apply to high-dimensional data of any other origin.

In the second part of this chapter, we outline the novel achievements proposed in this work which aim to improve feature selection and classification in high-dimensional data. Finally, we present the organization of this book.

Recent advancement of high-throughput experimental technologies has brought unprecedented opportunities in many areas of contemporary science. Life sciences is a perfect example of such an area where high-throughput techniques have not only revolutionized research and allowed us to pose novel scientific questions, but have also triggered development of new, dedicated methods in statistics and machine learning. These methods, broadly categorized as bioinformatics, have become an indispensable tool for analysis and interpretation of results from high-throughput assays in genomics or proteomics.

DNA microarrays are perhaps the most prominent example of high-throughput techniques used in genomics. Microarrays allow us to simultaneously measure expression levels of thousands of genes in a biological sample. By expression level of a gene we mean the quantity of the *transcript* (mRNA) which can be later translated into the protein coded by the gene concerned. Capacity of high-density microarrays is high enough to measure not only expression levels of all the genes of an organism, but also tens of thousands of non-coding regions (for instance, a human genome microarray from Agilent returns roughly 60,000 measurements, which encompass expression of all the 30,000 human genes as well as thousands of non-coding RNA sequences). Although microarrays can be now regarded as

the robust high-throughput technology, yielding reproducible results (as shown in comprehensive comparative studies by the Microarray Quality Consortium, (MAQC Consortium, 2006; Patterson *et al.*, 2006)), new alternative approaches are emerging, such as the SAGE (serial analysis of gene expression), or RNA-Seq (next-generation sequencing of transcripts used for quantitative expression profiling). These new technologies offer more flexibility in experiment design as they do not rely on the probes for transcripts to be specified in advance, however they still have their own intrinsic limitations, e.g. due to limited precision for low abundance transcripts (Łabaj *et al.*, 2011). Therefore, it seems that state-of-the-art high-throughput expression profiling will employ both RNA-Seq and microarrays in a combined, complementary approach (Łabaj *et al.*, 2011).

Organization of a typical high-throughput experiment in genomics involves measuring gene (or protein) expression profiles over a group samples (e.g. patients), where the number of samples n may reach up to a few hundreds, however $n \sim 30-100$ is more common. In any case, high-throughput experiments result in n vectors of dimensionality d, with $n \ll d$. Relatively small number of samples is due not only to the cost of high-throughput assays (which continuously decreases), but also to the infeasibility of gathering large, representative biological samples. In some experiments the samples are also labelled with a quantitative or qualitative target variable. Qualitative targets may represent such known characteristics of samples as as the disease status (e.g. tumor vs control), response to a therapy, risk of recurrence of a cancer, sample phenotype, etc. Note that the common convention in bioinformatics is to present results of high-throughput assays as $d \times n$ datasets, with columns representing samples and rows – features; in statistics or machine learning a transposed representation is more common, with samples (observation) occupying rows, and features (variables) – columns of datasets.

Based on results of high-throughput studies, we typically formulate research questions related to:

- class discovery,
- class comparison,
- class prediction.

Class discovery is related to identifying previously unknown subgroups among samples, which may be related to e.g. cancer subtypes or disease taxonomies (Bittner *et al.*, 2000). Class discovery has been performed using various methods of clustering (Eisen *et al.*, 1998; Tamayo *et al.*, 1999; Sturn *et al.*, 2002; Qin, 2006; Joshi *et al.*, 2008). Alternatively, class discovery may consist in *row-wise* clustering of expression data which reveals groups of presumably related features (genes) based on similar expression profiles.

Class comparison is related to the identification of genes which are differentially expressed between the groups of samples defined by the target variable, or significantly associated with the target. This is one of the most important applications of high-throughput techniques in genomics and has been extensively employed in numerous studies which aim to identify marker genes characteristic of different types of cancer or leukemia, or markers for targeted therapies (Golub *et al.*, 1999; Bittner *et al.*, 2000; West *et al.*, 2001; van't Veer *et al.*, 2002; Singh *et al.*, 2002; Chiaretti *et al.*, 2004; Nelson, 2004; Bild *et al.*, 2005; Xu *et al.*, 2005; Marincevic *et al.*, 2010; Franiak-Pietryga *et al.*, 2012b; Szmit *et al.*, 2012). Class comparison is typically realized by simultaneously testing the null hypothesis of no association of individual genes with the target, with the p-values of the tests multiple-testing adjusted in order to minimize the false-positive rate (Dudoit *et al.*, 2002b, 2003; Efron, 2007, 2008). However, it should be noted that some authors perform class comparison using multivariate approaches, where important features differentiating the classes are selected according to predictive performance of some classifiers (e.g. SVM). An example of such technique is the recursive feature replacement, RFR, (Fujarewicz and Wiench, 2003; Fujarewicz *et al.*, 2003; Simek *et al.*, 2004), used by Jarząb *et al.* (2005), or recursive feature elimination, RFE, (Guyon *et al.*, 2002), used by Fujarewicz *et al.* (2007) in the thyroid cancer study.

Class prediction consists in assigning new samples to classes represented by the target variable, based on gene expression profiles, with the classification models constructed from results of the high-throughput study. Clearly, this task is related to the class comparison problem, however the emphasis is on selection of the most informative subsets of features for classification of samples, rather than identifying all the genes which account for the differences between the classes. Numerous feature selection methods and classification models have been employed in class prediction studies, see specific examples and comprehensive overviews by Dudoit *et al.* (2002a); Cho and Won (2003); Guyon and Elisseeff (2003); Geman *et al.* (2004); Li T.*et al.* (2004); Lai *et al.* (2006); Saeys *et al.* (2007); Statnikov *et al.* (2008); Basford *et al.* (2012). One of the most severe challenges in class prediction is related to *overfitting*, i.e. poor prediction performance for new samples. Another issue is related to the difficulty with fair estimation of the generalization error of predictive models, given small number of samples (Simon *et al.*, 2003; Simon, 2003; Markowetz and Spang, 2005). Despite these difficulties, several authors have reported building predictive models using expression profiles from high-dimensional data, e.g. Golub *et al.* (1999); West *et al.* (2001); van't Veer *et al.* (2002); Xu *et al.* (2005). A few of these models have been commercialized, e.g. the MammaPrint

assay is the first microarray-based medical test, approved by the US Food and Drug Administration (FDA), designed for individualization of treatment of breast cancer patients (this test is based on the 70-gene expression signature identified by van't Veer *et al.* (2002)).

Several authors have raised concerns regarding stability and reproducibility of markers identified in different class comparison or class prediction studies (Xu *et al.*, 2005; Nelson, 2004; Ein-Dor *et al.*, 2005). For instance, comparing three related studies of breast cancer reported by (i) van't Veer *et al.* (2002), (ii) Sørlie *et al.* (2001, 2003) and (iii) Ramaswamy *et al.* (2002), it can be observed that only up to roughly 5% of differentially expressed genes identified by each of the studies are shared by the other experiments. Moreover, if we slightly change the set of samples in training data, we generally obtain different feature sets with equally good predictive performance as the original set reported by van't Veer *et al.* (2002), (Ein-Dor *et al.*, 2005).

Other authors also observe that marker genes from different studies of the same disease seem to be to a large extent study-specific, e.g. Xu *et al.* (2005); Nelson (2004); Miklos and Maleszka (2004); Lossos *et al.* (2004) (experiments concerning prostate cancer, schizophrenia and lymphoma).

This effect could be, to some extent, attributed to methodological differences concerning analysis of high-dimensional data (Subramanian and Simon, 2010; Dupuy and Simon, 2007), however the main problem is related to the small sample size (Ein-Dor *et al.*, 2006). Low reproducibility of results seems to be an inherent problem of class comparison/feature selection in the $n \ll d$ cases.

To overcome these difficulties, Subramanian *et al.* (2005) proposed to employ prior domain knowledge in class comparison studies. The idea was to focus on differential expression of *a priori* defined gene sets, grouping functionally related genes, rather than on differential expression of individual genes. The gene sets are defined as members of signalling pathways (as given in e.g. the KEGG pathway database), or as groups of genes with the same Gene Ontology category (as given in the GO database). This approach is motivated by the fact that lists of differentially expressed genes from class comparison studies are often too long, difficult to interpret, and highly variable across different studies. Another motivation is related to the fact that the actual differences between classes are often attributed to small changes observed over a group of related genes (e.g. in a signalling pathways) rather than to a substantial change in a few unrelated genes (Subramanian *et al.*, 2005). Hence if we express results of class comparison in terms of differentially expressed gene sets (e.g. activated

pathways), then we expect to improve interpretability as well as reproducibility and stability of results. See (Ackermann and Strimmer, 2009) and (Wu *et al.*, 2009) for a comprehensive review of the proposed approaches to gene set analysis.

Having presented the context and motivation of this work, we now present the key novel elements in this monograph. The main objective is to improve class prediction in high-dimensional ($n \ll d$) data by employing *a priori* domain knowledge in the process of feature selection. The key achievements in this monograph are as follows:

- We provide theoretical results which show limitations of data-driven feature selection in high-dimensional ($n \ll d$) data. These results express the probability of selecting the relevant features as a function of the sample size and dimensionality of the data. The conclusion from this analysis is that low stability (i.e. poor reproducibility) of data-driven feature selection is inherent in $n \ll d$ data. This motivates using additional, domain knowledge in the process of feature selection.
- We propose the method of classification in high-dimensional data based on activation of *a priori* defined feature sets. The feature sets represent the available domain knowledge about possible relationships among features. The feature selection algorithm will identify the feature sets which are activated, i.e. significantly associated with the target. If such feature sets are found, then the measures of activation of the feature set in individual samples will be used for the purpose of classification of samples. We propose the formulae which quantify the level of activation of the gene set in individual samples. Classification is then done based on these signatures of gene set activation calculated for individual samples.
- We provide a comprehensive methodological analysis of the available methods of gene set analysis. The methods are used to quantify *activation* of *a priori* defined gene sets. We clarify the models of statistical experiment implied by different algorithms as well as the null hypotheses actually assumed. We also analyze the models in terms of compliance with the actual biological experiment which produced the data. Based on this, we identify the methods which produce statistically sound and biologically interpretable results. We also provide a comprehensive numerical analysis of different methods of gene set analysis in terms of Type I error and power as a function of correlation among features and the signal strength (signal-to-noise ratio). Results of this analysis have been partly published in *Briefings in Bioinformatics*.

- We also provide two additional specific results pertaining to the aforementioned methodological analysis. First, we provide an improved version of the important gene set analysis method, GSA, proposed by Efron and Tibshirani (2007), with the modification related to correction of the flaw in estimation of significance in the original methods. Secondly, we propose a new interpretation of results produced by the popular methods of gene set analysis which use genes as sampling units. We show that results of these methods cannot be interpreted as p-values (as claimed by authors of these methods), however a heuristic, meaningful interpretation of the results can be proposed instead.
- We provide a comparative analysis of efficacy of (i) features selected with data-driven univariate or multivariate methods and (ii) features selected using gene set analysis methods (i.e. based on domain knowledge). The analysis is done in terms of predictive performance of classifiers as well as stability of features. We define several measures of stability of feature selection. We also overview recent results in the learning theory which relate CV_{loo}-stability (cross-validation leave-one-out stability) measures with predictivity conditions of classifiers. The concept of CV_{loo}-stability motivated the measures of stability used in this analysis.

This monograph is organized as follows. In Chapter 2 we discuss data driven methods of feature selection used in high-dimensional data analysis, focusing on univariate, multivariate and regularization-based techniques. These methods provide the baseline results in the analysis of efficiency of the proposed methods of features selection based on *a priori* domain knowledge. Chapters 3 through 5 form the methodological core of this book. In Chapter 3 we develop the theoretical analysis of the effect of small sample size on stability and reproducibility of feature selection in high-dimensional data, which motivates the proposed prior domain knowledge-based approach. Chapter 4 is devoted to the methodological analysis of the different approaches to gene set analysis, which will be employed as tools for domain knowledge based feature selection. In Chapter 5 we propose a generic algorithm for sample classification in high-dimensional data using domain knowledge-based feature selection. We compare this with the standard approach where feature selection is done in purely data-driven way. In Chapter 6 we provide numerical evaluation of the proposed approach and compare it with the standard approach in terms of generalization error and stability of feature selection.

Finally, we owe the Reader some clarification regarding terminology and conventions used in this monograph. Although the intention of the book is to

tackle generic problems concerning predictive modelling in high-dimensional data, the work was motivated by, and largely realized in the context of analysis of high-throughput data in genomics or proteomics. This inevitably has some impact on the language used in this work. For clarity of presentation, we often want to stick to the naming conventions common in bioinformatics. And so, we use the term "gene" interchangeably with the term "feature"; "gene expression" is simply the signal measured for a particular feature; "selection of differentially expressed genes" means "selection of features associated with the target"; a "(signalling) pathway" is another name for a "gene set", i.e. an *a priori* defined feature set; when we refer to a pathway as "activated", we mean that the gene set composed of the pathway members is significantly associated with the target; "phenotype" is another name for the "target variable". We also note that throughout this work we use the "tall" representation of high-dimensional datasets, common in bioinformatics, with rows of the dataset representing features (expression of genes) and column – subjects (samples) tested in the high-throughput assay. Note that by convention statistics and machine learning use the transposed representation of datasets.

We want to note that although these conventions and terminology are characteristic of bioinformatics, the very methodology proposed in this work, i.e. the algorithms of feature selection and classification based on prior domain knowledge are generic. These methods could be used e.g. in text mining studies where the task is to categorize documents based on high-dimensional vectors of attributes (terms), providing that prior domain (e.g. expert) knowledge is gathered defining sets of terms characteristic of some categories (this could be done e.g. for categorization of medical documents from the MEDLINE database, Yang and Pedersen (1997)).

This monograph builds on my previous research in such areas as machine learning, statistical data analysis, data mining and bioinformatics, as my experience gathered in these fields proved invaluable to tackle the specific problem of domain knowledge-based analysis of high-dimensional data. My research in these areas was published in a number of IF-journal papers: Maciejewski (2013); Pawłowski *et al.* (2013a,b); Rogalińska *et al.* (2013); Walkowicz *et al.* (2013); Franiak-Pietryga *et al.* (2012a,b); Król *et al.* (2012); Ostrzeszewicz *et al.* (2012); Sobieszczańska *et al.* (2012); Szmit *et al.* (2012); Walkowicz *et al.* (2011); Szmit *et al.* (2010); Wieteska *et al.* (2009); Maciejewski and Caban (2008); Maciejewski *et al.* (2008); Anders *et al.* (2006); Berenguel *et al.* (2005a,b).

Other publications which further develop specific ideas presented in this monograph or put them into practice are: Maciejewski (2012, 2011a,b); Maciejewski and Twaróg (2009); Maciejewski (2008a,b, 2007); Maciejewski and Jasińska (2005); Maciejewski *et al* (2005).

Chapter 2

Feature selection for sample classification based on high throughput data

In this chapter we present generic approaches to feature selection developed in machine learning as well as methods developed in bioinformatics literature which are focused on analysis of high throughput data. All the methods discussed here select subsets of features in a purely data-driven way, i.e. they do not include domain knowledge on possible associations among features in the process of feature selection. We discuss limitations of data-driven methods when dealing with data from high-throughput studies. These limitations motivate development of prior domain knowledge-based methods of feature selection proposed in this work.

2.1. Introduction

We introduce the following notation. Let $X = (x_{ij})$, $i = 1, ..., d, j = 1, ..., n$ denote the matrix with results of a high throughput study (e.g., gene expression data). Rows of this matrix, denoted $X_{i\bullet}$, $i = 1, ..., d$ represent features, e.g. expression of d genes, while columns denoted $X_{\bullet j}$, $j = 1, ..., n$ represent the n samples tested. We also define $Y = (y_j)$, $j = 1, ..., n$ as the $(1 \times n)$ target vector with labels of the samples. The methods of feature selection will be presented here in the context of binary classification, however generalization of many of the methods to multiclass, regression or survival time problems is possible. Here we assume that the samples belong to one of the classes represented by $y_i \in \{c_1, c_2\}$, $i = 1, ..., n$, and we denote the indices of the samples in each of the classes as $C_1 = \{i : y_i = c_1\}$ and $C_2 = \{i : y_i = c_2\}$. The number of samples in each class is denoted $n_1 = |C_1|$, $n_2 = |C_2|$. It is convenient to represent the vectors $(x_{ij} : j \in C_1)$ and $(x_{ij} : j \in C_2)$ of expressions of gene i in classes c_1 and c_2, respectively, as $V_i^{(1)} = (v_{ik}^{(1)}), k = 1, 2, ..., n_1$ and $V_i^{(2)} = (v_{ik}^{(2)}), k = 1, 2, ..., n_2$.

Feature selection is primarily done to reduce dimensionality of the feature space and thus to reduce the risk of overfitting of classification models. Overfitting is the major difficulty to overcome when building classifiers from data with large

number of features d and relatively small number of samples n. To illustrate this problem, let us consider randomly generated training data $X_{d\times n}$ with n samples and d features, with the binary class labels $y_i \in \{-1, 1\}$, $i = 1, ..., n$, randomly assigned to the samples. Let us consider fitting the simplest linear classifier (linear decision function) to the training data, i.e. fitting the d-dimensional hyperplane:

$$f(x_1, ..., x_d) = \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d,$$

where $\text{sign}(f(\mathbf{x}))$ is the classification of $\mathbf{x} \in R^d$. We want to find the hyperplane separating the training data, i.e. such that $\text{sign}(f(\mathbf{x})) > 0$ for $\mathbf{x}$ in class "+1", and $\text{sign}(f(\mathbf{x})) < 0$ for $\mathbf{x}$ in class "−1". This is equivalent to finding the coefficients $\beta_0, \beta_1, \ldots, \beta_d$ such that:

$$\begin{cases} y_1(\beta_0 + \beta_1 x_{1,1} + \beta_2 x_{2,1} + \ldots + \beta_d x_{d,1}) > 0 \\ y_2(\beta_0 + \beta_1 x_{1,2} + \beta_2 x_{2,2} + \ldots + \beta_d x_{d,2}) > 0 \\ \ldots \\ y_n(\beta_0 + \beta_1 x_{1,n} + \beta_2 x_{2,n} + \ldots + \beta_d x_{d,n}) > 0 \end{cases} \tag{2.1}$$

It is a well known fact from linear algebra that if the vectors $X_{\bullet j}$, $j = 1, \ldots, n$ are not linearly dependent, then for $d + 1 \geq n$ there always exist coefficients $\beta_0, \beta_1, \ldots, \beta_d$ satisfying the set of inequalities (2.1). This means that if sufficiently many dimensions are available (i.e. if $d + 1 \geq n$) then we can always fit a linear decision function which separates the training data with zero training error, even if there is no relationship between the features and Y (as in this example). This model has *overfitted* the data and thus has no generalization property and, consequently, is expected to realize 50% prediction error for new independent data. Note that for massive throughput data we usually have $d \gg n$, hence in such studies the major challenge is to avoid overfitting of classification models.

Some training algorithms are less prone to overfitting, as they perform feature subset selection rather than building models based on all inputs provided. Examples include decision trees or methods that use some form of regularization, e.g. Support Vector Machines or ridge regression. However, empirical studies show that for high dimensionality data even these methods benefit from prior feature selection. For instance, Kohavi and John (1997) report a number of high dimensionality studies where decision tree (ID3 algorithm) with prior features selection step outperforms the the decision tree trained using all the features; similar results are reported by Guyon *et al.* (2002) in the context of Support Vector Machines.

It should be also noted that the methods of feature selection discussed here are also used in the task of *class comparison* (Golub *et al.*, 1999), which consists in identification of subsets of genes, or genetic "signatures", which account for the differences between the groups of samples compared. Feature selection methods are then used to filter out relatively few genes most associated with the target (such as the phenotype, disease state or response to therapies, etc.) out of the vast data from the high throughput study. The purpose of this is to obtain insight into the biological process of interest. For this reason, methods which reduce dimensionality by projecting data onto the directions corresponding to the first few principal components (such as PCA or PLS, see e.g. Basford *et al.* (2012)) have not gained wide-spread application in class comparison and class prediction studies based on e.g. gene expression datasets. The key drawbacks of these methods are (i) interpretability problems: it is difficult to interpret the components obtained in terms of genetic "signatures" which provide insight into the biological processes, and (ii) these methods do not allow us to discard any of the features measured in the massive throughput study in order to focus in further analysis on relatively small subset of the most relevant features. For these reasons, we omit these methods from further discussion in this chapter.

Some authors proposed to use computationally intensive approaches to select the most informative subsets of features from high-throughput datasets. For instance, Li L.*et al.* (2001) and Li L.*et al.* (2004) used genetic algorithms coupled with the k-nearest neighbours classifier for feature selection from genomic or proteomic studies, respectively. Ooi and Tan (2003) and Peng *et al.* (2003) used genetic algorithms, coupled with the SVM classifier. Robbins *et al.* (2007) applied the heuristic, nature-inspired method (more specifically, ant colony) to the problem of features selection. Dramiński *et al.* (2008) proposed the Monte Carlo feature selection (MCFS) procedure, which ranks the features according to how frequently they tend to be selected by (many) decision trees fitted to randomly drawn subsets of observations represented in randomly selected subspaces of features. In (Dramiński *et al.*, 2010), the authors also showed how this procedure can be extended to discover interdependencies among the features identified as most important for classification. However, due to extreme computational burden of these methods given high-throughput genomic or proteomic datasets, as well as their intrinsic non-deterministic nature, we omit these methods from this work.

In the next part of this chapter, we present univariate and multivariate methods of feature selection. We will refer to these methods as the *standard* methods,

as they perform selection of features in the data-driven way, as opposed to the methods which use prior domain knowledge about relationships among features, which we propose in this work.

2.2. Univariate methods of feature selection

Univariate methods evaluate association of each feature with the target individually. The limitation of this is that these methods do not take possible relationships among features into account. Based on the calculated measure of association, the features are ranked from the most to the least associated with the target. Using the ranking list we can then select the set of most informative features by training a classifier based on the top k features ($k = 1, 2, ..., d$) and selecting the value of k which minimizes the expected prediction error of the classifier for the new data. This procedure is referred to as the *filter* approach, as opposed to the *wrapper* approaches to feature selection (Kohavi and John, 1997), discussed later.

Association of the i-th feature $X_{i\bullet}$ and the target Y can be expressed using different heuristic or statistical measures. Here we present some measures commonly used in bioinformatics or machine learning literature, (Sobczak and Malina, 1985; Lai *et al.*, 2006; Dudoit *et al.*, 2002a; Cho and Won, 2003). Golub *et al.* (1999) proposed the signal-to-noise measure which became popular in gene expression studies, and is defined as (see notation introduced on page 15):

$$SNR = \frac{\left|\bar{V}_i^{(1)} - \bar{V}_i^{(2)}\right|}{\mathrm{std}(V_i^{(1)}) + \mathrm{std}(V_i^{(2)})} \tag{2.2}$$

Other measures include the t-statistic and the Wilcoxon statistic (Dudoit *et al.*, 2002a), where the former is based on the assumption that $V_i^{(1)}$ and $V_i^{(2)}$ are normally distributed, and the latter is nonparametric. For more that two classes compared, the F-statistic or the Kruskall–Wallis statistics can be used, where the former is based on the normality assumption and the latter is nonparametric. For quantitative targets, Pearson or Spearman correlation coefficients are typically employed (Cho and Won, 2003). Sobczak and Malina (1985) propose to use the Sebestyen criterion for evaluation of individual features or feature sets in terms of separability between classes.

Other measures can be also envisaged which are based on some separability criteria between the densities estimated from $V_i^{(1)}$ and $V_i^{(2)}$, such as divergence

(Theodoridis and Koutroumbas, 2006), or mutual information between each feature and the target (e.g., Cho and Won (2003); Guyon and Elisseeff (2003)).

Remark I. One of appealing properties of the univariate methods is that, in addition to being used for feature selection in the filter procedures described previously, they can be employed for *class comparison*, i.e. for identification of differentially expressed genes (i.e. the features significantly associated with the target). To do this, one need to estimate *significance* of the calculated measure of association. Significance for a gene (feature) i is typically calculated as the p-value of the statistical test which assumes the null hypothesis that $X_{i\bullet}$ and Y were drawn from independent random variables (which for binary target is equivalent to testing the hypothesis that $V_i^{(1)}$ and $V_i^{(2)}$ were drawn from the same distribution). For heuristic measures of association (such as the SNR), for which the distribution of the test statistic under the null hypothesis is unknown, significance is typically estimated using the sample permutation test. This consists in estimating the distribution of the test statistic under the null hypothesis by repeatedly calculating the test statistic under (many) permutations of the class labels (or values in the vector Y). The p-value can be then calculated as the fraction of permutations for which the test statistic exceeds the original observed test statistic (such as the SNR):

$$p = \frac{1}{B}\sum_{i=1}^{B} I(t_i > t) \tag{2.3}$$

where t is the observed value of the test statistic, t_i is the value of the test statistic calculated for the i-th permutation, B is the number of permutations, and I is the indicator function (i.e. returns 1 if the condition is true, and 0 – otherwise).

It should be noted that this is doable only for targets which are exchangeable (e.g., for Y representing time-series data, no permutation test exist). Permutation tests are employed in some studies also for t- or F-statistics, if the data are not normally distributed.

Remark II. It should be noted that due to the very nature of statistical testing, the standard procedures which declare features as significantly associated with the target based on the p-value < 0.05 threshold, result in a considerable number of false positive findings. For instance, given high throughput data with d genes the expected number of false positive genes is $0.05 \times d$, which exceeds 1000

for the values of $d > 20,000$, commonly encountered in gene expression studies. For this reason, we need to apply *multiple testing correction* to control:

- the *family-wise error rate*, i.e. the probability that false positives appear in the list of rejected hypotheses (i.e. genes claimed differentially expressed), Hochberg (1988); Holm (1979), or
- the *false discovery rate*, i.e. the fraction of false positives in the list of rejected hypotheses, (Benjamini and Yekutieli, 2001; Benjamini and Hochberg, 1995).

The comprehensive overview of multiple testing correction in the context of high throughput data is available e.g. in Dudoit *et al.* (2003).

2.3. Multivariate methods of feature selection

As opposed to univariate methods which evaluate informativeness of individual features, multivariate methods evaluate association with the target of subsets of features. *Wrapper* approaches, proposed by Kohavi and John (1997), assess predictive performance of subsets of features using a given classification algorithm and attempt to find the subset maximizing predictive performance. Since the exhaustive search through all possible subsets is NP-hard, therefore wrapper methods perform a heuristic search through this space. Different search strategies have been proposed (e.g., greedy forward/backward selection, floating searches, genetic algorithms, simulated annealing) which, coupled with different classifiers, account for a diversity of wrapper multivariate methods. A comprehensive overview is given in Lai *et al.* (2006); Guyon and Elisseeff (2003); Kohavi and John (1997).

It should be noted that, in addition to filter and wrapper approaches, some authors also distinguish embedded methods, which perform feature selection at the stage of model fitting. Examples include algorithms for building decision trees or shrinkage/regularization-based regression.

Here we present the most popular of these methods. We later compare selected multivariate methods with the domain knowledge-based methods proposed in this work.

2.3.1. Recursive Feature Elimination – RFE

Recursive Feature Elimination was proposed by Guyon *et al.* (2002) and has since then gained much popularity in bioinformatics. It is an iterative multivariate method which ranks features from the weakest to the most informative. The

method starts with a linear classifier built using all the d features (the authors suggest to use either a linear discriminant classifier, or, preferably, a Support Vector Machine with a linear kernel). The general form of the fitted discriminant hyperplane is

$$D(\mathbf{x}) = \mathbf{w}\mathbf{x} + b$$

where $\mathbf{x} = [x_1, ..., x_d]$ is the vector of features and $\mathbf{w} = [w_1, ..., w_d]$ is the vector of weights. The features are then ranked using w_i^2 as the ranking criterion, and the feature with the smallest value of w_i^2 (regarded as the weakest feature) is removed from the model. In the next step the model is fitted again using the remaining $d-1$ features and the next weakest feature is removed from the model. The procedure is repeated until no features remain in the model. The features eliminated in consecutive steps form the list of features ranked from the weakest to the most informative. Alternatively, the sets of features used for model building in subsequent steps form a nested hierarchy of feature subsets $FS_1 \supset FS_2 \supset ... \supset FS_d$, where the informativeness of each of the subsets is determined based on the predictive performance of the classifier (e.g. SVM).

2.3.2. Pair-wise methods

Best pair selection

Best pair selection (Bo and Jonassen, 2002) is a filter approach which attempts to rank pairs of features (genes) based on their joint discriminatory power. For a given pair of genes, the method fits a linear class-decision boundary using the diagonal linear discriminant (DLD) method, and then projects all the samples on the DLD axis (which is perpendicular to the boundary). The gene-pair score is evaluated by taking the t-statistic calculated from the projected points. The method uses either a complete search where all pairs of genes are evaluated and the genes are ranked based on the gene-pair scores (without repetition), or, alternatively, a greedy approach where first individual genes are ranked, and next the best pair is formed using the top-ranking gene, then the next top-ranking gene etc.

Bo and Jonassen (2002) present results of empirical studies which suggest that pair-wise feature filtering methods may improve performance of classifiers built from massive throughput (microarray) data, as compared with the univariate filtering methods.

Top scoring pair

Top scoring pair (TSP) is another pair-wise method, proposed by Geman *et al.* (2004); Xu *et al.* (2005). The key motivation in developing this method was to ease the search for the marker genes for classification of samples, based on the integrated inter-study data from different microarray experiments. Typically, a microarray assay involves relatively small number of samples ($n \sim 10^2$, $n \ll d$ problem), hence it seems desirable to combine data from several independent massive throughput studies of the same phenomenon (e.g. cancer) to improve identification of marker genes, and consequently produce more stable markers. However, direct integration of data from different studies is difficult due to different microarray technologies and different transformation/normalization protocols typically employed by the separate studies. The TSP method attempts to overcome this by ranking all *pairs of genes* using a measure of discriminatory power which is invariant to monotonic transformations of expression data. Based on this ranking, a pair with the largest score (i.e. the largest measure of discrimination) is selected as the top scoring pair. Specifically, the TSP score for a pair of genes i, j, $1 \leq i$, $j \leq d$, $i \neq j$ is defined as:

$$\Delta_{ij} = |p_{ij}^{(1)} - p_{ij}^{(2)}|$$

where (see notation introduced in section 2.1)

$$p_{ij}^{(1)} = \frac{1}{n_1} \sum_{k=1}^{n_1} I(v_{ik}^{(1)} < v_{jk}^{(1)})$$
$$p_{ij}^{(2)} = \frac{1}{n_2} \sum_{k=1}^{n_2} I(v_{ik}^{(2)} < v_{jk}^{(2)})$$

Note that Δ_{ij} is high if, in terms of expression, gene i is consistently below gene j in class 1, and consistently above in class 2, or vice-versa. Since this simple score is based on relative rankings of expression of genes i and j, it is indeed invariant to monotonic transformations of data. Hence with this score we can realize feature selection based on larger datasets combining data from different experiments, although data from these experiments might not be mutually comparable.

2.3.3. Feature subset selection – greedy methods

Since the exhaustive search for the optimal subset of features (i.e. the one that realizes the best prediction performance given some classifier) is not feasible,

greedy methods have been proposed. Examples include *forward*, *backward* or *stepwise* selection of coefficients in regression models. Forward selection begins with the intercept term (β_0) and selects into the model a variable which best improves the model. The process continues until no variable is found which significantly improves the model. The current model (with k variables, and the corresponding vector of parameters estimates denoted $\beta^{(k)}$) and the next model (with $k+1$ variables and parameter estimates $\beta^{(k+1)}$) are compared in terms of the residual sums of squares

$$RSS(\beta) = \sum_{i=1}^{n} (y_i - f_\beta(x_i))^2$$

where f_β is the model fitted to data (x_1, y_1), ..., (x_n, y_n). The $k+1$-th parameter selected to extend the current model with k parameters is the one which maximizes $RSS(\beta^{(k)}) - RSS(\beta^{(k+1)})$. However, if the $k+1$-th parameter does not significantly improve the model, then the procedure stops and returns the final model with k variables. Significance is assessed based on the F-statistic (Hastie *et al.*, 2001):

$$F = \frac{RSS(\beta^{(k)}) - RSS(\beta^{(k+1)})}{RSS(\beta^{(k+1)})/(n-k-2)} \tag{2.4}$$

where the new model significantly improves the current model if the statistic exceeds the 0.95th quantile of the F distribution with $(1, n-k-2)$ degrees of freedom (which is equivalent to the p-value < 0.05 result of the statistical test assuming the null hypothesis that the new model does not improve the current model).

Backward feature selection (feature elimination) begins with the model containing all the variables and calculates the F statistic for each individual variable removed from the model. Then the variable with the smallest F is removed, providing the statistic is not significant. The procedure continues removing the weakest variable from the model until all variables remaining in the model produce significant F statistics.

Another modification of these procedures is the stepwise feature selection, which adds subsequent variables into the model, similarly to the forward selection. However, the variables already in the model are screened and the weakest variable with the insignificant contribution to the model (i.e. the one with the smallest and insignificant F) is removed from the model (as in the backward elimination).

2.3.4. Variable selection using regularization techniques – the lasso and the elastic net

The lasso (Tibshirani, 1996) and the elastic net (Zou and Hastie, 2005) are examples of variable selection methods based on regularization/penalization techniques. Regularization techniques were proposed to improve regression models fitted to high dimensional data and were also found useful in $n \ll d$ cases.

These methods were inspired by the ridge regression (Hastie *et al.*, 2001) which fits a linear model minimizing the residual sum of squares while imposing a bound on the size of the coefficients, i.e. (see notation introduced in section 2.1):

$$\beta^{ridge} = \arg\min_{\beta} \left(\sum_{j=1}^{n} (y_j - (\beta_0 + \beta_1 x_{1j} + ... + \beta_d x_{dj})) + \lambda \sum_{i=1}^{d} \beta_i^2 \right) \tag{2.5}$$

where $\beta = (\beta_0, \beta_1, ..., \beta_d)$ is the vector of model coefficients and λ is a fixed parameter which controls the effect of penalization. Equivalent formulation of this is to solve the optimization problem:

$$\begin{aligned} &\beta^{ridge} = \arg\min_{\beta} \left(\sum_{j=1}^{n} (y_j - (\beta_0 + \beta_1 x_{1j} + ... + \beta_d x_{dj})) \right) \\ &\text{subject to: } |\beta|^2 \leq s \end{aligned} \tag{2.6}$$

for some s (the parameter controlling the effect of penalization), where $|\beta|^2 = \sum_{i=1}^{d} \beta_i^2$ is the L^2 norm of the vector of parameters (note that β_0 is omitted in the penalization term).

Ridge regression is primarily used to improve prediction performance of the model as compared with the ordinary least squares regression (Hastie *et al.*, 2001), especially for high dimensionality data. However, ridge regression does not perform well as the feature selector, since it continuously shrinks all the coefficients, hence all the parameters are kept in the model. The lasso and the elastic net also impose a penalty of the model coefficients, however they produce a sparse model which makes them useful as feature selection methods.

The lasso

The lasso is fitted by solving a similar optimization task as given in Equation (2.6), however with the L_1 norm used instead of the L^2 norm in the penalization term:

$$\beta^{lasso} = \arg\min_{\beta} \left(\sum_{j=1}^{n} (y_j - (\beta_0 + \beta_1 x_{1j} + ... + \beta_d x_{dj})) \right)$$
$$\text{subject to: } |\beta|_1 \leq s \tag{2.7}$$

where $|\beta|_1 = \sum_{i=1}^{d} |\beta_i|$ is the L_1 norm of the vector of parameters.

As the lasso penalty is no longer *strictly* convex, while shrinking the coefficients, the lasso actually realizes variable selection (Tibshirani, 1996). However, feature selection done by the lasso has the following characteristics, which can be considered limitations of this method (Tibshirani, 1996; Zou and Hastie, 2005): (i) for $d > n$, the lasso selects at most n features; (ii) the lasso does not have a "group selection" property, i.e. from a group of highly correlated variables, the lasso selects only one variable and omits the other ones, however, it is not possible to determine which one will be selected.

These characteristics of the lasso may be undesirable in some applications, e.g. while analyzing microarray gene expression data, where we want to select a group of mutually highly correlated genes which form a signalling pathway, possibly with more then n elements.

The elastic net

The elastic net proposed by Zou and Hastie (2005) combines the penalties of the ridge regression and the lasso:

$$\beta^{elastic\,net} = \arg\min_{\beta} \left(\sum_{j=1}^{n} (y_j - (\beta_0 + \beta_1 x_{1j} + ... + \beta_d x_{dj})) \right)$$
$$\text{subject to: } a|\beta|_1 + (1-a)|\beta|^2 \leq s \tag{2.8}$$

which, for $a \in (0, 1)$, is both strictly convex (as the ridge regression) and singular at the vertexes (as the lasso). As shown by Zou and Hastie (2005), due to this form of penalty, the elastic net has the property of the lasso as the automatic feature selector. However, unlike the lasso, the elastic net has the "group selection" property, i.e. groups of strongly correlated features are either in or out of the model together. This property is related to the strictly convex penalty of the elastic net, as shown by Zou and Hastie (2005).

In Figure 2.1, contour plots of the penalties of the ridge regression, the lasso and the elastic net are illustrated for the two dimensional case (β_1, β_2), for $s = 1$ and different values of a in Equation (2.8). It can be clearly seen that the penalty of

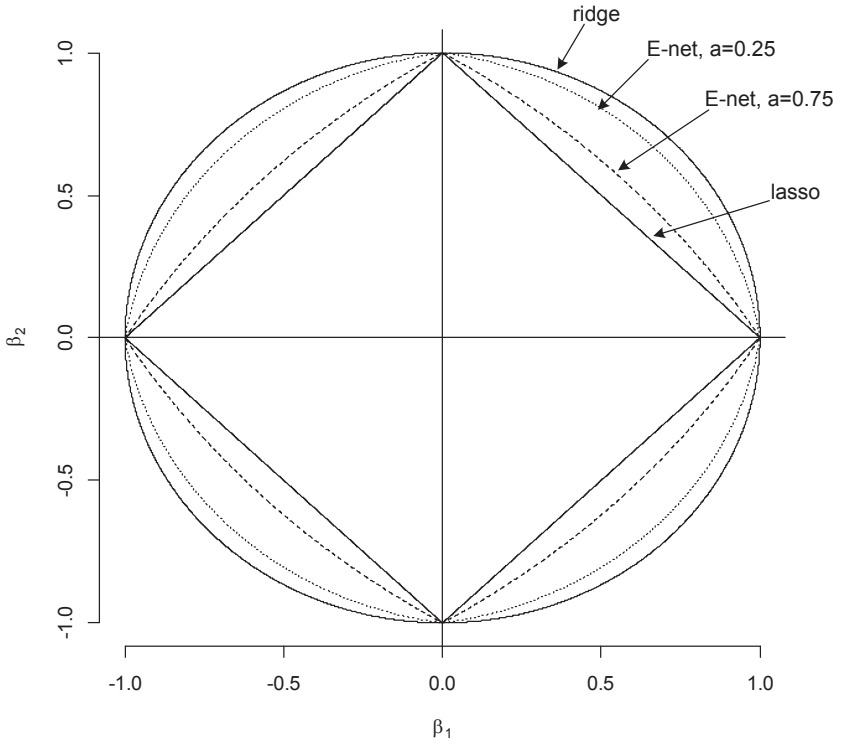

Fig. 2.1. Contour plots representing penalties in two dimensions (β_1, β_2) for the ridge regression, the lasso and the elastic net with different values of a (Equation (2.8))

the elastic net combines the characteristics of the ridge regression (strictly convex), and the lasso (singularities at the vertices), where changing the parameter a we bring the elastic net closer to the lasso (for $a \to 1$) or closer to the ridge regression (for $a \to 0$).

2.4. Discussion

Although numerous comparative studies have been reported regarding performance of different univariate and multivariate feature selection methods, conclusions are still not clear as to which of the methods should be preferred with high throughput data. For instance, Lai *et al.* (2006) compared several popular univariate (such as the t-test and SNR) and multivariate methods (such as greedy feature subset selection, RFE, TSP or best pair selection). The comparison in-

volved performance of different classifiers which used the features selected, and was based on a collection of data sets from real gene expression studies. Results reported by Lai *et al.* (2006) indicate that application of multivariate methods does not result in more informative features as compared with simpler univariate techniques. Also, none of the multivariate techniques clearly outperformed the other multivariate methods, although for some specific datasets performance of these methods differed significantly.

In the study by Bo and Jonassen (2002), performance of the best pair selection is compared with the standard univariate gene ranking (e.g. using the t-test) as well as with the multivariate forward search. Based on the empirical analysis involving two real life datasets from cancer-related studies, the authors conclude that the best pair method leads to improved performance over the standard methods.

Geman *et al.* (2004) empirically compared the top scoring pair method with the standard methods of feature selection used for the analysis of well known microarray studies (prostate cancer (Singh *et al.*, 2002), breast cancer (West *et al.*, 2001) and leukemia (Golub *et al.*, 1999)). This study indicates that the simple TSP method produces at least as good results as the standard methods of feature selection, however, the lists of marker genes, identified by the TSP are generally much shorter (and presumably more interpretable) as compared with the lists returned by the standard methods.

Other studies (Guyon and Elisseeff, 2003; Guyon *et al.*, 2002; Cho and Won, 2003; Dudoit *et al.*, 2002a,b) do not seem to bring convincing conclusions as to the preferable methods of selection of marker genes for classification of samples in high throughput studies.

Numerous high throughput assays have shown that application of standard univariate or multivariate methods of feature selection inevitably raises serious concerns related to stability and reproducibility of markers for sample classification (Xu *et al.*, 2005; Nelson, 2004; Ein-Dor *et al.*, 2005, 2006). It is commonly observed that the lists of most important features identified by different methods from the same data tend to show little overlapping, and, additionally, the lists returned by one particular method under small modifications of training data are generally not stable. Moreover, different studies of the same biological phenomenon (e.g. cancer) usually produce different, study-specific lists of marker genes.

In the next chapter, we show that these issues result from the $n \ll d$ data layout, i.e. from the small sample size as compared with the number of dimensions.

We analyze this problem from the theoretical standpoint and show that it is the $n \ll d$ data which inevitably harms any *data driven* method of feature selection in terms of stability/reproducibility of results. This leads to the conclusion that the solution to these weaknesses of feature selection cannot be achieved by finding the *right* method as long as we consider purely data driven approaches. The solution is rather to include additional domain-specific information on possible relationships among features in the process of feature selection for classification of samples. We elaborate on these domain knowledge-based methods in Chapter 4, and in Chapter 5 we employ them as tools for feature selection. Then in Chapter 6, we empirically compare the domain knowledge-based approach with the data driven methods presented in this chapter.

Finally, it is also noteworthy that purely data driven methods of feature selection, when applied e.g. to gene expression high throughput data, face another domain-specific limitation. It is commonly acknowledged that in many biological problems (e.g. diseases) studied with microarray techniques, the actual cause of the disease is related to relatively weak, but coordinated regulation in a group of functionally related genes, rather than to a strong change in some few unrelated genes (Subramanian *et al.*, 2005). If this is the case, then standard univariate or multivariate methods are very unlikely to be successful in discovering the real markers for sample classification, as virtually all of these methods will unavoidably start their search with the strongest features, some of which are likely to remain in the model. Data driven identification of subsets of weak features, regulated coordinately, would require that an exhaustive search through feature subsets is realized, which is in practice infeasible.

Chapter 3

Effect of small sample size – theoretical analysis

In this chapter we develop a theoretical model which allows to quantify instability of features selected from high dimensional data. The purpose of this is not only to explain the very nature of instability of features derived from high throughput data, but also to estimate the required sample sizes which guarantee generation of stable and relevant features.

In this chapter we use the following notation. Let $X = (x_{ij})$, $i = 1, ..., d$, $j = 1, .., n$ denote the matrix with results of a massive throughput study (e.g., gene expression data). Rows of this matrix, denoted X_i, $i = 1, ..., d$ represent features, e.g. expression of d genes, measured for n samples tested. We also define $Y = (y_i)$, $i = 1, ..., n$ as the $(1 \times n)$ target vector for the samples. Although Y can contain either qualitative or quantitative measurements, we assume here that Y is quantitative, i.e. $y_i \in R$, $i = 1, ..., n$.

We assume that the vectors X_1, ..., X_d and Y are samples of size n from the underlying random variables denoted $\mathcal{X}_1$, ..., $\mathcal{X}_d$, and $\mathcal{Y}$. Based on the data X we want to select features to be used as predictors. We consider a simple univariate feature selection procedure where features are selected based on their observed association with the target, i.e. the variables are ranked by the (absolute value of) association with the target, and the top k variables in the ranking list are then taken as predictors in regression or classification models. To explain instability of this feature selection procedure, we first assess probability that due to the limited sample size an irrelevant feature (i.e. not associated with the target) is selected instead of a relevant feature, and then we generalize this results to the case of selection of the top N "winning" features from large feature spaces.

3.1. Risk of selecting an irrelevant feature due to small sample size

We first focus on two variables $\mathcal{X}_i$ and $\mathcal{X}_j$. We assume that $\mathcal{X}_i$ and $\mathcal{Y}$ are independent and that $\mathcal{X}_j$ is associated with $\mathcal{Y}$. We denote by $\rho = \mathrm{cor}(\mathcal{X}_j, \mathcal{Y})$, the

actual correlation between the variables $\mathcal{X}_j$ and $\mathcal{Y}$, and without loss of generality we assume that $\rho > 0$. Obviously $\text{cor}(\mathcal{X}_i, \mathcal{Y}) = 0$. Given these two variables, we expect that the feature selection algorithm selects the variable $\mathcal{X}_j$ and omits the variable $\mathcal{X}_i$. Now we estimate the probability that this indeed happens, based on data X_i, X_j and Y.

Association of $\mathcal{X}_i$ and $\mathcal{Y}$ can be estimated from data based on the sample of size n (x_{i1}, y_1), ..., (x_{in}, y_n) as the sample correlation coefficient calculated as

$$r_i = \frac{\sum_{k=1}^{n}(x_{ik} - \bar{X}_i)(y_k - \bar{Y})}{\sqrt{\sum_{k=1}^{n}(x_{ik} - \bar{X}_i)^2}\sqrt{\sum_{k=1}^{n}(y_k - \bar{Y})^2}} \tag{3.1}$$

where $\bar{X}_i$, $\bar{Y}$ are means of X_i and Y. Similarly we calculate r_j as the sample correlation coefficient based on X_j and Y.

The probability that the feature selection algorithm selects the relevant feature j equals

$$p = \Pr(|r_j| > |r_i|) \tag{3.2}$$

To simplify analytical calculation of p we assume that the samples X_i, X_j and Y were drawn from normal distributions. Then using the Fisher transformation (Fisher, 1915, 1921) the transformed sample correlation

$$Z = \text{atanh}(r) = \frac{1}{2}\ln\frac{1+r}{1-r} \tag{3.3}$$

is approximately normally distributed, $Z \sim N(\mu, \sigma)$, with the parameters

$$\begin{aligned} \mu &= \frac{1}{2}\ln\frac{1+\rho}{1-\rho} \\ \sigma &= \frac{1}{\sqrt{n-3}} \end{aligned} \tag{3.4}$$

where n is the sample size and ρ is the true correlation between the random variables which generated the sample.

Since the Fisher transformation (Equation (3.3)) is an increasing function, the probability p (Equation (3.2)) equals

$$p = \Pr(|Z_j| > |Z_i|) \tag{3.5}$$

where $Z_i \sim N(\mu_i, \sigma_i)$, $Z_j \sim N(\mu_j, \sigma_j)$, with $\mu_i = 0$, $\mu_j = \frac{1}{2}\ln\frac{1+\rho}{1-\rho}$, and $\sigma_i = \sigma_j = \frac{1}{\sqrt{n-3}}$.

Then for independent (uncorrelated) features i, j the two dimensional random variable (Z_i, Z_j) has multivariate normal distribution $(Z_i, Z_j) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with the mean and covariance matrix

$$\boldsymbol{\mu} = \begin{bmatrix} 0 & \mu_j \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{2}\ln\frac{1+\rho}{1-\rho} \end{bmatrix} \tag{3.6}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_i^2 & 0 \\ 0 & {\sigma_j}^2 \end{bmatrix} = \begin{bmatrix} \frac{1}{n-3} & 0 \\ 0 & \frac{1}{n-3} \end{bmatrix} \tag{3.7}$$

The probability p (Equation (3.5)) can be now calculated by integrating the density of (Z_i, Z_j), denoted here $f(z_i, z_j)$, over $\{(z_i, z_j) \in R^2 : |z_j| > |z_i|\}$, i.e.

$$p = \iint\limits_{|z_j|>|z_i|} f(z_i, z_j)dz_i dz_j \tag{3.8}$$

Numerical calculation of the integral in Equation (3.8) is simpler after the change of variables:

$$\begin{bmatrix} u_i \\ u_j \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \times \begin{bmatrix} z_i \\ z_j \end{bmatrix} \tag{3.9}$$

which represents rotation counter-clockwise by 45°, as illustrated in Figure 3.1.

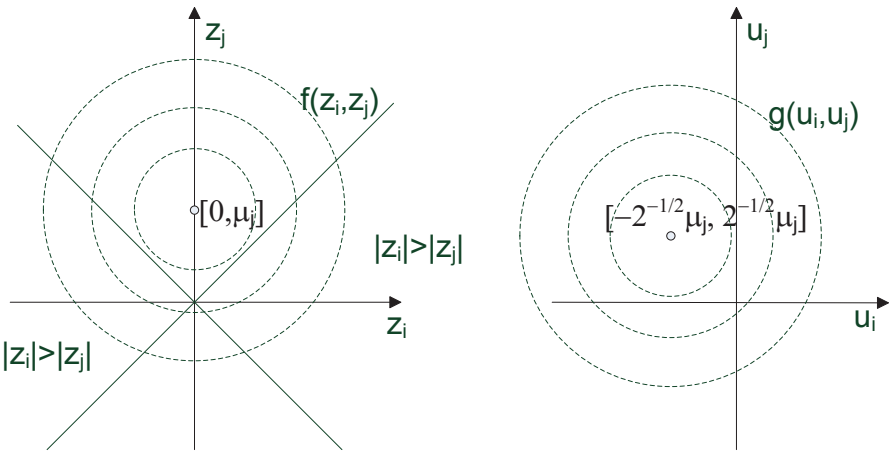

Fig. 3.1. Contour plot representing probability density $f(z_i, z_j)$ of the random variable (Z_i, Z_j) (left panel). After rotation (Equation (3.9)) $f(z_i, z_j)$ is transformed into $g(u_i, u_j)$ (right panel)

It can be observed that

$$\iint\limits_{|z_i|>|z_j|} f(z_i,\, z_j)dz_i dz_j = 2 \iint\limits_{u_i>0,\, u_j>0} g(u_i,\, u_j)du_i du_j$$

where the right-hand side integral can be readily calculated numerically.

This (considering Equation (3.8)) proves the following Theorem.

Theorem 1. *If X_i, X_j are samples of size n from normally distributed uncorrelated random variables $\mathcal{X}_i$ and $\mathcal{X}_j$, and Y is a sample of size n from normally distributed target variable $\mathcal{Y}$, where $cor(\mathcal{X}_j, \mathcal{Y}) = \rho$ and $\mathcal{X}_i$, $\mathcal{Y}$ are independent, then the probability p that the feature selection algorithm based on ranking features by absolute value of correlation with target will select feature $\mathcal{X}_j$ out of the pair $\mathcal{X}_i$, $\mathcal{X}_j$ equals*

$$p = 1 - 2 \iint\limits_{u_i>0,\, u_j>0} g(u_i,\, u_j)du_i du_j \tag{3.10}$$

where $g(u_i,\, u_j)$ is the density of multivariate normal distribution with the mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$:

$$\boldsymbol{\mu} = \left[-\tfrac{1}{2\sqrt{2}} \ln \tfrac{1+\rho}{1-\rho} \quad \tfrac{1}{2\sqrt{2}} \ln \tfrac{1+\rho}{1-\rho}\right] \tag{3.11}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \frac{1}{n-3} & 0 \\ 0 & \frac{1}{n-3} \end{bmatrix} \tag{3.12}$$

The probability p as a function of the sample size n and the correlation between the feature and the target (or the effect strength) is shown in Figures 3.2 and 3.3. These results clearly show that for the range of sample sizes often analyzed in microarray studies (e.g. n up to 50) there is a substantial risk that feature selection picks a variable that is completely unrelated with the target and omits the relevant feature. For instance, from Figure 3.2 we estimate that for the actual correlation $\rho = 0.3$ (moderate effect), this risk is about 30% (for the sample size $n = 20$), 15% (for $n = 50$), and 5% (for $n = 100$). Interestingly, this effect does not depend on the variance of the features $(X_i,\, X_j)$, but only on the sample size and the correlation of features with the target (effect strength). This phenomenon can be accounted for by the fact that limited (small) sample sizes inevitably lead to low accuaracy (due to high variance) in estimation of the actual relationship strength between features and the target, hence the features selected in replications of the experiment tend to be unstable.

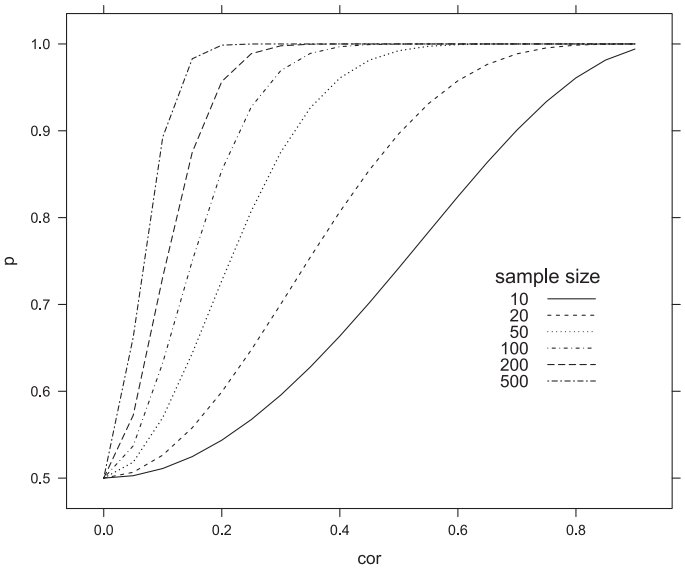

Fig. 3.2. Probability $p = Pr(|Z_i| < |Z_j|)$ of selecting the associated feature X_j and omitting the unassociated feature X_i as a function of the sample size and correlation between X_j and Y

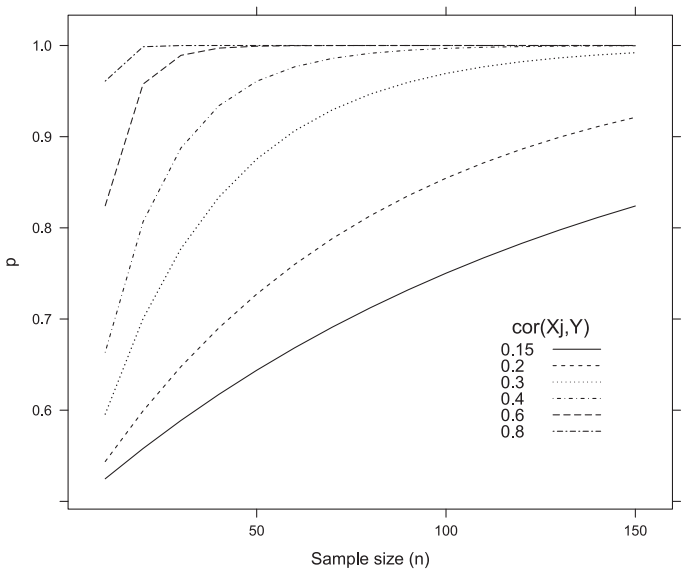

Fig. 3.3. Probability $p = Pr(|Z_i| < |Z_j|)$ of selecting the associated feature X_j and omitting the unassociated feature X_i as a function of the sample size and correlation between X_j and Y

This effect becomes even more challenging if the number of features searched by the feature selection algorithm increases. Note that in massive throughput studies the number of features searched commonly reaches $d = 10^3 - 10^4$ or more. We consider this case in the following section.

3.2. Feature selection from high dimensional data – effect of small sample size

We now consider the case where the target variable $\mathcal{Y}$ is associated only with a subset of variables $\mathcal{X}_i$, $i = 1, ..., d$, and we denote the set of indices of these variables S. We assume that these variables realize $\mathrm{cor}(\mathcal{X}_i, \mathcal{Y}) = \rho$ for $i \in S$, and without loss of generality we also assume that $\rho > 0$. The remaining variables are not associated with the target, i.e. $\mathrm{cor}(\mathcal{X}_i, \mathcal{Y}) = 0$ for $i \notin S$.

To simplify notation, we represent the rows of the data matrix X with indices in S by the matrix $V = (X_i)$, $i \in S$, whose rows V_1, V_2, ..., V_{n_V} represent samples from the $n_V = |S|$ variables associated with the target. Similarly, we represent samples from the variables not associated with the target by the matrix $W = (X_i)$, $i \notin S$, whose rows are denoted as W_1, W_2, ..., W_{n_W}, ($n_W = d - n_V$). In many applications (e.g. microarray gene expression studies) we often observe that $n_V \ll n_W$.

Now we analyze performance of a simple feature selection algorithm based on feature ranking. Given the data V_i, $i = 1, ..., n_V$, W_i, $i = 1, ..., n_W$ and Y the algorithm ranks the features by correlation with the target and selects N_{TOP} highest-ranked features. Here we consider the case where N_{TOP} is selected *a priori* and can be considered a parameter of the method. Obviously, we expect that the list of N_{TOP} selected features will be dominated by relevant features, i.e. features from the list V_i, $i = 1, ..., n_V$, with minimal share of the irrelevant features W_i, $i = 1, ..., n_W$. To quantify this we define the following measure of quality of feature selection based on high dimensional data:

$$p_L = \Pr(\text{in the list of } N_{TOP} \text{ features at least } L \text{ are relevant}) \tag{3.13}$$

We now calculate p_L analytically, which allows us to analyze how p_L changes as a function of the sample size n for a fixed L and N_{TOP}.

Let us denote the Z-transformed sample correlations with the target: $v_i = Z(\mathrm{cor}(V_i, Y))$ and $w_i = Z(\mathrm{cor}(W_i, Y))$, as defined by formula (3.3). In section 3.1, we derived the probability $p = \Pr(|v_i| > |w_j|)$, for a fixed i and j (Equation (3.10)). With high dimensional data the probability that the feature v_i will be

selected when compared with w_1, ..., w_{n_W} equals p^{n_W}, (as we assume that the features W_1, ..., W_{n_W} are independent), which quickly drops to 0 for the values of n_W encountered in massive throughput studies ($n_W \sim 10^3 - 10^4$).

We can obtain similar result by analyzing the order statistics $w_{(1)} < w_{(2)} < ... < w_{(n_W)}$. To simplify further analysis we focus of features positively associated with the target. The probability of selecting the relevant feature V_i when compared with W_1, ..., W_{n_W} equals then $\Pr(v_i > w_{(n_W)})$. As discussed in section 3.1, v_i is normally distributed with the parameters given by Equation (3.4). Distribution of the order statistic $w_{(n_W)} = \max(w_1, ..., w_{n_W})$ is also known and has the density

$$f_{w_{(n_W)}}(x) = n_W F^{n_W - 1}(x) f(x) \tag{3.14}$$

where F is the CDF and f is the density of w_j, which is also normally distributed with the mean = 0 and standard deviation given by Equation (3.4). In Figures 3.4 and 3.5 we graphically compare the density of v_i and $w_{(n_W)}$, for different values of n_W, which correspond to different dimensionality of feature space searched by the algorithm.

Note that the density of w_j is centered at 0 (not shown in Figures 3.4 and 3.5), however the density of $w_{(n_W)}$ shifts to the right as n_W increases. In Figure 3.4,

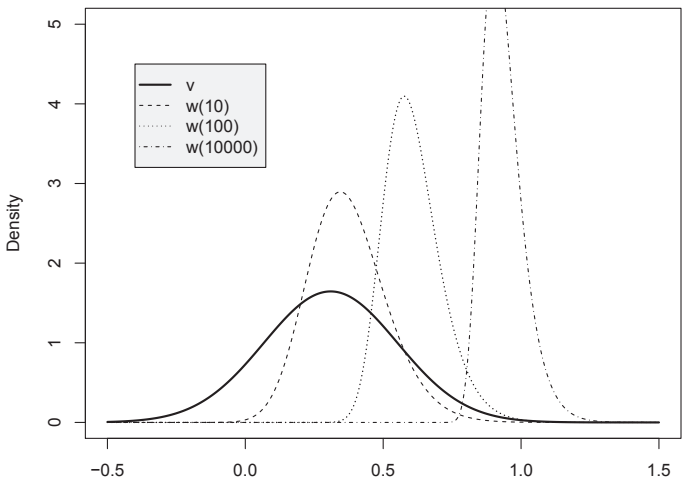

Fig. 3.4. Distribution of the observed correlation (v) of a relevant feature with the target compared with the maximum correlation of unrelated features ($w_1, ..., w_{n_W}$), for varying $n_W = 10, 100, 10000$. Plot created for the sample size $n = 20$ and correlation $\rho = 0.3$

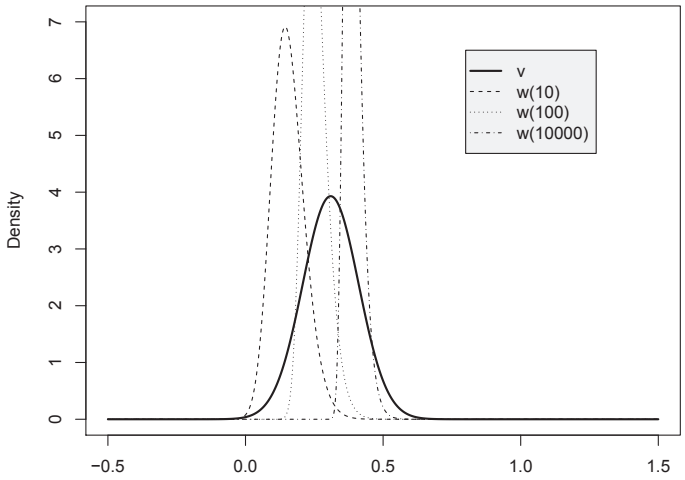

Fig. 3.5. Distribution of the observed correlation (v) of a relevant feature with the target compared with the maximum correlation of unrelated features $(w_1, ..., w_{n_W})$, for varying $n_W = 10, 100, 10000$. Plot created for the sample size $n = 100$ and correlation $\rho = 0.3$

we clearly see that for $n_W = 10$ the mean of $w_{(n_W)}$ exceeds the mean of v, hence the probability of selecting one of irrelevant features (w_1, ..., w_{n_W}) exceeds the probability of selecting the relevant feature v. For $n_W = 10000$ irrelevant features absolutely dominate. This effect diminishes with the growing sample size n, e.g. in Figure 3.5 we find that for $n_W = 10000$ the probability of selecting one of irrelevant features only slightly exceeds the probability of selecting the relevant feature.

Given the densities of v and $w_{(n_W)}$, we can quantify the probability

$$p_1 = \Pr(v > w_{(n_W)}) \tag{3.15}$$

that the relevant feature will be selected when ranked against the list of n_W irrelevant features. Since $\Pr(v > w_{(n_W)}) = \Pr(w_{(n_W)} + (-v) < 0)$ then the $p_1 = F_u(0)$ where F_u is the CDF of the random variable $u = w_{(n_W)} + (-v)$. Hence density of u can be obtained as the convolution of densities of $w_{(n_W)}$ and $(-v)$ (we denote these densities g and h, respectively):

$$f_u(t) = \int_{-\infty}^{\infty} g(t-x)h(x)dx \tag{3.16}$$

So the probability that a relevant feature demonstrates higher correlation with the target then n_W unrelated ('noisy') features is given by

$$p_1 = \int_{-\infty}^{0} f_u(x)dx \tag{3.17}$$

Both f_u and p_1 can be calculated numerically. Feasibility of this approach is illustrated in Table 3.1 where we calculated p_1 for the distributions v and $w_{(n_W)}$ depicted in Figures 3.4 and 3.5. Based on Table 3.1, we can quantify the effect of increasing the sample size n: e.g., increasing the number of samples from 20 to 100 raises the probability of selecting the relevant feature v when comparing it with $n_W = 10000$ irrelevant features from 6.587E−3 to 0.22.

Table 3.1. Probability $\Pr(v > w_{(n_W)})$ that the relevant feature will be selected when ranked against the list of irrelevant features as a function of the sample size n and the number of irrelevant features n_W

n	n_W		
	10	100	10000
20	0.41	0.13	6.587E−3
100	0.90	0.69	0.22

Now we will use similar reasoning to calculate the probability p_L as defined by formula (3.13). We consider selection of N_{TOP} features, given n_V actually relevant features (i.e. associated with the target) and n_W irrelevant features (not associated with the target). The value p_L describes the probability that in the list of N_{TOP} highest-ranked features at least L (for $L \leq n_V$) will come from the group of actually relevant features v_1, ..., v_{n_V}. We first note that if this is true then *at least* the following relevant features will be selected:

$$v_{(n_V-(L-1))},\ v_{(n_V-(L-2))},\ \ldots,\ v_{(n_V-1)},\ v_{(n_V)} \tag{3.18}$$

which is the last L order statistics from v_1, ..., v_{n_V}. The following (*at most* $N_{TOP} - L$) irrelevant features will also be selected:

$$w_{(n_W-(N_{TOP}-L-1))},\ w_{(n_W-(N_{TOP}-L-2))},\ \ldots,\ w_{(n_W-1)},\ w_{(n_W)} \tag{3.19}$$

which is the last $N_{TOP} - L$ order statistics from w_1, ..., w_{n_W}. Now we observe that these conditions are true if and only if

$$v_{(n_V-(L-1))} > w_{(n_W-(N_{TOP}-L))} \tag{3.20}$$

To prove this, observe that if inequality 3.20 is true, then before the $(N_{TOP} - L + 1) - th$ irrelevant feature is included in the list of 'winning' features, it is

guaranteed by (3.20) that at least L relevant features have already been selected. Truth of the reverse statement is obvious.

Hence we obtain the formula to calculate p_L:

$$p_L = \Pr(v_{(n_V-(L-1))} > w_{(n_W-(N_{TOP}-L))}) \tag{3.21}$$

This can be calculated numerically in a similar way as we used to calculate p_1 (defined by formula (3.15), and calculated by numerically integrating (3.17)). If we denote

$$\begin{aligned} k_V &= n_V - (L-1) \\ k_W &= n_W - (N_{TOP} - L) \end{aligned} \tag{3.22}$$

then the probability densities of the random variables $v_{(n_V-(L-1))}$ and $w_{(n_W-(N_{TOP}-L))}$ (which are the k_V-th and k_W-th order statistics, respectively) can be calculated as:

$$f_{w_{(k_W)}}(x) = n_W \binom{n_W-1}{k_W-1} F_w^{k_W-1}(x)(1-F_w(x))^{n_W-k_W} f_w(x) \tag{3.23}$$

$$f_{v_{(k_V)}}(x) = n_V \binom{n_V-1}{k_V-1} F_v^{k_V-1}(x)(1-F_v(x))^{n_V-k_V} f_v(x) \tag{3.24}$$

where F_w, F_v are the CDFs and f_w, f_v are the probability densities of w_i, $i = 1, ..., n_W$ and v_i, $i = 1, ..., n_V$, respectively ($w_i \sim N(0,\sigma)$, $v_i \sim N(\mu,\sigma)$, where μ, σ are given by Equation 3.4). Then we obtain the density of the r.v. $u = w_{(k_W)} + (-v_{(k_V)})$ as the convolution of the densities of $w_{(k_W)}$ and $(-v_{(k_V)})$ (where the density of $(-v_{(k_V)})$ is $f'_{v_{(k_V)}}(x) = f_{v_{(k_V)}}(-x)$):

$$f_u(t) = \int_{-\infty}^{\infty} f_{w_{(k_W)}}(t-x) f'_{v_{(k_V)}}(x)dx \tag{3.25}$$

Considering that $p_L = \Pr(v_{(k_V)} > w_{(k_W)}) = \Pr(u < 0)$, this proves the following Theorem.

Theorem 2. *If V_1, V_2, ..., V_{n_V} are samples of size n from n_V normally distributed* relevant *features (i.e. variables correlated with normally distributed target Y, with correlation equal ρ), and W_1, W_2, ..., W_{n_W} are samples of size n from n_W normally distributed features independent of target, then the probability that in the list of N_{TOP} features selected by the univariate algorithm based on ranking features, at least L features are relevant equals*

$$p_L = \int_{-\infty}^{0} f_u(x)dx \tag{3.26}$$

where f_u is calculated according to formula (3.25).

Note the the integral (3.26) and the convolution (3.25) can be calculated numerically.

This important result allows us to analyze effect of the sample size (n) on the number of noisy features returned by the feature selection procedure based on feature ranking. In Figures 3.6 through 3.8 we illustrate this relationship for some specific values of n_V and n_W. For instance, in Figure 3.6 we analyze selection of top 90 features from ca. 5000 features (e.g. transcripts) out of which only 100 are actually important (associated with the target). We observe that for relatively low correlation with the target ($\rho = 0.2$, bottom-left panel) at least 100 samples are required to guarantee that in the list of winning features about one third are relevant ($p_L \approx 1$ for $L = 30$). Obviously, with stronger effect

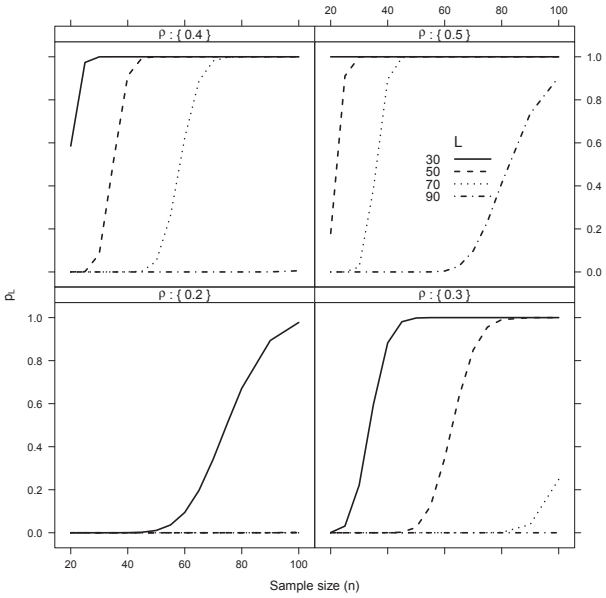

Fig. 3.6. Probability that at least L relevant features are selected in the list of top 90 features as a function of the sample size n and correlation ρ. Plot created for $n_V = 100$ and $n_W = 5000$

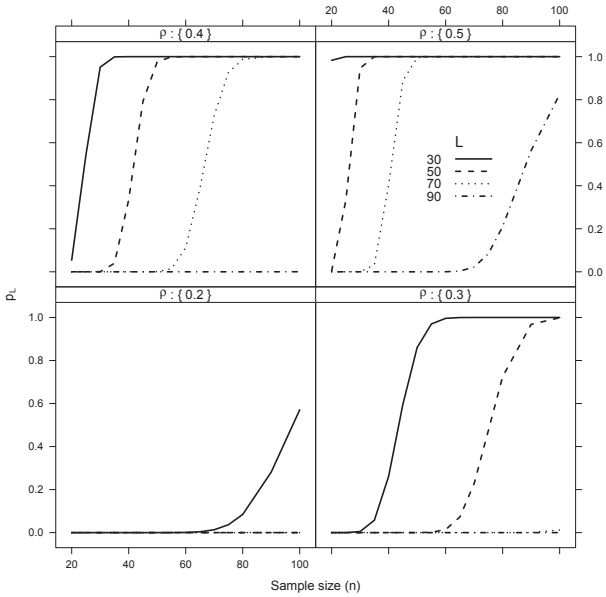

Fig. 3.7. Probability that at least L relevant features are selected in the list of top 90 features as a function of the sample size n and correlation ρ. Plot created for $n_V = 100$ and $n_W = 10000$

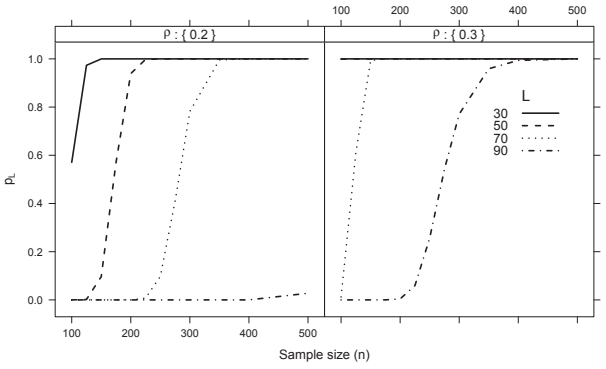

Fig. 3.8. Probability that at least L relevant features are selected in the list of top 90 features as a function of the sample size n and correlation ρ. Plot created for $n_V = 100$ and $n_W = 10000$

(i.e relationship between relevant features and the target), the number of required samples decreases (e.g. for $\rho = 0.3$ only about 40 samples guarantee one third of relevant features). For the strong effect ($\rho = 0.5$, top-right panel), the experiment design with $\sim 40 - 50$ samples guarantees that 70 out of 90 features selected are relevant.

In Figure 3.7 we analyze the effect of higher-dimensional data, with ca. 10 thousand features. We observe that now about 10-20 more samples are required to realize similar concentration of relevant features in the list of top winning features as in the previous example. For instance, comparing the bottom-right panels of Figure 3.6 and 3.7, we see that in order to guarantee 50 (out of 90) relevant features we need 80 samples for data of dimensionality of 5000 and about 100 samples if dimensionality is twice as much.

In Figure 3.8 we focus on the lower effect strength (ρ up to 0.3), as this range of correlation is commonly observed in microarray data. In this case hundreds of samples are needed to guarantee that majority of features selected are relevant, e.g., for $\rho = 0.3$ about 150 samples yield 70 relevant feature (in the list of 90), while for $\rho = 0.2$ the same requires 350 samples. To obtain all relevant features for $\rho = 0.3$ we need about 400 (the same would require about 1000 samples if $\rho = 0.2$, result not shown).

3.3. Discussion and conclusions

In this chapter, we analyzed simple feature ranking algorithms in terms of the probability of selection of the relevant features. We analytically estimated this probability as a function of the sample size and dimensionality of the feature space. This analysis shows that for the small sample size and high dimensional data ($n \ll d$ problems), irrelevant features can dominate the lists of features selected with univariate methods.

These results explain inherent problems with feature selection from $n \ll d$ data, irrespective of what method (uni- or multivariate) is used. The problems result from the small sample size, which leads to the high variance of the *observed* correlation of features with the target, and effectively hides the *actual* correlation of features with the target.

In order to simplify the analytical formulae derived in this chapter, we assumed that the features and normally distributed and independent, and we considered quantitative, normally distributed target. Although these results directly apply to regression (and logistic regression) problems, they are intended as the illustration of the general limitation of data-driven feature selection in high dimensional data.

These results can also be used at the stage of high-throughput experiment planning, to estimate the required sample size which guarantees the minimum expected concentration of relevant features in the results of class comparison. We showed that the minimum sample size which guarantees this grows with the

dimensionality of data. However, if the required sample size is not affordable in a high-throughput assay, then the solution is to use prior domain knowledge-based feature selection. In the next chapter we discuss gene set analysis methods which can be employed for prior biological knowledge-based feature selection.

Chapter 4

Prior domain knowledge-based methods of feature selection

In Chapter 3 and in section 2.1, we analyzed limitations of standard univariate and multivariate methods of feature selection. We concluded that for data from high throughput studies (i.e. for $n \ll d$ problems), purely data-driven methods are not able to guarantee stability and reproducibility of the selected feature sets. Moreover, data-driven methods are unlikely to identify sets of *weakly* activated features which, working as a group of related features often account for the real differences between samples analyzed in class comparison/class prediction studies (Subramanian *et al.*, 2005).

In order to overcome these limitations of data-driven methods, we need to incorporate in the process of feature selection additional information, which is not derived from the experiment data but is used in the analysis as *a priori* domain knowledge about possible functional relationships in the set of features. In the context of bioinformatics, on which we primarily focus in this work, such domain knowledge about relationships among features (genes) is available in signalling pathway or gene ontology databases, such as the KEGG (Kyoto Encyclopedia of Genes and Genomes), Biocarta or Gene Ontology, and is currently being actively developed. Based on these databases, we derive gene sets which include genes – members of signalling pathways, or genes which share common gene ontology terms (i.e. are involved in the same molecular function or biological process, or are related to the same cellular component), or share common chromosome location.

Numerous methods of gene set analysis which allow us to estimate activation of pathways, or otherwise related gene sets, have been proposed in bioinformatic literature. The methods are primarily employed to obtain insight into the nature of the underlying biological process or disease. In these applications, the main incentive is to improve interpretability of differential expression studies based on data from high throughput assays. In this work, we adopt a different perspective – we focus on gene set analysis methods as the means of including prior domain knowledge in the process of feature selection for sample classification. We expect that by including *a priori* knowledge, we will be able to identify subsets

of features undergoing small, coordinated changes, which underlie the real differences between samples, and which are likely to be missed by the standard univariate or multivariate approaches. Moreover, we expect to improve stability and reproducibility of feature selection.

In this chapter, we primarily focus of methodological aspects of feature (gene) set analysis. It is known that different methods of gene set analysis tend to produce very disparate results, however, differences between the methods were studies mainly empirically. This motivated us to study this problem on the theoretical basis.

In the next section, we briefly review the current research pertaining to the competing approaches to gene set analysis. In section 4.2, we identify four groups of gene set analysis methods, which are based on fundamentally different methodological assumptions. In section 4.3, we discuss models of the statistical experiment that the different groups of methods actually imply, and we show which of the models comply with the actual biological experiment which provided the data. In section 4.4, we empirically compare the groups of methods in terms of the power and type I error (false positive rate). This comparison is based on data with known characteristics regarding signal to noise level and correlation among features. Finally, we analyze the relationship between power of the methods and correlation among features, in order to explain discrepancies between different empirical studies (section 4.6).

Finally, we want to note that application of the methods discussed in this chapter is by no means restricted to bioinformatics or genomics. The algorithms can be used generically to study relationships between *a priori* defined subsets of numeric feature and the target. For instance, in text categorization tasks feature set analysis methods presented here could be used to improve feature selection and classification of documents (e.g. from the MEDLINE database), where the documents are represented by ca. 72.000 features (Yang and Pedersen, 1997). In this context, *a priori* defined feature sets could represent sets of terms characteristic of some subject areas.

Results reported in this chapter were partly published in the journal *Briefings in Bioinformatics* (Maciejewski, 2013).

4.1. Introduction to gene set analysis methods

The main purpose of gene set analysis is to overcome the major limitations of purely data-driven class comparison based on high-throughput data. The limita-

tions are related primarily to poor interpretability and reproducibility of results, and they underlie the specific problems commonly observed with class comparison based on $n \ll d$ data (Subramanian *et al.*, 2005):

- Class comparison studies may produce long lists of statistically significant features (e.g. differentially expressed genes). Interpretation of such lists, which is expected to reveal the actual cause of the difference between the classes compared, is often infeasible or largely subjective.
- On the other hand, class comparison studies may produce empty lists of features which remain statistically significant after the multiple testing correction. This is especially likely if the signal in the relevant features is low as compared with the noise, and if d is large.
- It is commonly observed that different high-throughput studies of the same problem (e.g. disease) reveal lists of significant features which show very low overlapping. This problem in $n \ll d$ data can be mainly attributed to the small sample size, as shown in Chapter 3.
- It is often postulated that the actual effect underlying the difference between the classes compared in high-throughput assays in genomics is due to relatively low but coordinated change in a group of related features rather than a big change is some unrelated features. For instance, referring to high-throughput assays in genomics, Subramanian *et al.* (2005) argue that "An increase of 20% in all genes encoding members of a metabolic pathway may dramatically alter the flux through the pathway and may be more important than a 20-fold increase in a single gene". Standard, data driven methods of class comparison are virtually unable to reveal such important feature sets, focusing of highly variable but irrelevant features.

Gene set analysis attempts to ease these interpretability and reproducibility related difficulties by employing *a priori* domain knowledge about groups of features which are presumably related functionally. In this way, the underlying differences between the classes can be expressed in terms of activation of feature sets (such as signalling pathways in case of genomic studies).

Numerous approaches have been proposed which differ in how activation of a gene set is defined and estimated from data. Roughly, some methods analyze association of genes in a gene set with the target while ignoring the remaining genes in the data; other methods compare genes in the gene set with the remaining genes in terms of association with the target. The former methods are known as *self-contained*, and the latter – as *competitive* methods (Nam and Kim, 2008). Goeman and Bühlmann (2007) made an important distinction between the null hypotheses tested by different methods:

- Self-contained null hypothesis assumes that no genes in the gene set concerned are differentially expressed.
- Competitive null hypothesis assumes that the genes in the gene set are not more associated with the target (i.e. differentially expressed) than the remaining genes.

Rejection of the null hypothesis is interpreted as activation of the gene set. The numerous methods of gene set analysis propose different test statistics to verify the self-contained and competitive hypotheses, as well as different procedures to estimate statistical significance (i.e. the p-value) of the test statistic. More specifically, the methods of gene set analysis differ in terms of the model of statistical experiment realized by the methods, i.e. in terms of the following assumptions which underlie the test procedure:

- The methods differ in terms of the definition of the random variables which are actually compared in the test. The data which are used to calculate the test statistic are then regarded as independent samples from these random variables.
- The methods also differ in the actual meaning and interpretation of the null hypothesis tested.
- The methods differ in the the way the statistical significance is evaluated, or more specifically, how the distribution of the test statistic under the null hypothesis (null distribution) is obtained. The methods use one of the following approaches: (i) the null distribution is assumed to follow some known parametric distribution, (ii) the null distribution is obtained by permutation of samples, or (iii) the null distribution is obtained by permutation of genes.
- Since statistical significance (or the p-value) of the test is related to replications of the experiment (as the p-value denotes the probability, assuming H0, of obtaining a more extreme value of the test statistic under many replications of the experiment), the methods *de facto* differ in how replication of the experiment can be realized. We show that some methods imply that replication of the experiment is realized by taking more samples, while some other methods imply that replication of the experiment means taking more genes from a gene set.

Interestingly, authors of many methods do not explicitly state these assumptions; often the underlying assumptions implicitly arise from the statistical hypothesis testing procedure. In this chapter we clarify these important assumptions underlying the numerous methods of gene set analysis.

We first (section 4.2) systematize the methods into the following categories:

- self-contained,
- competitive which use sample permutations to derive the null distribution,
- competitive which use gene permutations to derive the null distribution,
- methods which assume parametric null distribution.

Then in section 4.3, we provide the methodological analysis of the methods in these groups in terms of the model of the statistical experiment. We also analyze compliance of the models with the organization of the actual high-throughput experiment which produced the data. This analysis is intended to indicate which of the methods produce (biologically) interpretable results, i.e. we want to identify the methods whose significant p-value actually denotes significant association of the feature set with the target. These methods will be then used in Chapter 5 in order to incorporate prior domain knowledge in the process of feature selection.

4.2. Mathematical formulation of gene set analysis methods

The mathematical formulation of gene set analysis methods will be done using the following notation. We denote the matrix containing results of a high-throughput study as $W_{d\times n}$, where the columns, denoted $W_{\bullet i}$, $i = 1, ..., n$, represent the d-dimensional vectors of features measured for the n samples tested in the assay. The target values associated with the samples are represented by the vector $Y = (Y_j), i = 1, ..., n$. For instance, in the context of a gene expression study, W can be the matrix with expression of d genes observed for n samples (patients), where the disease status of the samples (e.g. type of leukemia, or cancer vs healthy) is represented by Y.

It is convenient to represent an *a priori* defined gene set (e.g. a signalling pathway) as the set of indices, G, of rows of the matrix W which correspond to the genes in the gene set concerned (see also Remark 5 on page 103 for some technical issues related to this mapping). We denote the number of elements in G as m. We also denote $G^C = \{i : 1 \leq i \leq d \wedge i \notin G\}$ as the complement of G. We represents the subset of rows in W which correspond to G as the $(m \times n)$ matrix $X = (W_{i\bullet}), i \in G$, and similarly $X^C = (W_{i\bullet}), i \in G^C$, where $W_{i\bullet}$ denotes the i-th row of W.

Many genes set analysis methods define the gene set score (i.e. the measure of association of the gene set with the target) as the aggregate of the individual gene

scores (i.e. measures of association of genes in the gene set with the target). If we define the gene score function as τ, then the gene scores are denoted as $t = (t_i)$, where $t_i = \tau(X_{i\bullet}, Y)$, $i = 1, \ldots, m$, and as $t^C = (t_i^C)$, where $t_i^C = \tau(X_{i\bullet}^C, Y)$, $i = 1, \ldots, (d-m)$. Unless otherwise stated, for the binary target, τ will be defined as the t-statistic. We denote the p-values associated with the scores t and t^C as $p = (p_i), i = 1, \ldots, m$ and $p^C = (p_i^C), i = 1, \ldots, (d-m)$.

In sections 4.2.1 through 4.2.4, we present the following groups of methods of gene set analysis: self-contained, competitive with randomization of samples, competitive with randomization of genes, and parametric. We discuss the key methodological differences between these groups of methods, focusing on definition of the gene set score and the procedure to assess statistical significance of results.

4.2.1. Self-contained methods

The null hypothesis of the self-contained methods assumes that no genes in the gene set are associated with the target. The hypothesis is tested based on the data (X, Y), with the matrix X^C ignored. Rejection of the null hypothesis (i.e. p-value < 0.05) is interpreted as *activation* of the gene set.

Here we present the most prominent examples of self contained methods.

1. Globaltest, GT, (Goeman *et al.*, 2004).
   The test is based on the generalized linear model of relationship between X and Y: $g(E(Y_i|\beta)) = \beta_0 + \beta_1 x_{1,i} + ... + \beta_m x_{m,i}$, where g is the link function and β denotes the vector of coefficients in the model. The null hypothesis assumes that the genes are not associated with Y, i.e. $H_0 : \beta_1 = \beta_2 = ... = \beta_m = 0$. The test statistic is derived as:

$$GT = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{\mu_2} \left[X_{i\bullet}(Y - \mu)' \right]^2 \tag{4.1}$$

   where μ is the first and μ_2 is and second central moment of Y. Significance of the test statistic is assessed using permutation of samples, or using the parametric distribution (asymptotic normal distribution or, for small samples, scaled χ^2 distribution).
2. SAM-GS, (Dinu *et al.*, 2007).
   The SAM-GS is the gene set analysis method based on the popular class-comparison SAM algorithm proposed by Tusher *et al.* (2001). The gene set score is defined as:

$$SAMGS = \sum_{i=1}^{m} \left(\frac{\bar{X}_{i\bullet}^{(0)} - \bar{X}_{i\bullet}^{(1)}}{\mathrm{std}(X_{i\bullet}) + s_0} \right)^2 \tag{4.2}$$

where $X_{i\bullet} = (x_{ij})$, $j = 1, ..., n$, is the i-th row of X, $\bar{X}_{i\bullet}^{(0)} = \mathrm{mean}(x_{i,j} : Y_j = 0)$ and $\bar{X}_{i\bullet}^{(1)} = \mathrm{mean}(x_{i,j} : Y_j = 1)$ denote the mean signal for gene i in class 0 and 1, respectively, and s_0 is a small constant added to stabilize the gene set score for the rows with small variability signal. Significance of the test statistic is assessed using permutation of samples.

3. Q2, (Tian *et al.*, 2005).
   The gene set score is defined as the aggregate of the individual gene scores in G, i.e.

$$Q2 = \frac{1}{m} \sum_{i=1}^{m} t_i \tag{4.3}$$

   with significance estimated using randomization of samples.
4. SC.GSEA, self-contained version of the GSEA method proposed by Subramanian *et al.* (2005).
   The self-contained version of the popular GSEA method (see point 1 on page 50) was suggested by Goeman and Bühlmann (2007). The idea is to compare the p-values calculated for the members of G, $(p_i), i = 1, ..., m$, versus the uniform distribution, since under the null hypothesis (members of G not associated with the target), the p-values should follow the uniform distribution. Therefore, the SC.GSEA gene set score is defined as the Kolmogorov–Smirnov statistic comparing $(p_i), i = 1, ..., m$, versus the uniform distribution, with the significance assessment based on (i) permutation of samples or (ii) the analytical Kolmogorov–Smirnov distribution.

A number of similar self-contained methods were also proposed by Fridley *et al.* (2010).

Most of the self-contained methods use sample randomization for significance assessment. Significance (and the p-value) of the test statistic S (where S is the gene set score such as GT or Q2, etc.) is obtained by comparing the value of S observed in the test, denoted S_0, with the distribution of the statistic calculated under the null hypothesis which states that the features in X and the target Y are independent. Hence, the null distribution of S can be obtained by repeatedly calculating S under many permutations of samples, with the corresponding p-value obtained as:

$$p = \frac{1}{B} \sum_{i=1}^{B} I(S_i > S_0) \tag{4.4}$$

where S_i is the value of the gene set score S observed under the the i-th permutation of samples (e.g. permutation of values in Y), I denotes the indicator function and B is the number of permutations. It should be noted that the permutation test is available only in the the sample labels are exchangeable (e.g. if they are drawn as independent realizations from some random distribution).

4.2.2. Competitive methods with randomization of samples

The null hypothesis tested by the competitive methods assumes that the genes in G are not more often associated with the target than the genes outside G. The hypothesis is tested based on the complete data available from the experiment, i.e. (X, X^C, Y). Rejection of the hypothesis (i.e. p-value < 0.05) indicates that the gene set includes significantly more differentially expressed genes than the remaining collection of genes in the experiment, and therefore can be declared as *activated*.

Competitive methods use either gene or sample randomization for assessment of significance of results. In this section, we focus on sample randomization methods. Here we present the most prominent of these methods.

1. GSEA – Gene Set Enrichment Analysis, (Subramanian *et al.*, 2005).
   GSEA is one of the first and best-known methods of gene set analysis proposed in bioinformatics. The GSEA creates the sorted list of p-values calculated for all the genes in the study (i.e. for all the rows of W). We expect that for the null hypothesis, the ranks of genes in G should be uniformly distributed along the list of all genes. Therefore, the GSEA gene set score (referred to as the Enrichment Score, ES) is defined as the Kolmogorov–Smirnov statistic comparing ranks of the p-values of genes in G vs the uniform distribution. (Note that using *ranks* of the p-values makes the method competitive, while using raw p-values, as proposed in the SC.GSEA, makes the method self-contained).
   Significance of the GSEA statistic is obtained using randomization of samples.
2. SAFE, (Barry *et al.*, 2005).
   The SAFE method directly compares the vectors t vs t^C, which is motivated by the assumption that for the null hypothesis the vectors should not differ

in terms of some commonly used measures of similarity. More specifically, the SAFE gene set score is defined as the Kolmogorov–Smirnov or as the Wilcoxon rank-sum statistic comparing the vectors t and t^C.
Significance of the gene set score is assessed using randomization of samples.

3. GSA – Gene Set Analysis, (Efron and Tibshirani, 2007).
The GSA procedure, proposed as the extension of the GSEA method, defines a different gene set score and uses an enhanced procedure of significance assessment. The authors argue that these modifications lead to improved power of the method, which they conclude from empirical experiments. GSA has become one of the most prominent gene set analysis methods.
The GSA gene set score (known as the *maxmean* statistic) is defined as

$$S_{max} = \max\left\{\left|\frac{\sum_{i=1}^{m} I\left(t_i > 0\right) t_i}{m}\right|, \left|\frac{\sum_{i=1}^{m} I\left(t_i < 0\right) t_i}{m}\right|\right\} \tag{4.5}$$

Note that S_{max} represents the mean association of up- or down-regulated features (whichever dominate), however, the mean is taken over all m features in G rather than the actual number of the up- or down-regulated features. This is supposed to weaken the impact of single features with very strong effect. In (Efron and Tibshirani, 2007), the authors also consider simpler statistics, such as $S = \frac{1}{m}\sum_{i=1}^{m} t_i$ (as in Equation (4.3)), or $S = \frac{1}{m}\sum_{i=1}^{m} |t_i|$.
Prior to assessment of significance, which is done using randomization of samples, the test statistic is standardized, which consists in adjusting it using all the rows in W. The p-value of the standardized S statistic is calculated as

$$p = \frac{1}{B}\sum_{i=1}^{B} I\left(\frac{S_i - mean^*}{stdev^*} > \frac{S - mean_S}{stdev_S}\right) \tag{4.6}$$

where the standardization terms $mean_S$ and $stdev_S$ denote the mean and standard deviation of the gene scores calculated for all genes in W, S_i denotes the gene set score calculated for the i-th permutation of samples, B is the total number of permutations, and the standardization terms $mean^*$ and $stdev^*$ denote the mean and standard deviation of individual gene scores calculated over all genes in W and over a large number of permutations. Similar restandardization formula is available for the S_{max} statistic – see (Efron and Tibshirani, 2007).

Note that if the raw gene set scores S or S_{max} were used in the significance assessment procedure (e.g. if $p = \frac{1}{B}\sum_{i=1}^{B} I(S_i > S)$), then the GSA would clearly become a self-contained method. Standardization of the statistics makes the method competitive, as the standardized statistics measure association with the target of the features in G, relative to the average association with the target of all the features in (X, X^C).

4. GSA2 – modified version of GSA. The modified version of the GSA, proposed in (Maciejewski, 2013), aims to correct the inconsistency in the significance assessment procedure employed by the original GSA method. Note that the GSA uses sample randomization to assess significance of the standardized statistic $(S - mean_S)/stdev_S$ – Equation (4.6). The purpose of the permutation procedure is to empirically obtain the null distribution of the test statistic, or more specifically, as realized by Equation (4.6), to assess the probability that a more extreme value of the test statistic $((S - mean_S)/stdev_S)$ would be observed if we repeatedly performed the experiment, assuming that the null hypothesis holds. To do this, we need to calculate the value of test statistic under permutation of samples. Note however, that in formula (4.6), we do not compare the observed test statistic $(S - mean_S)/stdev_S$ with its value calculated for a permutation of samples, but with the value $(S_i - mean^*)/stdev^*$, which does not provide the null distribution of the test statistic $(S - mean_S)/stdev_S$. Hence, the original GSA method does not properly assess the p-value associated with the standardized gene set score.
To correct this we propose that the p-value is assessed as

$$p_{mod} = \frac{1}{B}\sum_{i=1}^{B} I\left(\frac{S_i - mean_{Si}}{stdev_{Si}} > \frac{S - mean_S}{stdev_S}\right) \tag{4.7}$$

where the $mean_{Si}$ and $stdev_{Si}$ are the mean and standard deviation of the individual gene scores calculated for all the genes under the i-th permutation of samples. Note that the left-hand side of the inequality represents the value of the standardized statistic under the i-th permutation of samples. Significant p-value (i.e. $p_{mod} < 0.05$) indicates that the gene is enhanced (i.e. contains higher concentration of differentially expressed genes than the whole dataset W).

4.2.3. Competitive methods with randomization of genes

Numerous competitive methods have been proposed which compare differential expression in X and X^C (or which compare t vs t^C) and use randomization of genes (rows of W) for the assessment of significance, i.e. for derivation of the distribution of the test statistic under the null hypothesis. Significant p-values (i.e. p-value < 0.05) are interpreted by these methods as the indication that the gene set contains more differentially expressed genes than its complement.

In the following section, we analyze methodological problems related to this interpretation. We show that significance assessment based on gene randomization relies on unrealistic assumptions related to the model of statistical experiment, which makes interpretation of the p-values problematic.

Here we present some of the most prominent methods which rely on gene randomization.

1. Q1, (Tian *et al.*, 2005).
   The method uses the same statistic as the self-contained procedure Q2 – Equation (4.3):

$$Q1 = \frac{1}{m}\sum_{i=1}^{m} t_i \tag{4.8}$$

   Assessment of significance of Q1 is done using randomization of genes (rows of W), with the p-value calculated similarly as in Equation (4.4), i.e. $p = \frac{1}{B}\sum_{i=1}^{B} I(Q1_i > Q1)$, where $Q1_i$ is the value of the statistic for the i-th permutation of genes, and B is the number of permutations.
   Note that the gene randomization procedure makes the method competitive, since it actually compares t vs t^C.
2. Functional Class Score, (Pavlidis *et al.*, 2004).
   The idea of the Functional Class Score is to compare p vs p^C, i.e. the p-values corresponding to the rows of X with the p-values corresponding to the rows of X^C. The gene set score is defined as

$$FCS = \frac{1}{m}\sum_{i=1}^{m} -\log(p_i) \tag{4.9}$$

   Significance is assessed using gene randomization, which makes this method competitive (note that sample randomization would make FCS the self-contained procedure).

Similar methods which aggregate individual gene scores for the members of G and use gene permutations for significance assessment were also proposed by Volinia *et al.* (2004) and Breslin *et al.* (2004).

4.2.4. Parametric methods

Numerous methods have been proposed which attempt to use some standard statistics with known null parametric distributions to compare X vs X^C in terms of association with the target. Some methods perform *overrepresentation analysis* to verify if the genes in G are over-represented in the list of genes in W which are declared as differentially expressed. Technically, the methods compare the two binary vectors of length d, L_D and L_G, whose elements represent whether the corresponding row in W is differentially expressed and whether it is a member of the gene set G. The comparison is based on the contingency table, with the p-value of the null hypothesis of no association of L_D and L_G obtained from the Fisher's exact test or from the hypergeometric test. This approach is implemented in several gene set analysis tools, e.g. Scheer *et al.* (2006); Al-Shahrour *et al.* (2005, 2007). Note that the over-representation approach requires that each gene is declared as either differentially expressed or not, which is based on (somewhat arbitrary) threshold assumed on the p-value; it turns out that generally the gene set analysis is very sensitive to this threshold.

Several threshold-free parametric methods have been proposed. For instance, they attempt to directly compare t vs t^C using some standard statistical tests, or they test some statistic calculated from t, for which the null distribution is known. In the next section, we provide methodological analysis of these methods and show that their underlying assumptions lead to difficulties with interpretation of the p-values returned by some of the methods.

Here we present some of the threshold-free parametric approaches.

1. T-profiler, Boorsma *et al.* (2005).
   This method uses the t-test to compare the vectors of individual gene scores in X and X^C. The test statistic is defined as

$$t_G = \frac{\mu_G - \mu_{G^C}}{s\sqrt{\frac{1}{m} + \frac{1}{d-m}}} \tag{4.10}$$

   where μ_G and μ_{G^C} denotes the mean gene score in G and in G^C, respectively, and s is the standard deviation of the gene scores calculated from the pool of $\{G, G^C\}$. In the original work, the authors proposed to use

the log-ratio of gene expressions in the classes compared as the gene score, however, other gene scores could be used, such as t and t^C (which would give $\mu_G = \bar{t}$, $\mu_{G^C} = \bar{t}^C$).
The p-value of the test is obtained from the t-distribution with $m-2$ degrees of freedom.

2. Parametric tests proposed by Irizarry *et al.* (2009).
Irizarry *et al.* (2009) proposed to compare the vectors of individual gene scores, t and t^C, using some well-known statistical tests with known null distributions. This approach was intended to provide a more powerful and computationally simpler procedure as compared with the popular GSEA algorithm, which uses the Kolmogorov–Smirnov statistic and permutation-based significance assessment.
More specifically, Irizarry *et al.* (2009) proposed to compare t vs t^C using the Wilcoxon rank sum statistic and the χ^2 statistic, which test for the difference in location and scale, respectively. The idea was to provide an alternative to the less powerful Kolmogorov–Smirnov test. Significance of the tests for location and scales can be then obtained from the Wilcoxon or the χ^2 null distribution.
3. PAGE – parametric analysis of gene expression, (Kim and Volsky, 2005).
The idea is to compare the mean gene score calculated for the genes in G with the distribution of the gene scores calculated for all the genes in the experiment. Kim and Volsky (2005) proposed to use the fold-change as the gene score (i.e. $\bar{X}_{i\bullet}^{(0)}/\bar{X}_{i\bullet}^{(1)}$, see notation introduced in point 2 on page 48). The test statistic is defined as the z-score:

$$z = \frac{1}{\delta}(\mu - \mu_G)\sqrt{m} \tag{4.11}$$

where μ_G denotes the mean fold-change calculated for the genes in G, and μ and δ denote the mean and standard deviation of fold-changes calculated for all the genes in W. Under the null hypothesis, z should follow the standard normal distribution $z \sim N(0,1)$, which is used to assess significance of the test.
4. Category method, Jiang and Gentleman (2007).
The method defines the gene set score as the average over G of the t-statistics calculated as the individual gene scores:

$$z = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} t_i \tag{4.12}$$

Assessment of significance is based on the assumption that under the null hypothesis, z should follow the normal distribution $z \sim N(0,1)$. Similar method was also proposed by Irizarry *et al.* (2009).

5. Fisher's method for combining p-values, Fridley *et al.* (2010).
   The gene set score defined by this method aggregates the p-values for the genes in G, i.e.

$$F = -2\sum_{i=1}^{m} \log(p_i) \tag{4.13}$$

   This procedure is based on the Fisher's method for combining p-values from independent tests. Since under the null hypothesis, the p-value p_i follows the uniform distribution on $[0,1]$, hence $-2\log(p_i)$ follows the χ^2 distribution with 2 degrees of freedom, and thus, under the self contained null hypothesis, the test statistic F follows the χ^2 distribution with $2m$ degrees of freedom. This distribution can be used to assess the statistical significance of F. Note however that if significance of F was assessed using permutation of samples, then this method would become self-contained (similar to the Q2, Equation (4.3)).

Next, we will analyze methodological differences between the gene set analysis methods presented in this section. We will develop models of statistical experiment explicitly or implicitly assumed by the methods, which provide the context for the interpretation of p-values returned by the methods.

4.3. Methodological analysis – assumptions underlying different gene set analysis methods

We analyze the underlying statistical assumptions which are implied by the different methods of gene set analysis. Although all the methods express their results in terms of p-values, i.e. rely on statistical hypothesis testing, many methods do not precisely state the key assumptions pertaining to the statistical hypothesis test performed. We analyze these assumptions related to the definition of the random variables tested, the size and independence of the random samples taken from the variables, the definition and meaning of the null hypothesis. It should be noted that these assumptions can be inferred from the definition of the test

statistic and/or from the procedure used to obtain the null distribution of the test statistic. For instance, if a gene set analysis method employs the t-test (see the T-profiler, Equation (4.10)), or the Wilcoxon test (see the parametric test, point 2 on page 55) to compare the individual gene scores in G vs the scores in G^C, (i.e. t vs t^C), then this actually means that t and t^C are considered as the samples of size m and $(d-m)$, respectively, composed of iid elements drawn from their underlying random variables, whose distributions we want to compare. In other words, the method relies on the implicit assumption that the elements in t are independent, i.e. that the genes in G are independent (which is a questionable assumption). Similarly, if we analyze the significance assessment procedure, i.e. the method to obtain the null distribution of the test statistic, we can infer the actual null hypothesis assumed. For instance, the gene permutation procedure used to derive the null distribution of the Q1 statistic (Equation (4.8)) actually means that the genes in G come from the same distribution as the genes outside G, as the permutation test *de facto* implies that the vectors representing "individual gene score" and "assignment of a gene to the gene set" are independent. In other words, we assume that the genes in G come the same distribution (which is again a questionable assumption).

We also note that all the methods of gene set analysis express their results in terms of the p-values, i.e. a gene set is declared as significant if the p-value of a specific test is sufficiently small (e.g. p-value < 0.05). However, the p-value is interpretable only in the context of repetitions of the experiment. For instance, p-value < 0.05 means that if we repeated the experiment many times, then only the fraction of 5% repetitions of the experiment would return the data which produce the test statistic more extreme than the actually observed. Therefore, it is important to analyze the methods in terms of how the experiment could be repeated. We will show that the models of statistical experiment underlying some of the methods define repetition of the experiment quite differently than perceived by the researcher who performs the actual experiment (e.g. in genomics).

In sections 4.3.1 through 4.3.4, we clarity the models of statistical experiment explicitly or implicitly assumed by the different groups of gene set analysis methods presented in the previous section. In section 4.3.5, we discuss applicability of different methods for (i) testing self-contained or competitive hypotheses, as formulated by Goeman and Bühlmann (2007) (see page 45), and for (ii) generation of features for sample classification.

4.3.1. Model 1 of statistical experiment – self-contained methods

We first analyze the self-contained methods which use sample randomization for significance assessment. These methods define the test statistic based on (X, Y), ignoring X^C.

The data analyzed by these methods can be regarded as n independent samples $(X_{\bullet i}, Y_i)$, $i = 1, ..., n$, coming from the underlying random variables: $\mathcal{X} \in R^m$ and $\mathcal{Y}$, where $\mathcal{X}$ represents expression of m genes in G, and $\mathcal{Y}$ represents the target (e.g. disease status) of the subjects (e.g. patients) tested.

Assessment of significance relies on the null distribution of the test statistic obtained using (many) sample permutations. Note that this procedure to obtain the null distribution is valid for, and *de facto* implies, the null hypothesis which assumes that expression of genes $\mathcal{X}$ and the target $\mathcal{Y}$ are independent.

Summarizing, self-contained methods with randomization of samples realize the following model of the statistical experiment:

- The data $(X_{\bullet i}, Y_i)$, $i = 1, ..., n$, are considered as n iid samples from the random variables $\mathcal{X} \in R^m$ and $\mathcal{Y}$.
- Interpretation of the random variables: $\mathcal{X}$ represents expression of genes in G, and $\mathcal{Y}$ represents the target (disease status, phenotype) of the subjects tested.
- The null hypothesis states that $\mathcal{X}$ and $\mathcal{Y}$ are independent.
- Repetition of the experiment can be done by measuring gene expressions and the target for new subjects.

This model applies to the methods presented in section 4.2.1, such as the Globaltest, SAM-GS, Q2, or the SC.GSEA (with the null distribution based on permutation of samples). We note that the same model would apply if we used sample randomization to assess significance of the FCS or Category statistics (see Equations (4.9) or (4.12)), i.e. these methods would become self-contained. We also note that Model 1 is not valid if the SC.GSEA uses the Kolmogorov–Smirnov distribution instead of the sample randomization procedure. In the next section, we provide an appropriate model for this case and for similar methods.

4.3.2. Model 2 of statistical experiment – based on analytical distribution of p or t

Several methods rely on testing whether the elements of $(p_1, ..., p_m)$, or the elements of $(t_1, ..., t_m)$ follow some known analytical distribution. For

instance, the SC.GSEA method (point 4 on page 49) tests if $(p_1, ..., p_m)$ come from the uniform distribution on $[0, 1]$. Similarly, the Fisher's method (Equation (4.13)) is based on combining p-values from independent tests, where the p-values are assumed to follow the uniform distribution on $[0, 1]$. Another similar procedure is proposed by Irizarry *et al.* (2009), who argue that under the self-contained null hypothesis, the elements of $(t_1, ..., t_m)$ should follow the normal distribution.

All these approaches regard the elements of $(p_1, ..., p_m)$, or the elements of $(t_1, ..., t_m)$, as independent samples from some distribution, where the distribution is derived analytically assuming that we perform m independent tests of a hypothesis which is assumed to be true. For instance, the methods which test uniformity of $(p_1, ..., p_m)$, rely on the well-known fact that if we perform m independent tests of a (true) null hypothesis, then the resulting p-values are expected to be uniformly distributed on $[0, 1]$.

Therefore, if a gene set analysis method attempts to verify the self-contained null hypothesis by testing whether the elements of $(p_1, ..., p_m)$ come from the uniform distribution, the method *de facto* assumes that we perform m independent tests of association between a gene in the gene set G and the target, where each of the subsequent m tests are based on the data $(X_{i\bullet}, Y)$, $i = 1, ..., m$. The same applies to other related methods which test for some analytical distribution of $(t_1, ..., t_m)$, or of some aggregate of p or of t.

Hence we obtain the following model of statistical experiment which applies to these methods:

- The data (X, Y) represent results of m independent tests, $(X_{i\bullet}, Y)$, $i = 1, ..., m$. Each of the tests is based on n samples $(X_{i,j}, Y_j)$, $j = 1, ..., n$, where $X_{i,j}$, Y_j are taken from the random variables $\mathcal{X} \in R$ and $\mathcal{Y}$. This means that the genes in G are assumed to be independent and to follow the same distribution.
- Interpretation of the random variables: $\mathcal{X}$ represents expression of each of the genes in G, and $\mathcal{Y}$ represents the target (disease status, phenotype) of the subjects tested.
- The null hypothesis states that $\mathcal{X}$ and $\mathcal{Y}$ are independent.
- Repetition of the experiment can be done by measuring gene expressions and the target for new subjects.

This model underlies the following methods of gene set analysis: the SC.GSEA method (point 4 on page 49) employing the Kolmogorov–Smirnov distribution, the Category methods (Equation (4.12)), the Fisher's method for combining independent p-values (Equation (4.13)).

4.3.3. Model 3 of statistical experiment – based on comparison of t vs t^C

Several methods of gene set analysis rely on comparing the vectors of association of genes with the target: t against t^C. For instance, Irizarry *et al.* (2009) propose to directly compare t vs t^C using standard statistical tests for location or scale, such as the t-test, the Wilcoxon rank-sum test or the χ^2 test. The null hypotheses tested by such procedures assume that the random variables which generated the samples t and t^C do not differ in terms of some specific parameter, such as their means (for the case of the t-test), etc. Hence these methods *de facto* assume that t and t^C include independent samples of size m and $(d-m)$, respectively, from some underlying distributions denoted $\mathcal{T}$ and $\mathcal{T}^{\mathcal{C}}$. The tests then compare $\mathcal{T}$ and $\mathcal{T}^{\mathcal{C}}$ assuming the null hypothesis that the distributions do not differ.

Similar tests are realized by competitive methods which employ gene randomization, such as the Q1 (Equation (4.8)) or the FCS (Equation (4.9)). These methods indirectly compare t vs t^C, as the test statistic (such as Q1) is calculated using only genes in G, i.e. ignoring t^C or p^C. However, the gene randomization procedure used to assess significance makes these methods competitive, as is actually compares the vectors t vs t^C. Note that by using gene randomization to estimate the null distribution of the test statistic (such as Q1), we *de facto* assume that the measure of association t and the binary variable which denotes assignment of a gene to the gene set are independent. This assumption further means that we treat the elements of t, and the elements of t^C as the iid samples from some underlying random variables $\mathcal{T}$ and $\mathcal{T}^C$, and that the null hypothesis states that $\mathcal{T}$ and $\mathcal{T}^{\mathcal{C}}$ have the same distribution.

Hence we obtain the following model of statistical experiment:

- The data t and t^C represent iid samples of size m and $(d-m)$, respectively, from the underlying random variables denoted $\mathcal{T} \in R$ and $\mathcal{T}^C \in R$. This *de facto* means that the genes in G (and also in G^C) are assumed to be independent and to follow the same distribution.
- Interpretation of the random variables $\mathcal{T}$ and $\mathcal{T}^C$ is unclear. Informally, we could refer to these variables as association of genes in G or in G^C with the target, however, it is difficult to precisely formulate the meaning of these random variables in accordance with the actual experiment performed.
- The null hypothesis states that the variables $\mathcal{T}$ and $\mathcal{T}^C$ have the same distribution.
- This model of statistical experiment leads to serious difficulties with interpretation of how repetition of the experiment could be done. Typically, the

statistical experiment is repeated by taking more samples from the underlying random variables which are compared or tested. However, 'taking more samples from the variable $\mathcal{T}$' seems meaningless, since the gene set G is of fixed size m, hence $\mathcal{T}$ cannot generate $m+1$ samples. The same applies to the variable $\mathcal{T}^C$, as the number of gene in the experiment (e.g. on the microarray) is fixed.

This model of statistical experiment is realized by the gene set analysis methods which use genes (rather than subjects tested) as the sampling units, i.e. by competitive methods with gene randomization (e.g. Q1 or FCS), by the methods which compare t vs t^C using statistical tests (such as the t-test, the Wilcoxon test, or Kolmogorov–Smirnov test, see point 2 on page 55), or by the PAGE method (Equation (4.11)).

4.3.4. Model 4 of statistical experiment – competitive methods with sample randomization

We now analyze assumptions pertaining to the model of statistical experiment realized by competitive methods which use sample randomization for assessment of significance (section 4.2.2). These methods use X, X^C and Y to calculate the test statistic (similarly to some parametric methods), however an important difference is that these methods estimate the null distribution empirically, using permutations of samples, instead of using the analytical distribution of the test statistic. This considerably changes the model of statistical experiment which underlies these methods as compared with parametric methods of gene set analysis.

The data (Y, X, X^C) analyzed by these methods can be regarded as n independent samples $(Y_i, X_{\bullet i}, X^C_{\bullet i})$, $i = 1, ..., n$, from the underlying random variables denoted $\mathcal{Y}$ (which represents the phenotype of a subject tested), $\mathcal{X} \in R^m$ and $\mathcal{X}^C \in R^{(d-m)}$ (which represent the subject's expression of genes in G and expression of genes in the complement of G, respectively). Since we use sample randomization to generate the null distribution of the test statistic, we *de facto* assume the null hypothesis which states that the variables $\mathcal{X}$ and $\mathcal{Y}$ are independent and that $\mathcal{X}^C$ and $\mathcal{Y}$ are independent.

Summarizing, we obtain the following model of statistical experiment:

- The data $(Y_i, X_{\bullet i}, X^C_{\bullet i})$, $i = 1, ..., n$, represent iid samples from the underlying random variables $\mathcal{Y}$, $\mathcal{X} \in R^m$, $\mathcal{X}^C \in R^{d-m}$.
- Interpretation of the random variables: $\mathcal{Y}$ represents the target (disease status or phenotype), $\mathcal{X}$ represents expression of genes in G, and $\mathcal{X}^C$ represents expression of the remaining genes.

- The null hypothesis assumes that the variables $\mathcal{X}$ and $\mathcal{Y}$ are independent and the variables $\mathcal{X}^C$ and $\mathcal{Y}$ are independent.
- Repetition of the experiment can be done by taking new subjects and measuring their target as well as expression of m genes in G and $(d-m)$ genes outside G.

This model describes the actual hypothesis tested by such methods as the GSEA, SAFE or the GSA2 (point 4 and page 52). Note however, that the original version of the GSA procedure improperly estimates the null distribution of its test statistic and the p-value (Equation (4.6)), therefore this model, strictly, does not apply to the original GSA.

4.3.5. Discussion – applicability of different methods for testing self-contained or competitive hypotheses

Gene set analysis methods aim to quantify activation of gene sets which is defined in terms of self-contained or competitive null hypotheses (Goeman and Bühlmann, 2007). The former are related to whether expression of genes in the gene set is associated with the target, and the latter are related to whether the gene set contains significantly more differentially expressed genes than its complement. Having clarified the models of statistical experiment realized by the different methods, we now want to discuss which of the methods are appropriate for testing the different types of null hypotheses, and which of the methods are based on possibly unrealistic assumptions regarding independence or distribution of genes (features). Based on this, we want to identify the methods which seem preferable for testing activation of gene sets as well as for prior domain knowledge-based feature selection.

The gene set analysis methods which realize Model 1 (section 4.3.1) test the self-contained null hypothesis, i.e. significant p-values of these methods indicate that the gene set concerned contains genes associated with the target. We also note that the sample considered by the methods directly corresponds to the actual sample tested in the high-throughput study, composed of n subjects for which we measure expression of m genes and the value of target. Repeating the experiment could be done by simply taking measurements from new subjects. Therefore, we conclude that Model 1 of statistical experiment directly follows the organization of the actual (biological) study and produces (biologically) interpretable results. We also note that this model does not assume nor implies any unrealistic requirements such as independence or the same distribution of genes in the gene set.

The methods based on Model 2 (section 4.3.2) also test the self-contained null hypothesis, i.e. significant p-values indicate that genes in the gene set are associated with the target. The sample considered by this model directly corresponds to the actual (biological) sample, hence repetition of the experiment could be done be taking measurements from new subjects, therefore the p-value is (biologically) interpretable. We note however, that Model 2 relies on the assumption that the genes in the gene set concerned are independent and that they all follow the same distribution. This requirement is unrealistic in practice, e.g. in high-throughput assays in genomics we expect that gene sets (signalling pathways) include (co)related genes.

The methods based on Model 3 (section 4.3.3) claim to test the competitive null hypothesis. We note however that this model of statistical experiment does not follow the organization of the actual high-throughput study. In this model, we test samples of size m and $(d-m)$ from the underlying random variables, $\mathcal{T}$ and $\mathcal{T}^C$, which we informally define as association of genes in G, and in G^C with the target. However, it is unclear how this definition could be related to the organization of the actual (biological) experiment. For instance, it is unclear how taking new (biological) samples, i.e. testing new subjects, could possibly produce more samples from these variables, as the size of the gene set is fixed. It seems that the major difficulty with this model lies in the fact that it uses genes as the sampling units rather than (biological) subjects which are sampled and tested in the (biological) study. This leads to difficulties with interpretation of the p-values produced by these methods, as what Model 3 defines as repetition of the statistical experiment (i.e. taking new samples from $\mathcal{T}$ and $\mathcal{T}^C$) is not related to what the researcher perceives as repetition of the actual experiment (i.e. taking measurements from new (biological) subjects). Additionally, we observe that Model 3 relies on the assumption that the genes in G are independent and identically distributed. Hence, we conclude that methods based on Model 3 do not produce meaningful results (p-values) which could be interpreted in the context of competitive or self-contained null hypothesis.

Model 4 (section 4.3.4) is similar to Model 1, as it directly corresponds to the organization of the actual (biological) study. The only difference between these models lies in the fact that in Model 1 we measure expression of m genes in the gene set and the target for the n subjects tested, while in Model 4 we additionally measure expression of the $(d-m)$ genes in the complement of the gene set. This leads to the slightly different null hypothesis which, under Model 4, assumes that the genes in G are not associated with the target (which is the self-contained null

tested by Model 1), but additionally that genes in G^C are not associated with the target. We also note that, unlike Models 2 or 3, this model does not assume independence and the same distribution of genes in G. Therefore, we conclude that significant p-values produced by the methods based on Model 4 indicate that either G or G^C contains genes associated with the target. This is equivalent to testing the self-contained null hypothesis providing that the genes in G^C are not associated with the target (which is often a reasonable assumption).

We conclude that gene set analysis methods based on Model 1 (such as the Globaltest, SAM-GS or Tian's Q2) clearly test the self-contained null hypothesis. Methods based on Model 2 (such as the Category or various techniques which rely on combining the p-values in G), or methods based on Model 4 (such as the GSEA, GSA2) also test the self-contained null hypothesis, however they rely on some additional assumptions. Model 2 assumes that genes in the gene set are iid (which is a highly unrealistic assumption in practical applications). Model 4 tests the self-contained null under the assumption that the genes in G^C are not associated with the target (which is often a reasonable assumption).

We also conclude that since Model 3 of statistical experiment is not in line with the organization of the actual (biological) study, the methods based on this model (such as the PAGE or Tian's Q1) fail to produce the p-values which could be interpreted in the context of either competitive or self-contained null hypothesis. Additionally, these methods rely on the strong assumption that the genes in G are independent and identically distributed.

For these reasons, the preferable methods of gene set analysis to be employed in the task of prior domain knowledge-based feature selection are the ones based on Model 1 or Model 4 of statistical experiment. In section 4.4, we will additionally compare these methods empirically in terms of the power and type I error, focusing on correlated and low signal-to-noise data.

4.3.6. Heuristic interpretation of competitive methods based on Model 3

We showed that the methods of gene set analysis based on Model 3 consider genes as sampling units in the procedure of statistical hypothesis testing. Since in the actual study we consider the subjects (e.g. patients) as the sampling units, this discrepancy leads to serious difficulties with interpretation of the random variables tested and of the p-values obtained. Despite of these methodological problems, methods based on Model 3 are commonly implemented in numerous gene set analysis tools (as shown in the review by Nam and Kim (2008)). Another disap-

pointing conclusion drawn in section 4.3.5 is that none of the methods concerned address the competitive null hypothesis in statistically sound way. Therefore in this section we want to propose a new interpretation of the popular gene sampling methods, which allows us to compare gene in G with the genes in G^C in terms of association with the target, as formulated by the competitive hypothesis. This interpretation is heuristic, i.e. does not rely on the statistical hypothesis testing (and hence does not produce a p-value).

In this model we compare genes in G with the genes in G^C in terms of association with the target. Given a sample of n subjects, we measure association of the genes in G and in G^C with the target, denoted t and t^C, respectively. We define $f(t)$ as some aggregate of elements of t (e.g. $f(t) = (\sum_{i=1}^{m} t_i)/m$ or $f(t) = (\sum_{i=1}^{m} |t_i|)/m$).

We address the research question similar to the one formulated as the competitive null hypothesis: we want to verify whether G contains more genes associated with the target than other subsets of genes of size m drawn from G^C. To quantify this, we randomly draw many subsets of size m from G^C and calculate the heuristic measure which compares genes in G with the genes in these subsets:

$$s = \frac{1}{B} \sum_{i=1}^{B} I\left(f(\tau_i) > f(t)\right) \tag{4.14}$$

where the vector τ_i contains the measures of association of the m genes in the i-th subset with the target, and B is the total number of subsets selected.

Small values of s indicate that gene sets randomly selected from G^C are unlikely to be stronger associated with the target than the genes in G, which we can interpret as activation (enhancement) of the gene set G. More specifically, s can be interpreted as the fraction of gene sets drawn from G^C which are more activated than G. Clearly this is not a p-value, as our new interpretation has nothing to do with the statistical hypothesis testing.

4.4. Empirical evaluation of power and type I error

In this section, we compare performance of different gene set analysis methods as a function of signal-to-noise and correlation of genes in the gene set. We will compare the methods in terms of (i) type I error (i.e. the probability that a gene set which is not activated will be erroneously identified as activated) and (ii) power (i.e. the probability that a gene set that is actually activated is identified

as such). It should be noted that some of the methods (i.e. the methods based on Models 2 and 3) rely on the assumption that gene expressions are independent. Since this is unrealistic in practice, in this empirical study we want to analyze performance of the methods if data do not meet this assumption.

The second purpose of this study is to quantify power of different methods given data with very low signal to noise level. We want to observe which of the methods are sensitive enough to detect activation of gene sets whose individual features (genes) are weakly regulated and thus are hardly detectable using standard feature selection methods described in sections 2.2 and 2.3. Finally, we want to compare power of different methods if gene sets analyzed contain only a small fraction of genes strongly regulated, with the remaining genes not associated with the target. If excessive power is demonstrated by some methods given such data, then this has to be regarded as a drawback of the methods, when the purpose is to discover weak but coordinated change of expression of the members of a gene set.

This empirical study is based on simulated data only, as we want to evaluate performance of the methods given data with controlled characteristics regarding signal-to-noise ratio, mutual correlation of members of a gene set, etc. It should be noted that in addition to this empirical evaluation based on simulated data, we also provided empirical comparison of selected methods based on real data from high throughput assays. Results are reported e.g. in the papers (Maciejewski, 2011a, 2012).

4.4.1. Analysis of the false positive rate

We first quantify the false positive rate (type I error) under varying correlation of genes in the gene set. In this study no genes are associated with the target (i.e. differentially expressed), but some of the genes are correlated. The purpose of this is to compare performance of different methods, in particular of the methods which assume mutual independence of features (see Models 2 and 3 – sections 4.3.2 and 4.3.3), given data with dependent features.

Since all the methods are based on statistical hypothesis testing, then for the true null hypothesis (as assumed in this study) the methods are expected to falsely reject the fraction of 5% null hypotheses in the series of repeated experiments, provided that rejection is based on the p-value $p < 0.05$ threshold. We want to observe which of the methods realize excessive (i.e. higher than 5%) false positive rate and how this depends on the mutual correlation of genes in the gene set of interest. As we will see in this study, some of the methods tend to declare gene sets as activated based solely on correlation of members, and not on their association with the target.

We use a simulated data set with $n = 30$ samples and $d = 1000$ genes. We define one gene set G which contains the first $m = 40$ genes. In this study no genes in G or in G^C are associated with the target, however the genes in G are correlated. Expression of the genes in G is generated from the multivariate normal distribution with the mean equal 0 (for each of the m genes). The covariance matrix has diagonal elements equal 1 and non-diagonal elements equal r (this parameter in varied the study). Note that since variances of genes are 1, r is also the correlation of genes in G. The $d - m$ genes in G^C are generated from the standard normal distribution N(0,1).

We generate this dataset 500 times and record the p-values returned by different gene set analysis methods. The false positive rate of each method is estimated as the fraction of experiments (out of 500) in which we observed significant p-value (i.e. $p < 0.05$).

The false positive rates as a function of correlation r are reported in Table 4.1.

Table 4.1. False positive rates for different methods of gene set analysis as a function of mutual correlation of the members of G

Method	Model	Correlation in G				
		0	0.2	0.4	0.6	0.8
GT	1	0.050	0.036	0.038	0.062	0.066
Q2	1	0.054	0.034	0.046	0.044	0.050
SC.GSEA.Perm	1	0.048	0.052	0.054	0.052	0.054
SC.GSEA.KS	2	0.044	0.130	0.318	0.556	0.822
Category	2	0.066	0.482	0.632	0.682	0.772
Q1	3	0.050	0.096	0.180	0.220	0.280
PAGE	3	0.030	0.148	0.320	0.606	0.702
t vs t^C, Wilcoxon	3	0.056	0.516	0.644	0.716	0.802
GSEA	4	0.034	0.070	0.066	0.062	0.060
GSA	unclear	0.056	0.072	0.070	0.086	0.096
GSA2	4	0.050	0.057	0.047	0.060	0.050
SAFE.W	4	0.053	0.043	0.060	0.040	0.057
SAFE.KS	4	0.057	0.050	0.063	0.057	0.040

The methods are denoted as in sections 4.2.1–4.2.4, and are grouped by the model of statistical experiment actually performed, as defined in sections 4.3.1–4.3.4. Specific settings: GT (Equation (4.1)) employs the asymptotic null distribution; SC.GSEA.Perm (point 4 on page 49) estimates significance based on permutation of samples; SC.GSEA.KS (point 4 on page 49) estimates significance based on the Kolmogorov–Smirnov distribution; t vs t^C, Wilcoxon (point 2 on page 55) uses the Wilcoxon test to compare t vs t^C; SAFE.W and SAFE.KS (point 2 on page 50) use the Wilcoxon or Kolmogorov–Smirnov distribution, respectively.

We observe that for uncorrelated genes, all the methods realize the false positive rate equal approx. 0.05, which is expected considering the fact that in this study the null hypothesis is true, and that we reject the hypothesis at the p-value$<$ 0.05 threshold. The only methods that seem slightly too conservative are the GSEA and PAGE, since their false positive rate is approx. 0.03. For correlated genes, the methods based on Models 2 or 3 realize excessive false positive rates, and we observe that the false positive rates increase for growing correlation of gene set members. We conclude that the methods which rely on the assumption that members of the gene set are independent, tend to declare as activated the gene sets with only correlated members, but not related with the target. Note that this conclusion holds for both self-contained and competitive null hypothesis, as in our study both are true: self-contained (G does not include genes associated with the target), and competitive (G does not include more genes associated with the target than G^C).

This effect is not observed for the methods based on Models 1 or 4, which do not rely on the assumption that gene set members are independent: correlation of gene set members does not boost the false positive rate. In this study we also compare the original GSA method (point 3 on page 51) with the modified version GSA2 (point 4 on page 52). Note that, strictly, the GSA is not based on Model 4, since the permutation procedure does not estimate the null distribution of its gene set score. Hence, what the method returns as the p-value, is not truly a p-value; this flaw in significance assessment may account for excessive false positive rates of the GSA. We clearly see that GSA2, i.e. the version modified according to Equation (4.7) brings the false positive rate to the expected 5%, irrespective of the correlation level.

Our numerical experiment confirms that significance of the GSA statistic should be evaluated according to Equation (4.7), as proposed in this work, rather than according to Equation (4.6), as proposed by Efron and Tibshirani (2007).

4.4.2. Power under the self-contained hypothesis

In this experiment we compare power of different methods given data with varying number of differentially expressed genes in the gene set and varying psignal-to-noise ratio (i.e. effect strength). We will also control the level of mutual correlation of gene set members. We want to observe which of the methods are most sensitive to discover as activated gene sets with low signal-to-noise level. On the other hand, we want to see which of the methods are over-sensitive in that they tend to identify as activated gene sets with only very few strongly activated

genes and the remaining majority unactivated. This analysis is motivated by the concern expressed by some authors, e.g., Nam and Kim (2008), that self-contained methods may be over-sensitive if the gene sets include only a few genes associated with the target.

Since we analyze the methods in the context of prior domain knowledge-based feature selection, we focus in this study on the power related to the self-contained hypothesis. We want to quantify sensitivity of different methods for detection of gene sets associated with the target rather than for detection of gene sets which contain more differentially expressed members than their complements. The former, i.e. sensitivity under the self-contained hypothesis is more informative in the context of feature selection. The latter, i.e. sensitivity under the competitive hypothesis is analyzed in (Maciejewski, 2013).

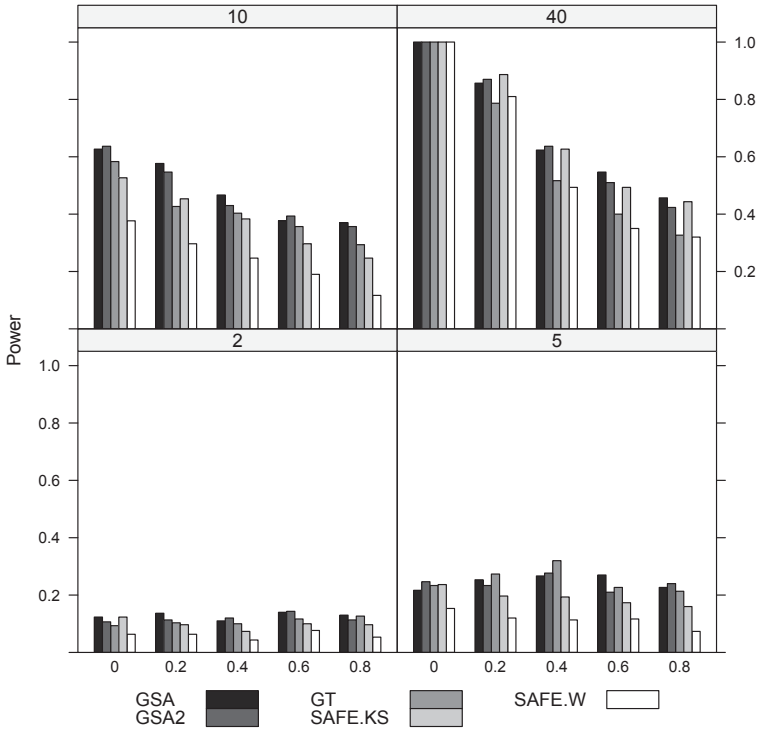

Fig. 4.1. Power of selected methods as a function of correlation and the number of differentially expressed genes (n.DE) in the gene set. Small effect, $\Delta = 0.5$. Power is quantified as the fraction of experiments in which the gene set is declared as activated

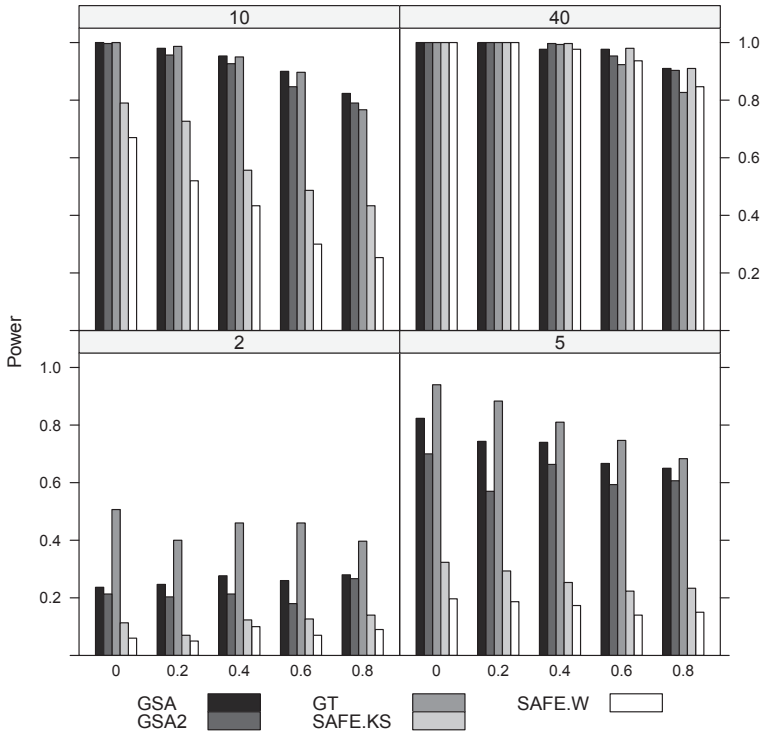

Fig. 4.2. Power of selected methods as a function of correlation and the number of differentially expressed genes (n.DE) in the gene set. Medium effect, $\Delta = 1$. Power is quantified as the fraction of experiments in which the gene set is declared as activated

This analysis is done using a dataset with $n = 30$ samples and $d = 1000$ genes, out of which $m = 40$ genes constitute a gene set G. We assume that $n.DE$ genes in this gene set are differentially expressed and correlated. Expression values of the remaining $m - n.DE$ genes in G, and of the $d - m$ genes in the complement of G are independent and not associated with the target, and are generated from the standard normal distribution N(0, 1). Expression of the $n.DE$ genes in G is generated from the multivariate normal distribution, with the covariance matrix as in the previous study (i.e. diagonal elements equal 1, non-diagonal elements equal r). However, the mean for each of the genes in the first group of $\frac{n}{2} = 15$ samples equals 0, while in the second group the mean equals Δ.

In this study, we vary the number of differentially expressed genes in G, $n.DE = 2, 5, 10, 40$, the correlation coefficient, $r = 0, 0.2, 0.4, 0.6, 0.8$, and the parameter Δ which represents signal-to-noise ratio, or the effect strength, $\Delta = 0.5, 1, 1.5$.

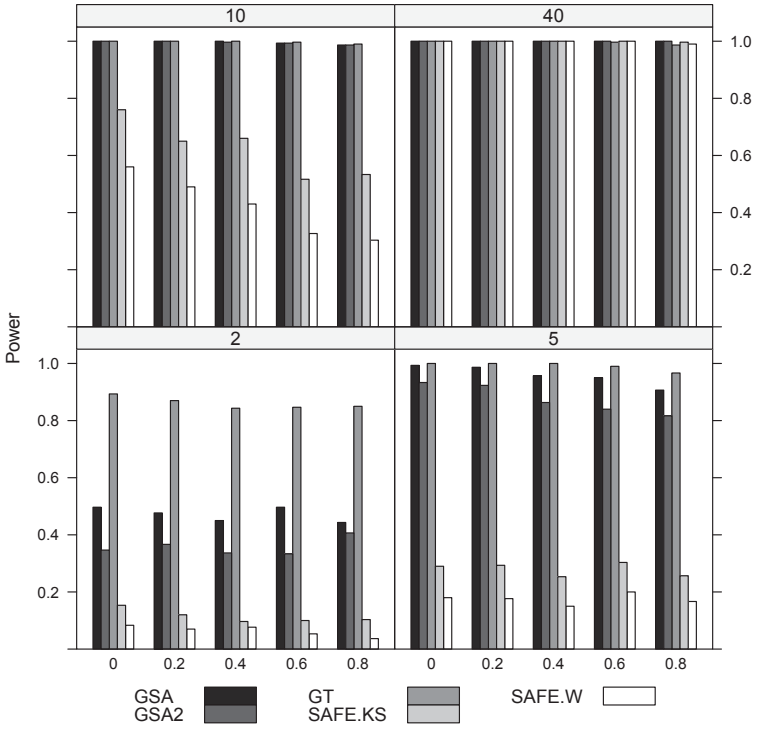

Fig. 4.3. Power of selected methods as a function of correlation and the number of differentially expressed genes (n.DE) in the gene set. Strong effect, $\Delta = 1.5$. Power is quantified as the fraction of experiments in which the gene set is declared as activated

We quantify the power of different methods as a function of these parameters. We estimate power of a method as the fraction of repetitions of the experiment, in which the method yields significant p-value, $p < 0.05$, which means that the gene set is declared as activated.

In this analysis we omit the methods based on Model 2 or 3, due to their excessive false-positive rates under correlation of genes. We focus on selected, most prominent methods which implement Model 1 or 4 (and we also include the important and popular GSA method). Results for subsequent levels of effect strength, $\Delta = 0.5$, 1, 1.5, are presented in Figures 4.1–4.3. In each of the figures, the four panels correspond to the subsequent values of $n.DE = 2$, 5, 10, 40. In the figures, we denote the methods as in Table 4.1.

We first observe that the methods demonstrate significantly different power if the gene sets contain a small number of members associated with the target (see bottom panels in Figures 4.1–4.3, which correspond to $n.DE = 2$ or 5; com-

pare also panels for $n.DE = 10$). For small effect (Figure 4.1), the GSA (GSA2) method generally realizes slightly better power than other methods, while for stronger effects – the Globaltest (GT) demonstrates highest power, with the most remarkable difference between the methods observed for $n.DE = 2$. For instance, for $n.DE = 2$ and $\Delta = 1.5$, the Globaltest realizes power ~ 0.9, while the GSA2 and SAFE realize power ~ 0.4 and ~ 0.1, respectively. These results indicate that if a gene set happens to contain only a few genes strongly associated with the target, then the GT algorithm is likely to declare the gene set as significant, which we can interpret as over-sensitivity of this method, if the task is to identify gene sets with moderate effect detected over many genes.

We also observe that all the methods seem to loose power with growing correlation within the gene set; this effect is most remarkable for the small effect (Figure 4.1, top panels). In section 4.6, we analyze this effect in detail.

4.5. Discussion

In this chapter, we identified the models of statistical experiment which are actually realized by different methods of gene set analysis. These models explicitly state the null hypothesis tested by each of the methods and provide proper interpretation of the p-values produced by the methods. Based on this we draw the following conclusions:

- The methods which use gene randomization or parametric models for estimation of significance of gene set scores (i.e. methods based on Models 2 and 3, which includes such popular methods as Tian's Q1 statistic, Functional Class Score, PAGE, etc.), do not produce p-values which are interpretable in the context of either self-contained or competitive hypotheses. These methods are based on the model of statistical experiment which does not reflect organization of the actual high-throughput study which generated the data, and/or rely on unrealistic assumptions regarding independence and distribution of genes in gene sets. As a consequence, these methods tend to declare as significant the gene sets which include only correlated genes, but not associated with the target.
- Only the methods based on Model 1 or 4 produce statistically interpretable results. This group includes self-contained methods or competitive methods which use sample randomization for estimation of significance (most prominent examples are the Globaltest, GSEA, GSA (GSA2), or SAFE). We note however that the important GSA algorithm does not properly estimate statistical significance of gene set scores, which results in slightly excessive type I error. A corrected version, GSA2, should be used instead. The type I error

of these methods is not affected by correlation of genes in gene sets, however the methods in this group may significantly differ in terms of power. In the algorithms which implement domain knowledge-based feature selection, developed in Chapter 5, we will employ gene set analysis methods which belong to this group.

4.6. Comment about the power of self-contained methods as a function of correlation of features

Our numerical studies show that the power of self-contained methods generally decreases with increasing correlation of genes in the gene set analyzed. This is consistently observed for small (Figure 4.1, top panels), medium (Figure 4.2, top panels) and strong effect (Figure 4.3, top left panel), providing the power curve does not saturate at the value of 1 (as in Figure 4.3, top right panel) or around 0 (as in Figure 4.1, bottom left panel, where the effect of correlation is weak due to very few correlated genes involved). Similar observation can be made for the case of testing the competitive null hypothesis, see e.g. (Maciejewski, 2013), where we showed that power of such methods as GSA, GSA2 or SAFE gets remarkably lower for correlated genes.

This effect of loosing power with increasing correlation of gene set members is commonly observed in several comparative studies which attempt to empirically evaluate gene set analysis methods (see e.g., Ackermann and Strimmer (2009); Liu *et al.* (2007)). Despite slightly different organizations of the simulation experiments, these authors consistently report that correlation in gene sets leads to decreased power of gene set analysis methods.

However, it is interesting to observe that a different organization of the simulation experiment may bring opposite conclusions, as shown in the study by Fridley *et al.* (2010). Fridley *et al.* (2010) used the quantitative target Y generated from the normal distribution with the mean proportional to the subject's gene expression. In this case the power increases with growing correlation of genes in the gene set, as shown in (Fridley *et al.*, 2010), and in Table 4.2.

More specifically, in Table 4.2 we summarize a simple numerical experiment where we compare the two simulation settings which lead to contradicting conclusions regarding the effect of correlation of features:

- The study by Fridley *et al.* (2010) with the following specific settings: expression for $n = 30$ samples $X_{\bullet i}, i = 1, ..., n$ (see notation introduced in section 4.2) is generated from the multivariate normal distribution $MVN(0, \Sigma)$ of dimensionality $m = 40$ (i.e. we have 40 genes in the gene set). The covari-

Table 4.2. Mean p-value returned by the Globaltest for the continuous or binary target, under increasing correlation and strength of effect

Experiment	Effect strength	Correlation in G			
		0	0.1	0.3	0.5
Fridley et al.	$\beta = 1$	0.031	4.97E−08	2.63E−23	9.07E−31
	$\beta = 2$	0.028	8.10E−12	2.00E−25	5.26E−36
	$\beta = 3$	0.029	1.33E−07	1.19E−24	7.54E−30
Classification	$\Delta = 0.5$	0.00031	0.018	0.094	0.171
	$\Delta = 1.0$	6.12E−14	2.92E−06	0.00064	0.0085
	$\Delta = 1.5$	5.10E−19	2.48E−10	5.40E−06	9.18E−05

ance matrix $\Sigma_{m\times m}$ has the diagonal elements equal 1 and the remaining elements equal r (in the study we make $r = 0,\ 0.1,\ 0.3,\ 0.5$). Note that since the variance of genes is 1, the correlation of genes in the gene set also equals r. The value of quantitative target Y_i for each sample $X_{\bullet i}$ is generated from the normal distribution

$$\begin{aligned} &Y_i \sim N(\mu_i, \sigma) \\ &\text{with: } \mu_i = \beta X_{\bullet i}, \sigma = 1 \end{aligned} \tag{4.15}$$

where β is the effect strength ($\beta = [1,\ ...,\ 1]_{1\times m}$ or $\beta = [2,\ ...,\ 2]_{1\times m}$, or $\beta = [3,\ ...,\ 3]_{1\times m}$), which, for simplicity, is denoted in Table 4.2 as $\beta = 1$, etc.

- The classification study similar to the one used in section 4.4.2. We generate expression of $n = 30$ samples divided into two groups of 15 samples with $Y = 0$ in one group and $Y = 1$ in the other group. Expression of $m = 40$ genes in the gene set is generated for a sample using the multivariate normal distribution $MVN(0, \Sigma)$ (for the samples in the first group), and using $MVN(\Delta, \Sigma)$ (for the samples in the other group). Δ is the effect strength; in the study we set $\Delta = 0.5,\ 1,\ 1.5$. The covariance matrix $\Sigma_{m\times m}$ is as in the previous study by Fridley *et al.* (2010).

In the numerical experiment we demonstrate power of the Globaltest method under these simulation scenarios. Although we focus on the Globaltest only, the observations we make and conclusions we draw hold for other self-contained methods.

We generate the data sets 500 times and observe the fraction of experiments out of 500 where the p-value calculated by the Globaltest is significant, $p < 0.05$, (i.e. we identify the gene set as activated). This we interpret as power of the Globaltest. We also record the mean p-value calculated over 500 replications of the experiment. The mean p-value as a function of correlation of genes r and the

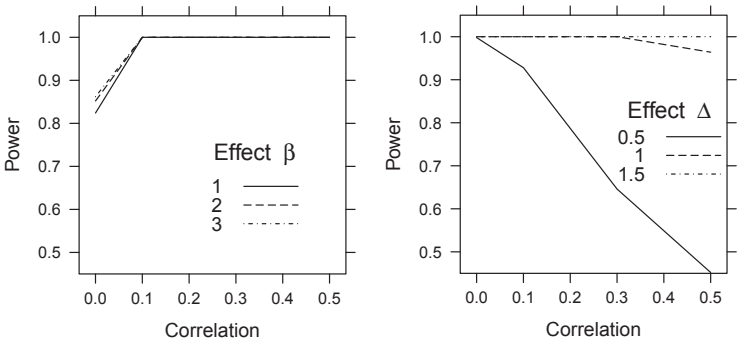

Fig. 4.4. Power (measured as the fraction of significant gene sets with p-value < 0.05) demonstrated by the Globaltest as a function of correlation of genes in the gene set. In the Fridley *et al.* experiment (left panel) the target Y is continuous; in the classification experiment (right panel) Y is binary

effect strength (β or Δ) is reported in Table 4.2 while the power is reported in Figure 4.4.

In the case of the study by Fridley *et al.* (2010), we observe that the mean p-value decreases with the growing correlation of genes. This effect is very strong, e.g., the mean p-value about 0.03 for uncorrelated genes drops to about 10^{-7} for the correlation of 0.1 (Table 4.2, top section), which leads to the growth of power to 1 (Figure 4.4, left panel).

On the contrary, for the classification study the power of the Globaltest decreases with the growing correlation of features (Figure 4.4, right panel). This can be accounted for by observing that growing correlation boosts the p-values (Table 4.2, bottom section). For instance, for the small effect ($\Delta = 0.5$), correlation of 0.3–0.5 makes the method virtually loose power (power = 0.45 for the correlation of 0.5).

The effect demonstrated by the study reported by Fridley *et al.* (2010) is quite easy to explain. Recalling that the Globaltest tests the linear relationship between Y_i and $X_{\bullet i}$, $i = 1, ..., n$ (see point 1 on page 48), we observe that by increasing the correlation in the gene set, i.e. including correlated values in the vector $X_{\bullet i}$ in Equation (4.15), we ensure stronger linear relationship between the target Y_i and expression of genes in the samples $X_{\bullet i}$, $i = 1, ..., n$, as compared with the case where the elements of the vectors $X_{\bullet i}$ are independent. Now we want to account for the effect of loosing power with correlation of genes, observed in the classification study (Figure 4.4, right panel, Table 4.2, also Figures 4.1–4.3) and in virtually all similar comparative studies published in literature, e.g. (Ackermann and Strimmer, 2009; Liu *et al.*, 2007).

This effect can be accounted for analytically by looking at the overlapping of the two multivariate distributions used to generate expression data in the classes of samples, under increasing correlation. To explain this, we introduce the following notation: let us denote the density of the two multivariate random variables $\mathcal{X}_0$, $\mathcal{X}_1 \in R^m$ we used to generate expression data for samples in the two classes as $f_0(x_1, ..., x_m)$ and $f_1(x_1, ..., x_m)$, where f_0 was used to generate the vectors (samples) $X_{\bullet i}$ for which $Y_i = 0$, and f_1 – to generate the remaining samples. In the studies related to the power of self-contained methods under correlation of features, it is often assumed that f_0 and f_1 are multivariate normal distributions (MVN):

$$\begin{aligned} \mathcal{X}_0 &\sim MVN(\mu_0, \Sigma) \\ \mathcal{X}_1 &\sim MVN(\mu_1, \Sigma) \end{aligned} \tag{4.16}$$

where the distributions have the same covariance matrices but are shifted in location. As before, we assume that the diagonal elements of $\Sigma_{m \times m}$ are equal 1, and non-diagonal elements are equal r. Note that r can be interpreted as correlation of gene expressions, as variances of gene expressions are 1.

4.6.1. Simple model of activation of a gene set

The simple model of the effect of activation of a gene set (or activation of as signalling pathway) can expressed as:

$$\begin{aligned} \mu_0 &= [0, ..., 0]_m \\ \mu_1 &= [\Delta, ..., \Delta]_m \end{aligned} \tag{4.17}$$

which means that the (correlated) expressions of genes in the activated pathway are (slightly) shifted by possibly small effect Δ, as compared with the baseline expressions of the inactivated pathway. This model is used by some authors, e.g. Subramanian *et al.* (2005) actually mean this model when they argue that "An increase of 20% in all genes encoding members of a metabolic pathway may dramatically alter the flux through the pathway and may be more important than a 20-fold increase in a single gene".

Now, we show that this simple and quite obvious model of activation of a gene set (or signalling pathway) leads to the aforementioned effect of loosing power by the gene set analysis methods under correlation of genes. To show this, we note that self-contained methods test association between the target and gene expressions in the gene set based on the data $(X_{\bullet i}, Y_i)$, $i = 1, ..., n$, where the samples $(X_{\bullet i})$, $i = 1, ..., n$ were generated partly from $\mathcal{X}_0$, and partly from $\mathcal{X}_1$. Clearly,

the higher the overlapping of the densities of the variables $\mathcal{X}_0$ and $\mathcal{X}_1$, the more similar the vectors of expression $X_{\bullet i}$ tend to be between the classes. Hence the methods such as the Globaltest are less likely to discover the relationship between Y and the gene expressions, which results in reduced power of the method. The amount of overlapping of the densities of $\mathcal{X}_0$ and $\mathcal{X}_1$ can be calculated as:

$$\begin{aligned} b &= \int_{f_1<f_2} \cdots \int f_1(x_1, ..., x_m)dx_1...dx_m + \int_{f_1>f_2} \cdots \int f_2(x_1, ..., x_m)dx_1...dx_m \\ &= \int \cdots \int \min(f_1(x_1, ..., x_m), f_2(x_1, ..., x_m))dx_1...dx_m \end{aligned} \tag{4.18}$$

Note that b measures the Bayes error of the classifier built from the data $(X_{\bullet i}, Y_i)$, $i = 1, ..., n$, assuming equiprobable classes (Hastie *et al.*, 2001).

Therefore it is informative to observe the overlapping b as a function of correlation r and dimensionality of data m. We calculated b numerically for the dimensionality $m = 2, ..., 8$ and report the results in Table 4.3, for the effect strength $\Delta = 1$.

We clearly see that for a fixed dimensionality of the gene set, increasing correlation of features leads to higher overlapping of the densities f_0 and f_1. Hence the samples generated from the random variables $\mathcal{X}_0$ and $\mathcal{X}_1$ (i.e. the data vectors $\{X_{\bullet i} : Y_i = 0\}$ and $\{X_{\bullet i} : Y_i = 1\}$) are more difficult to distinguish. Thus in many realizations of the simulation study reported in this section (or in section 4.4.2), the relationship between $X_{\bullet i}$ and Y_i, $i = 1, ..., n$, turns out to be weak (with the p-value of the Globaltest exceeding 0.05), which leads to reduced power of the gene set analysis methods concerned.

Table 4.3. The overlapping b (Equation (4.18)) of the densities of $\mathcal{X}_0$ and $\mathcal{X}_1$ (Equation (4.16)) for $\mu_0 = [0, ..., 0]_m$ and $\mu_1 = [1, ..., 1]_m$, as a function of the number of genes in the gene set (m) and their correlation (r)

Dimensionality m	Correlation r				
	0	0.2	0.4	0.6	0.8
2	0.479	0.519	0.55	0.576	0.598
3	0.386	0.465	0.518	0.559	0.591
4	0.317	0.429	0.5	0.55	0.587
5	0.264	0.406	0.489	0.544	0.585
6	0.221	0.387	0.482	0.543	0.584
7	0.184	0.336	0.492	0.555	0.572
8	0.166	0.348	0.486	0.542	0.581

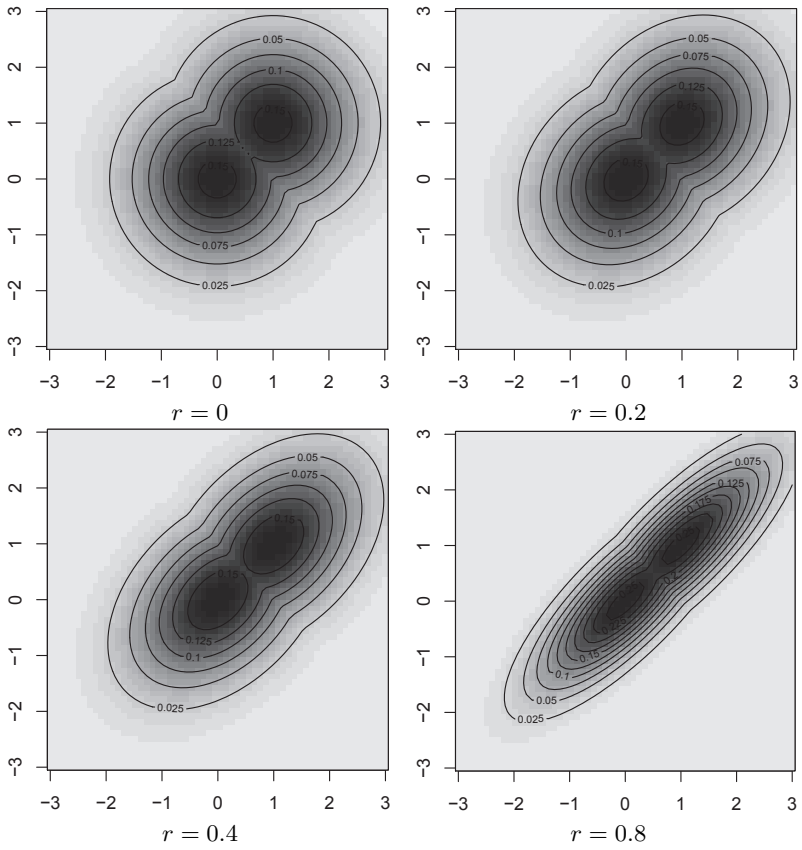

Fig. 4.5. Contour plots of the densities of the random variables $\mathcal{X}_0$, $\mathcal{X}_1$ (Equation (4.16)) as a function of correlation of features r. The density of $\mathcal{X}_0$ is centered at $\mu_0 = [0, 0]$, and the density of $\mathcal{X}_1$ is centered at $\mu_1 = [1, 1]$

To understand the nature of the effect reported in Table 4.3, we present visualization of the overlapping densities under increasing correlation of the features – Figure 4.5. The plot is done for two-dimensional data (i.e. the axes are labeled by x_1, x_2 – expressions of two genes in the gene set). The figures illustrate the effect shown in Table 4.3, of increasing overlapping of densities of the two classes of samples.

It also interesting to observe that although for uncorrelated features the overlapping of densities decreases with growing dimensionality m (see Table 4.3, column for $r = 0$), increasing correlation again leads to more overlapping (hence poor separability) of the classes.

4.6.2. Activation of a gene set requires suppression of inhibitors

In the previous section, we assumed a simple model of activation of a gene set (or signalling pathway), as given by Equation (4.17). This model essentially means that a slight, coordinated up-regulation of the members of the pathway is equivalent to activation of the entire pathway, which seems reasonable and biologically justified (Subramanian *et al.*, 2005).

However, we can envisage another model where in order to activate a pathway, we need to ensure simultaneous effects of:

- up-regulation of some members of the pathway, which yield the product of the pathway,
- down-regulation (suppression) of the members of the pathway which play the role of inhibitors, i.e. block expression of other genes in the pathway.

This scenario is also biologically relevant (personal communication with dr M. Jank, SGGW).

This motivates the following model (refer to Equation (4.17)):

$$\begin{aligned} \mu_0 &= [0, ..., 0]_m \\ \mu_1 &= [\Delta, -\Delta, \Delta, ...]_m \end{aligned} \tag{4.19}$$

where the negative elements in the vector μ_1 are related to the inhibitor genes and positive elements – to the up-regulated genes (here we assume for simplicity that the number of inhibitor genes is $\frac{m}{2}$, so that the scalar product $\left\langle \mu_1^{Equation(4.17)}, \mu_1^{Equation(4.19)} \right\rangle = 0$, i.e. the vectors are perpendicular).

If we assume that f_0 and f_1 have the same covariance matrices Σ (the same as in section 4.6.1), we actually postulate that the densities f_0 and f_1 are shifted relative to each other along the direction perpendicular to the direction of high variability in data. Note that in the previous case (Equation (4.17)) the densities were shifted along the direction of high variability in correlated data (as illustrated in Figure 4.5).

It is now interesting to observe how the measure of overlapping b changes under the correlation r. We provide the results in Table 4.4, based on numerical integration of Equation (4.18) for m up to 8. We clearly see that increasing correlation reduces the overlapping b and thus improves separability of classes represented by the variables $\mathcal{X}_0$ and $\mathcal{X}_1$. We illustrate this effect graphically by the contour plots of the densities f_0 and f_1 plotted for the dimensionality $m = 2$, see Figure 4.6. We illustrate the overlapping densities for correlation $r = 0.8$ under the previous model (Figure 4.7), and under the current model (Figure 4.8).

Table 4.4. The overlapping b (Equation (4.18)) of the densities of $\mathcal{X}_0$ and $\mathcal{X}_1$ (Equation (4.16)) for $\mu_0 = [0, .., 0]_m$ and $\mu_1 = [1, -1, 1, ..]_m$, as a function of the number of genes in the gene set (m) and their correlation (r)

Dimensionality m	Correlation r				
	0	0.2	0.4	0.6	0.8
2	0.479	0.429	0.361	0.264	0.114
3	0.479	0.429	0.361	0.264	0.114
4	0.317	0.264	0.197	0.114	0.025
5	0.318	0.264	0.197	0.114	0.025
6	0.221	0.171	0.114	0.053	0.006
7	0.224	0.165	0.113	0.053	0.006
8	0.166	0.108	0.065	0.024	0.001

Table 4.5. Comparison of the mean p-values returned by the Globaltest for the data generated from the multivariate normal distribution as in Figure 4.5 or Figure 4.6. The simulated data are generated as in section 4.4.2, with $n = 30$ and $m = n.DE = 2$; the mean p-value is calculated from 500 replications of the simulation

Effect Δ	Correlation r				
	0	Model in Figure 4.5		Model in Figure 4.6	
		0.4	0.8	0.4	0.8
0.5	0.2005	0.2317	0.2469	0.1514	0.1258
1.0	0.01487	0.02944	0.04617	0.008013	0.004926
1.5	0.0001711	0.001768	0.003039	4.627e−05	2.785e−05

It is clearly noticeable that in the previous model the classes become virtually indistinguishable due to their densities merging together, while in the current model the densities are perfectly separable.

In order to demonstrate performance of the Globaltest under these two models, we generated gene expression data as in the power related study (section 4.4.2), with $n = 30$ and $m = n.DE = 2$, and recorded the mean p-value of the Globaltest over 500 replications of the experiment. Results are given in Table 4.5. This study confirms that correlation in data can lead to higher p-values of the Globaltest (and other self-contained methods, as shown in section 4.4.2), i.e. to lower power of these methods. However, correlation in data can also lead to higher power (slightly smaller p-values), as shown in the two last columns in Table 4.5. This should be attributed to better separability of the classes which are compared by self-contained methods of gene set analysis.

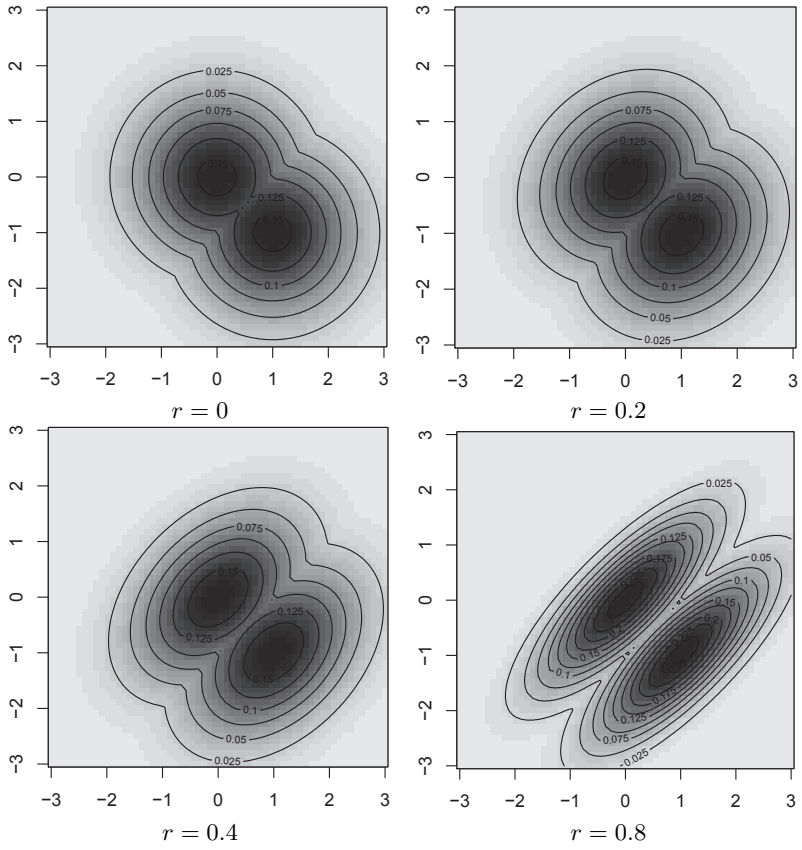

Fig. 4.6. Contour plots of the densities of the random variables $\mathcal{X}_0$, $\mathcal{X}_1$ (Equation (4.16)) as a function of correlation of features r. The density of $\mathcal{X}_0$ is centered at $\mu_0 = [0, 0]$, and the density of $\mathcal{X}_1$ is centered at $\mu_1 = [1, -1]$

However, it should be noted that the latter effect is harder to demonstrate for higher dimensional data (i.e. for larger numbers of correlated genes in the gene sets $n.DE$). For instance, for 20-dimensional, correlated data (i.e. for simulated data with $m = n.DE = 20$), we observe the mean p-values as shown in Table 4.6. We again observe that comparing to uncorrelated data, correlation between genes always leads to increased p-values of the Globaltest, i.e. to reduced power, although the strength of this effect strongly depends on whether the densities of the classes compared were shifted along the direction of high variability in correlated data (i.e. along the direction of the first principal component calculated from the data), or along some other line presumably perpendicular to the this direction.

Table 4.6. Comparison of the mean p-values returned by the Globaltest for the data generated from the 20-dimensional multivariate normal distributions with means given by Equations (4.17) and (4.19). The simulated data are generated as in section 4.4.2, with $n = 30$ and $m = n.DE = 20$; the mean p-value is calculated from 500 replications of the simulation

Effect Δ	Correlation r				
	0	Model: Equation (4.17)		Model: Equation (4.19)	
		0.4	0.8	0.4	0.8
0.5	0.004826	0.1476	0.2507	0.03691	0.08884
1.0	1.205e−09	0.004744	0.03854	4.371e−05	0.001944
1.5	9.23e−15	4.55e−05	0.002059	2.902e−10	5.219e−06

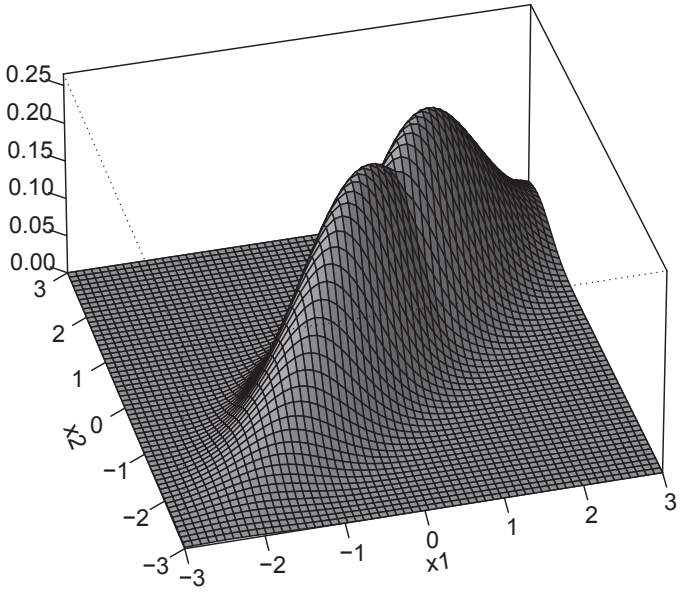

Fig. 4.7. 3D representation of the densities shown in Figure 4.5, for $r = 0.8$

The former scenario is represented by the two middle columns in Table 4.6 (we observe strong increase in the p-values), and the latter – by the two last columns (where we observe moderate increase in the p-values).

This effect defies simple explanations and intuitions proposed for low dimensional data and illustrated in Figures 4.5 through 4.8. One of the obstacles in applying these models for higher-dimensional data, and consequently in demonstrating the effect shown in Table 4.5 for higher-dimensional data, seems to result from the fact that the number of samples available in typical high throughput

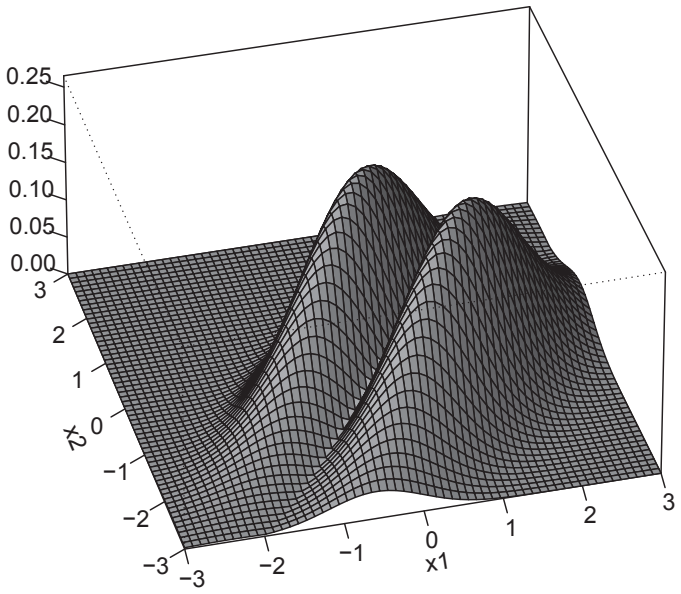

Fig. 4.8. 3D representation of the densities shown in Figure 4.6, for $r = 0.8$

studies (e.g., $n \sim 10^2$ in gene expression data) becomes prohibitively small to observe density related effects in high-dimensional data. This was analyzed by Scott and Thompson (1983) and named by these authors as the *empty space phenomenon*; this effect can be regarded as one of the symptoms of the *curse of dimensionality*, well known in machine learning literature (Theodoridis and Koutroumbas, 2006; Hastie *et al.*, 2001).

Chapter 5

Predictive modelling based on activation of feature sets

In the previous chapter, we discussed different methods of gene set enrichment analysis. The methods allow us to assess which of the sets of genes (features) given by *a priori* domain knowledge are activated, i.e. are most strongly associated with the target variable. As such, the methods focus on the activation of the entire gene sets. However, in order to use the the knowledge about pathway activation for classification of samples, we need to assess gene set activation scores per individual samples. We propose different approaches to estimation of these per-sample enrichment *signatures*, and discuss how these signatures can be related to the target in order to build domain knowledge-based predictive models.

In the second part of this chapter, we propose the algorithm for classification of samples based on the per-sample signatures of gene set activation, i.e. based on features generated using prior domain knowledge. The algorithm allows us to estimate the generalization error of the classifier as well as stability of feature selection. We propose several measures to quantify stability of feature selection.

We provide an overview of the recent developments in the learning theory which relate stability measures to predictivity conditions of classifiers. This theory was an inspiration of the stability measures proposed in this chapter.

In Chapter 6 we will empirically compare the proposed classification procedure will the standard approach based on data-driven feature selection. For the purpose of this analysis, we describe the algorithm for classification in $n \ll d$ data which is based on standard methods of feature selection.

Numerous papers have been devoted to the problem of estimation of gene expression signatures, or markers, for prediction of such targets as subtypes of cancers, expected response to therapies or risk of recurrence of a cancer. For instance, West *et al.* (2001) proposed to construct gene expression signatures by preselecting 100 top most differentially expressed genes and calculating the first singular value decomposition (SVD) factor from expression of these genes. This

factor was then used as a "supergene" to predict the phenotype (e.g. clinical status of breast cancer), with the prediction based on a logistic regression model. Bild *et al.* (2005) demonstrated methodology to construct signalling pathway related signatures. They deregulated a pathway using adenoviruses and identified sets of genes whose expression realized highest correlation with pathway-related effects. Then they represented these genes by a "metagene" calculated as the first principal component and showed empirically that this metagene can be used as a pathway related signature to predict the status or phenotype in various lung, ovarian or breast cancers.

Edelman *et al.* (2006) proposed a method to calculate *enrichment score* for each sample based on genes from an a priori specified gene set. They estimated association of all genes in the sample with the (binary) target using the class membership likelihood ratio and ranked all the genes in the sample by the this measure of association. Then they defined the *enrichment score of a gene set in a sample* as the Kolmogorov–Smirnov statistic comparing ranks of the genes in the gene set with the uniform distribution. This idea was clearly inspired by the enrichment score ES proposed the the GSEA gene set analysis method (Subramanian *et al.*, 2005), compare the GSEA statistic, point 1 on page 50. However, it should be noted that the original enrichment score of the GSEA is defined to measure enrichment (activation) of the entire gene set, while the enrichment score proposed by Edelman *et al.* (2006) refers to the activation status of the gene set in individual samples.

In the papers (Maciejewski, 2011a,b), we evaluated quality of the features selected as the sets of genes in top most activated pathways. We compared these features with the features selected with purely data driven (or "standard") methods. The comparison was done using data from real gene expression studies and involved both predictive performance and stability of the feature selection. In the analysis we focused on selected self-contained and competitive methods of gene set analysis. In the paper (Maciejewski, 2012), we proposed to aggregate the genes in the top winning pathways into "supergenes", where an activated pathway contributes two supergenes per the sample j:

$$\begin{aligned} v_j^+ &= \frac{1}{m}\sum_{i=1}^{m} I(t_i \geq 0) x_{ij} \\ v_j^- &= \frac{1}{m}\sum_{i=1}^{m} I(t_i < 0) x_{ij} \end{aligned} \tag{5.1}$$

where x_{ij}, $i = 1, ..., m$, is expression of gene i in the gene set concerned, m is the size of the gene set, and $I(t_i \geq 0)$ or $I(t_i < 0)$ means that the gene is up- or down-regulated, respectively (see notation introduced in section 4.2, and the definition of the GSA statistic, point 3 on page 51). This idea was inspired by the suggestion made in the original report which defined the GSA method (Efron and Tibshirani, 2006). This approach was evaluated in terms of predictive performance and stability of feature selection based on real data from gene expression studies.

In this section we want to extend these ideas and propose several approaches to deriving signalling pathway-based signatures which reflect activation of the pathways in individual samples. We will calculate the signatures for the pathways which, given results of high throughput study, prove significantly associated with the target, as indicated by the gene set analysis methods based on models 1 or 4 (as only these methods provide meaningful measures of association, see section 4.3). In this way, we will employ prior domain knowledge in the procedure of feature selection. Next, we will discuss how these pathway-based signatures can be employed for classification of samples and we will provide the algorithm to evaluate classifiers obtained in terms of the expected predictive performance for new data and in terms of stability.

In the next chapter we will empirically compare the proposed methods with data-driven methods (discussed in Chapter 2). The comparisons will be done using simulated data in order to ensure controlled characteristics of data in terms of correlation of features and strength of the inter-class differences.

Numerous ideas discussed in this chapter were published in (Maciejewski, 2008a,b; Maciejewski and Twaróg, 2009; Maciejewski, 2011a,b, 2012).

5.1. Classification based on signatures of gene set activation in individual samples

Let us assume that a given, *a priori* defined gene set is declared as activated (enriched) by a gene set analysis method based on results of a high throughput study. We want to assess activation scores of this gene set per individual samples. The idea is that the scores, or "signatures", should indicate to what extend the gene set is activated in a particular sample, which, in turn, could be used to predict the target for this sample.

We will develop these signatures and the procedure to classify samples based on the signatures, using the following notation. We denote results of a high throughput study as $W_{d \times n}$, where the columns represent samples and rows represent

features (e.g. expression of genes). We denote the class labels of the samples as $Y_{1\times n}$; we consider here the binary classification problem: $Y_i \in \{0, 1\}$, $i = 1, ..., n$. We represent an *a priori* defined gene set as the set of row indices of W, and denote this set as S. Let $X = (x_{ij})$, $i = 1, ..., m$, $j = 1, ..., n$ represent the subset of rows of W corresponding to S, i.e. $X = (W_{i\bullet})$ for $i \in S$, where m is the number of elements in S.

Let $x_{i\bullet}^{(0)} = (x_{i,1}^{(0)}, ..., x_{i,n_0}^{(0)}) = (x_{ij} : Y_j = 0)$ and $x_{i\bullet}^{(1)} = (x_{i,1}^{(1)}, ..., x_{i,n_1}^{(1)}) = (x_{ij} : Y_j = 1)$ denote expressions of gene i in class 0 and in class 1, respectively. We regard $x_{i\bullet}^{(0)}$ and $x_{i\bullet}^{(1)}$ as n_0 and n_1-element samples from some underlying random variables, denote here as $\mathcal{X}_i^{(0)}$ and $\mathcal{X}_i^{(1)}$, respectively. We denote probability density functions of these distributions as $f_i^{(0)}$ and $f_i^{(1)}$, and cumulative distribution functions as $F_i^{(0)}$ and $F_i^{(1)}$, respectively.

Assuming that the gene set S is significantly associated with the target Y, we want to use the data (X, Y) to classify a sample $u = (u_1, ..., u_m)$ whose coordinates represent expressions of the members of the gene set S.

5.1.1. Method 1

We assume that the probability densities $f_i^{(0)}$ and $f_i^{(1)}$, for $i = 1, ..., m$ can be estimated with the densities of the normal distribution:

$$\begin{aligned} f_i^{(0)}(x) &= \frac{1}{\sigma_i^{(0)}\sqrt{2\pi}} \exp\left(-\frac{(x-\mu_i^{(0)})^2}{2(\sigma_i^{(0)})^2}\right) \\ f_i^{(1)}(x) &= \frac{1}{\sigma_i^{(1)}\sqrt{2\pi}} \exp\left(-\frac{(x-\mu_i^{(1)})^2}{2(\sigma_i^{(1)})^2}\right) \end{aligned} \tag{5.2}$$

where the parameters $(\mu_i^{(0)}, \sigma_i^{(0)})$ and $(\mu_i^{(1)}, \sigma_i^{(1)})$ are calculated as the mean and standard deviation of $x_{i\bullet}^{(0)}$ and $x_{i\bullet}^{(1)}$, respectively.

Given a new sample $u = (u_1, ..., u_m)$ we can calculate two signatures which indicate association of the sample with the profile of gene expressions characteristic of the class 0 and class 1:

$$\begin{aligned} s^{(0)} &= \sum_{i=1}^{m} \log\left(f_i^{(0)}(u_i)\right) \\ s^{(1)} &= \sum_{i=1}^{m} \log\left(f_i^{(1)}(u_i)\right) \end{aligned} \tag{5.3}$$

Note that higher value of $f_i^{(0)}(u_i)$ as compared with $f_i^{(1)}(u_i)$ indicates that expression of the gene i in sample u is more similar to the profile of expression characteristic of class 0, and vice-versa. Hence $s^{(0)}$ and $s^{(1)}$ can be regarded as the overall measures of similarity of u with regard to expression profile of the gene set in class 0 and class 1, respectively. We use the log transformation in formula (5.3) to improve numerical stability of the calculations.

Given the signatures $s^{(0)}$ and $s^{(1)}$, we can readily classify the sample u as:

$$g(u) = \arg\max_{k=0,1} \left(s^{(k)} \right) \tag{5.4}$$

where g is the classification rule motivated by the naive Bayes classifier. The classifier assigns the sample u to the class which realizes bigger value $s^{(k)}$, i.e. to which u is more similar according to the similarity measures $s^{(0)}$ and $s^{(1)}$.

5.1.2. Method 2

In Method 1, we use the similarity measure between a sample $u = (u_1, ..., u_m)$ and the profile of gene expressions in class k based on the vectors $(f_1^{(k)}, ..., f_m^{(k)})$, for $k = 0, 1$, where the components of these vectors are the density functions of subsequent features in S, calculated at $(u_1, ..., u_m)$. As such, the vectors $(f_1^{(0)}, ..., f_m^{(0)})$ and $(f_1^{(1)}, ..., f_m^{(1)})$ can be compared (as done in Equation (5.4)), however, direct interpretation of the elements of the vectors in not straightforward.

Here we propose an alternative measure (signature) of similarity between $u = (u_1, ..., u_m)$ and the profiles of gene expression specific to class 0 and 1. The purpose is to ease interpretability of the components of the measure.

As in model 1, we assume that the distributions which generated samples $x_{i\bullet}^{(0)}$ and $x_{i\bullet}^{(1)}$, $i = 1, ..., m$, can be approximated by the normal probability distribution with the parameters $(\mu_i^{(0)}, \sigma_i^{(0)})$ and $(\mu_i^{(1)}, \sigma_i^{(1)})$, where the parameters are calculated as the mean and standard deviation of $x_{i\bullet}^{(0)}$ and $x_{i\bullet}^{(1)}$, respectively. The cumulative distribution functions of these distribution are:

$$\begin{aligned} F_i^{(0)}(x) &= \Phi_{\mu_i^{(0)}, \sigma_i^{(0)}}(x) \\ F_i^{(1)}(x) &= \Phi_{\mu_i^{(1)}, \sigma_i^{(1)}}(x) \end{aligned} \tag{5.5}$$

where $\Phi_{\mu,\sigma}$ denotes the cumulative distribution function of the normal distribution. We define the measure of similarity between μ_i, i.e. expression of gene i in the sample concerned and the profiles of expression of this gene in classes 0 and 1 as:

$$\begin{aligned} p_i^{(0)} &= \min\left(F_i^{(0)}(u_i), 1 - F_i^{(0)}(u_i)\right) \\ p_i^{(1)} &= \min\left(F_i^{(1)}(u_i), 1 - F_i^{(1)}(u_i)\right) \end{aligned} \tag{5.6}$$

Note that the values $p_i^{(0)}$ and $p_i^{(1)}$ express the probabilities of drawing the observed value of expression u_i, *or a more extreme value*, from the distribution characteristic of the given class 0 or 1. Obviously, $p_i^{(0)}, p_i^{(1)} \in (0, 0.5]$, where small values of $p_i^{(k)}$ indicate that it is unlikely that u_i comes from the distribution $f_i^{(k)}$. The meaning of $p_i^{(0)}$, $p_i^{(1)}$ is illustrated in Figure 5.1, where

$$\begin{aligned} p_i^{(1)} &= \int_{-\infty}^{u_i} f_i^{(1)}(x)dx = F_i^{(1)}(u_i) = Pr\{\mathcal{X}_i^{(1)} < u_i\} \\ p_i^{(0)} &= \int_{u_i}^{\infty} f_i^{(0)}(x)dx = 1 - F_i^{(0)}(u_i) = Pr\{\mathcal{X}_i^{(0)} \geq u_i\} \end{aligned}$$

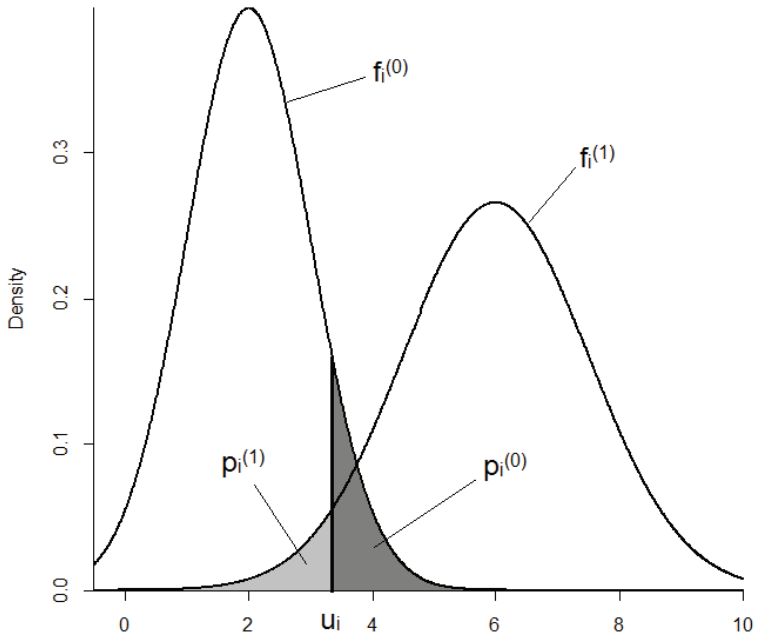

Fig. 5.1. Illustration of the measures of association proposed in Equation (5.6), where u_i is the value of expression of gene i and the probabilities $p_i^{(0)}$ and $p_i^{(1)}$ indicate how extreme u_i is with regard to profiles of expression of this gene in class 0 ($f_i^{(0)}$) and in class 1 ($f_i^{(1)}$)

Now given a sample $u = (u_1, ..., u_m)$, we can use the vectors $(p_1^{(0)}, ..., p_m^{(0)})$ and $(p_1^{(1)}, ..., p_m^{(1)})$ as signatures which indicate association of u with the profile of gene set expressions characteristic of class 0 or 1, respectively. By aggregating these vectors as in Method 1:

$$\begin{aligned} s2^{(0)} &= \sum_{i=1}^{m} p_i^{(0)}(u_i) \\ s2^{(1)} &= \sum_{i=1}^{m} p_i^{(1)}(u_i) \end{aligned} \tag{5.7}$$

we obtain measures that can be used to classify the sample u as:

$$g2(u) = \arg\max_{k=0,1} \left(s2^{(k)} \right) \tag{5.8}$$

The classification rule $g2$ assigns the sample to the class which realizes bigger similarity measure $s2^{(0)}$, $s2^{(1)}$.

5.1.3. Method 3

Method 3 is similar to Method 2, with different method of modelling the densities $f_i^{(0)}$ and $f_i^{(1)}$: instead of using normal probability distribution (Equation (5.2)) we will use nonparametric kernel density estimates. Technically, we use the estimates with the Gaussian kernels, i.e.

$$\begin{aligned} f_i^{(0)}(x) &= \frac{1}{n_0 h} \sum_{j=1}^{n_0} \varphi\left(\frac{x - x_{ij}^{(0)}}{h} \right) \\ f_i^{(1)}(x) &= \frac{1}{n_1 h} \sum_{j=1}^{n_1} \varphi\left(\frac{x - x_{ij}^{(1)}}{h} \right) \end{aligned} \tag{5.9}$$

where φ denotes the probability density function of the standard normal distribution, and the smoothing parameter (bandwidth) h is calculated using the heuristic rule proposed by Silverman (1986), i.e., $h \sim$ minimum of the standard deviation and the interquartile range, and inversely proportionate to the sample size raised to $-\frac{1}{5}$ power, Equation 3.31 in (Silverman, 1986).

Since nonparametric density estimation has been extensively studied in literature, clearly many other choices of the kernel φ and the bandwidth h are available – see e.g. (Devroye *et al.*, 1996) for a comprehensive overview of this vast topic.

Given the densities $f_i^{(0)}$ and $f_i^{(1)}$, we assess how extreme expression u_i of gene i is with regard to the profile of expression of this gene in class 0 and 1, i.e.

$$\begin{aligned} p_i^{(0)} &= \min\left(c_i^{(0)}, 1 - c_i^{(0)}\right) \\ p_i^{(1)} &= \min\left(c_i^{(1)}, 1 - c_i^{(1)}\right) \end{aligned} \tag{5.10}$$

where $c_i^{(k)} = \int_{-\infty}^{u_i} f_i^{(k)}(x)dx = Pr\{\mathcal{X}_i^{(k)} < u_i\}$, $k = 0, 1$.

The sample can be now classified based on the signature vectors $(p_1^{(0)}, ..., p_m^{(0)})$ and $(p_1^{(1)}, ..., p_m^{(1)})$, with the classification procedure as in Method 2:

$$g3(u) = \arg\max_{k=0,1}\left(s3^{(k)}\right) \tag{5.11}$$

where the aggregated signatures $s3^{(0)}$ and $s3^{(1)}$ are calculated as:

$$\begin{aligned} s3^{(0)} &= \sum_{i=1}^{m} p_i^{(0)}(u_i) \\ s3^{(1)} &= \sum_{i=1}^{m} p_i^{(1)}(u_i) \end{aligned} \tag{5.12}$$

The classification rule $g3$ assigns the sample to the class to which the sample is closer in terms of the similarity measure $s3^{(0)}$, $s3^{(1)}$.

5.1.4. Comment on assumptions of methods 1–3 and on alternative parametric approach

The methods presented in sections 5.1.1–5.1.3 were developed for the binary classification problem $Y_i \in \{0, 1\}$, $i = 1, ..., n$. We note however that the methods can be extended to encompass multi-class problems with $Y_i \in C = \{c_1, c_2, ..., c_k\}$. To do this, we need to estimate expression profiles of a gene i characteristic of the k classes $f_i^{c_1}, f_i^{c_2}, ..., f_i^{c_k}$, as done in Equations (5.2), (5.5) or (5.9) for the binary case. Based on the profiles, we calculate the aggregated signatures which represent association of a tested sample u with the k classes $s^{(c_1)}$, $s^{(c_2)}$, ..., $s^{(c_k)}$, as done in Equations (5.3), (5.7) or (5.12). Then extension to the multi-class case is straightforward:

$$g(u) = \arg\max_{c \in \{c_1, c_2, ..., c_k\}}\left(s^{(c)}\right)$$

Although analysis of multi-class experiment is technically feasible, the binary classification problem is definitely more important when considering class prediction from real high throughput data such as gene expression results. The number of samples commonly tested in such studies, n up to several tens, is usually deemed too small to provide enough training data for multi-class problems, so typical experiment organization assumes only two groups.

It should be noted that the classification procedures proposed here resemble the memory based learning (MBL) classifiers in that we do not search through a space of parametric models $\mathcal{H}$ in order to fit the model $f \in \mathcal{H}$ which is expected to minimize the expected prediction error. Instead, we use the training data to build the profiles of subsequent features in the (two) groups of samples compared, and then we classify a new sample by directly comparing the sample with these profiles.

Later we will compare this approach with an alternative method where we first perform feature selection based on pathway activation and then we fit a parametric model from a given hypothesis space $\mathcal{H}$ in the low-dimensional data spanned by the members of the winning pathways. This latter approach was analyzed in (Maciejewski, 2011a,b), where we compared different families of models (e.g. Support Vector Machines, logistic regression, random forests, as well as nonparametric models) built with the genes from the most activated pathway directly taken as features. We showed then that this approach may outperform classifiers based on features selected using data-driven methods. This analysis will be extended in Chapter 6.

In the next section, we discuss the problem of generalization of classifiers built from high throughput data, and in section 5.3 we present a generic framework where either MBL-like methods (such as the ones proposed in sections 5.1.1 through 5.1.3), or parametric classifiers (e.g. SVMs) can be employed for class prediction from high throughput data.

5.2. Assessment of predictivity in classification based on high dimensional data

We showed previously (section 2.1) that the major challenge in building predictive models from high dimensional data (such as data from high-throughput studies with $d \gg n$) is related to *overfitting*, which refers to the fact that models which accurately fit to the training data are likely to demonstrate poor performance given new independent data. In this section, we want to elaborate on this effect in

order to (i) understand conditions under which overfitting can be controlled and (ii) find methods to provide reliable assessment of classifier performance assuming that the training data is sparse in terms of the number of samples available. We discuss both classical results from the learning theory which are focused on the properties of the hypotheses space, as well as recent results which are focused on the properties of the learning algorithms rather than the hypothesis space. These new results are interesting as they develop stability measures of the learning algorithm and show relationship between these measures and generalization property of classifiers.

Finally, we consider data-reuse methods employed for model selection and estimation of generalization measures.

In predictive modelling we want to find the functional relationship between the feature vectors (x, inputs) and targets (y, outputs) given the training data $S = \{(x_1, y_1), ..., (x_n, y_n)\}$. We assume that inputs and outputs are independent samples from the underlying random variables $X \in R^d$ and Y, respectively, and that (X, Y) are governed by an unknown joint probability distribution denoted $\mu(X, Y)$.

A learning algorithm is a map which assigns to the training data S a function $f \in \mathcal{H}$, where $\mathcal{H}$ is the hypotheses space comprising the space of functional relationships f between X and Y that we assume. To measure the (in)accuracy of prediction of Y using $f(X)$, we define the loss function $L(Y, f(X))$, which takes 0 for $Y = f(X)$ and some positive values otherwise which represent punishment for misprediction.

The most important quality measure of the fitted model f is the *expected prediction error* for new data, known also as the *generalization error*, defined as the mean error (loss) in classification calculated with regard to the distribution of (X, Y) (Hastie *et al.*, 2001):

$$EPE(f) = E(L(Y, f(X))) = \int L(y, f(x))d\mu(x, y) \tag{5.13}$$

Obviously, since μ is unknown, in practice EPE cannot be calculated. However, we can calculate the *empirical error* as the average error observed for the training samples:

$$E_{emp}(f) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i)) \tag{5.14}$$

5.2.1. Empirical assessment of predictivity

Reliable estimation of the EPE is one of the major challenges in predictive modelling unless we deal with a "data rich" scenarios where the number of samples n is sufficiently large. Then the EPE can be estimated empirically based on the partition of the training data not used for selection of the model. More specifically, it is typically recommended to randomly split the training data into 3 partitions (Bishop, 1995; Hastie *et al.*, 2001): *train*, *validation* and *test*, where the train and validation partitions are used to fit a model and estimate its generalization error, respectively, for a series of models searched by the learning algorithm. The last step is to use the test partition to obtain the unbiased estimate of the EPE for the final model selected in the previous step (i.e. for the model which realizes the smallest prediction error observed for the validation partition).

In should be noted that if we simplify this procedure and use only two partitions: train and test for model fitting and selection, and fail to obtain the final unbiased estimate of the EPE using an independent (third) partition, then we risk an over-optimistic estimate of the EPE.

5.2.2. Predictivity conditions based on the learning theory

For the cases when n is not large enough to use the above procedure, we can refer to results from the statistical learning theory which provide conditions for generalization of the predictive models. The idea is to impose some constraints on the hypothesis space $\mathcal{H}$ (i.e. excluding some "weird" functions from $\mathcal{H}$) so that it can be guaranteed that small empirical error implies small generalization error (Vapnik and Chervonenkis, 1991; Vapnik, 1999; Poggio *et al.*, 2004). Technically, the learning theory develops such constrains for the class of *empirical risk minimization* (ERM) algorithms, i.e. the algorithms which fit a function $\hat{f}$ such that:

$$E_{emp}(\hat{f}) = \min_{f \in \mathcal{H}} E_{emp}(f) \tag{5.15}$$

assuming that the minimum exists (although results also hold if $E_{emp}(\hat{f}) = \inf_{f \in \mathcal{H}} E_{emp}(f)$).

The key result states that the ERM algorithm (i) *generalizes*, and (ii) *is consistent* if and only if the hypothesis space $\mathcal{H}$ is the uniformly Glivenko–Cantelli (uGC) class of functions.

Generalization means that the function $\hat{f}$ returned by the algorithm realizes the empirical error which converges to the generalization error:

$$E_{emp}(\hat{f}) \xrightarrow{P} EPE(\hat{f}) \tag{5.16}$$

where the convergence is in probability as the sample size n increases, i.e. $\forall \epsilon > 0 \lim_{n\to\infty} Pr\{|E_{emp}(\hat{f}) - EPE(\hat{f})| > \epsilon\} = 0$.

Consistency means that the generalization error for the function $\hat{f}$ returned by the algorithm is asymptotically close to the smallest generalization error that can be achieved in $\mathcal{H}$, i.e.

$$\forall \epsilon > 0 \lim_{n\to\infty} Pr\left\{EPE(\hat{f}) > \inf_{f\in\mathcal{H}} EPE(f) + \epsilon\right\} = 0 \tag{5.17}$$

The constraint imposed on the family of functions $f \in \mathcal{H}$ which guarantees generalization and consistency of the ERM algorithms (i.e. $\mathcal{H}$ being the uGC class) means that, loosely, for any distribution μ, $X \sim \mu$, the expected value $E(f(X))$ can be approximated by the average value of f calculated over a sufficiently large sample from X, i.e. (Poggio *et al.*, 2004):

$$\forall \epsilon > 0 \lim_{n\to\infty} Pr\left\{\sup_{f\in\mathcal{H}} \left|\frac{1}{n}\sum_{i=1}^{n} f(x_i) - \int_X f(x)d\mu(x)\right| > \epsilon\right\} = 0 \tag{5.18}$$

Note that for binary f this condition is equivalent to the requirement that the family has the finite Vapnik–Chervonenkis (VC) dimension, the well-known condition for predictivity in classification.

It should be noted that this classical theory applies only to ERM learning, i.e. to the algorithms which minimize the empirical error over a fixed hypothesis space. Therefore it cannot be used for non-ERM algorithms such as support vector machines, regularization based learning or k-nearest neighbours classifiers, etc. (Poggio *et al.*, 2004).

For this reason, the generalized theory of learning was recently developed which applies to ERM and non-ERM algorithms (Mukherjee *et al.*, 2002; Poggio and Smale, 2003; Poggio *et al.*, 2004; Mukherjee *et al.*, 2006; Wibisono *et al.*, 2009). The key difference is that the classical theory assumes ERM learning and imposes constraints on the hypotheses space to ensure generalization, while the generalized theory characterizes properties of a learning algorithm (not necessarily ERM) in terms of *stability* which are necessary and sufficient for generalization. Technically, the theory defines the following stability measures of the learning algorithm (i.e. the map from the learning sets to $\mathcal{H}$) (Mukherjee *et al.*, 2002; Poggio *et al.*, 2004):

- Cross-validation leave-one-out stability ($\mathrm{CV_{loo}}$): the learning map is distribution independent, $\mathrm{CV_{loo}}$-stable if

$$\lim_{n\to\infty} \sup_{1\leq i\leq n} |L(y_i, f_S(x_i)) - L(y_i, f_{S^i}(x_i))| = 0 \tag{5.19}$$

  in probability, for all distributions μ, where f_S denotes the function fitted by the algorithm using the complete training set S, and f_{S^i} denotes the function fitted using S with the sample (x_i, y_i) removed.
- Cross-validation leave-one-out stability of the empirical and expected error ($\mathrm{CVEEE_{loo}}$): the learning map is $\mathrm{CVEEE_{loo}}$-stable if
  - is $\mathrm{CV_{loo}}$-stable,
  - realizes stability of the expected error:

$$\lim_{n\to\infty} \sup_{1\leq i\leq n} |EPE(f_S) - EPE(f_{S^i})| = 0 \quad \text{in probability,} \tag{5.20}$$

  - realizes stability of the empirical error:

$$\lim_{n\to\infty} \sup_{1\leq i\leq n} |E_{emp}(f_S) - E_{emp}(f_{S^i})| = 0 \quad \text{in probability.} \tag{5.21}$$

The key results of the generalized theory are that (i) for the ERM learning algorithms, the $\mathrm{CV_{loo}}$ stability is the necessary and sufficient condition for generalization and consistency, and (ii) for any learning algorithm, $\mathrm{CVEEE_{loo}}$ stability is the sufficient condition for generalization (but not for consistency).

It should be noted that the theoretic results guarantee asymptotic, as the samples size $n \to \infty$, conformity of the empirical error (which we can measure) with the expected error (which we want to estimate). Additionally, the theory also provides the estimates of the over-optimism of the training error, i.e. of $EPE - E_{emp}$ for finite sample sizes n (Mukherjee *et al.*, 2002; Poggio *et al.*, 2004). However, practical application of these estimates for the purpose of assessment of the EPE is difficult due to (i) not very fast rate of convergence of E_{emp} to EPE typically observed (Poggio *et al.*, 2004; Mukherjee *et al.*, 2006), which is a limitation considering small values of n commonly available in high-throughput studies, and (ii) the difficulty in calculating the theoretical bounds on generalization of the models (Hastie *et al.*, 2001).

However, results of the general theory of learning are important as they introduce stability as an important criterion for generalization of predictive models. The specific stability criteria, based on the cross-validation leave-one-out procedure (Equations (5.19)–(5.21)) motivated the measures of stability of feature selection that we later introduce and investigate in sections 5.3 and 6.

Finally, we want to mention the classic results from the learning theory pertaining to the nonparametric classification (and regression) procedures, such as the nearest neighbours or the procedures based on kernel estimates of the class densities. These results provide conditions for Bayes risk consistency of these nonparametric procedures, i.e. for the convergence (as the sample size $n \to \infty$) of the error rate of these procedures to the Bayes error (Greblicki, 1978; Greblicki and Pawlak, 1987). Interestingly, Bayes risk consistency can also be ensured even if some of the training observations have mislabelled class designations, as shown e.g. by Greblicki (1980).

5.2.3. Data reuse methods for assessment of predictivity

An alternative approach to the assessment of predictivity when dealing with small sample sizes is to employ data-reuse methods such as bootstrap or cross-validation (Hastie *et al.*, 2001). The idea is to directly estimate the generalization error EPE by repeatedly, randomly generating a training subsets from the available n samples and evaluating predictive performance of the models fitted to the training subsets. Then the EPE is estimated as the average of the prediction errors observed over the fitted models.

More specifically, in the bootstrap procedure, B datasets each of size n are randomly selected with replacement from the original dataset of n samples. Then B models are fitted based on the generated datasets and are evaluated on the original dataset. The final estimate of the EPE is obtained by combining the B estimates of the prediction error.

In cross-validation, the available n samples are split into K partitions of roughly similar size. Then in K steps of the cross-validation procedure, each of the partitions is once used as the test partitions with the remaining partitions used for fitting the model. In this way, we get K estimates of the prediction error for subsequent models, which we can average to obtain the final estimate of the EPE. In the *leave-one-out* cross-validation, i.e. for $K = n$, we obtain approximately unbiased estimate of the EPE, however variance of the estimate tends to be higher. By contrast, in the *K-fold* cross-validation (with K typically taken between 5 and 10), the estimate of the EPE will show higher bias but smaller variance (Hastie *et al.*, 2001).

Obviously, the leave-one-out cross-validation is computationally more expensive, however it allows to calculate different measures of stability of the learning process, as shown in section 5.2.2. For this reason, the algorithm for class predic-

tion that we present in the next section employs the leave-one-out cross-validation which is the basis of several measures of feature selection stability that we propose.

When dealing with high dimensional data (with $d \gg n$), where dimensionality reduction/feature selection plays the key role, it is essential that *internal* rather than *external* cross-validation is used (Allison *et al.*, 2006; Markowetz and Spang, 2005). The difference lies in where the feature selection step is performed: in the external procedure, feature selection is done once prior to the cross-validation loop, while in the internal procedure each step of cross-validation includes feature selection as the part of model fitting. This distinction is ignored by some authors who tend to choose computationally simpler, external cross-validation for assessment of generalization in class prediction based on high dimensional data. However, this can lead to surprisingly high over-optimism in estimation of the generalization error, as shown e.g. in (Simon *et al.*, 2003; Simon, 2003).

For this reason, we employ internal cross-validation in the algorithms presented in sections 5.3 and 5.5. This, additionally, allows us to estimate several measures of leave-one-out stability of the feature selection, similar to the stability measures discussed in section 5.2.2.

5.3. Algorithm of sample classification based on prior domain knowledge

We provide a generic procedure which allows us to incorporate prior domain knowledge about possible associations between features into the process of building a predictive model based on results of a high throughput study. The procedure returns (i) the predictive model, (ii) estimates of the expected prediction error when using the model for new, independent data, (iii) the set of features which are actually used in prediction, (iv) measures of stability of the selected feature set under small changes of data.

We assume that results of a massive throughput study, such as gene expression microarray assay, are given as (W, Y) where $W_{d\times n}$ represents the matrix of expressions of d genes (features) measured from n samples, and $Y_{1\times n}$ represents the class labels of the samples, $Y_i \in C = \{c_1, ..., c_k\}$, $i = 1, ..., n$. Although we restrict this algorithm to the classification problem (rather than regression problem, with quantitative Y), we do not require that Y is binary, except when explicitly assumed. Note however, that in vast majority of practical problems in bioinformatics for which the method discussed here can be employed (such as class prediction studies based on gene expression data), it is commonly assumed that

targets are binary ($Y_i \in \{0, 1\}$, which represents e.g. disease vs. control, response no response to therapy, recurrence vs. no recurrence of a disease, etc.).

We also assume that the prior domain knowledge about association between features is given in the form of a database of gene sets $\mathbf{S} = \{S_1, S_2, \ldots, S_M\}$. The sets may include groups of related genes, such as the members of signaling pathways, or groups of genes with common gene ontology terms. For clarity of presentation of the algorithm, we assume that the gene sets $S_1, \ldots, S_M$ are specified as sets of indices of rows in W corresponding to the genes in subsequent gene sets. See Remark 5 (page 103) for explanation of some technical issues related to this assumption.

The class prediction algorithm takes as input:

1. (W, Y),
2. $\mathbf{S}$,

and produces on output:

1. The classifier $g : R^r \to C$, which assigns the class label to a (new) sample based on expression of r selected genes,
2. The subset of r features used by the classifier g (which we represent by the set of their indices $S \subset \{1, ..., d\}$),
3. EPE – prediction error expected when new, independent data is classified with g,
4. Measures of stability of the feature selection, as defined in section 5.4.

The class prediction algorithm is realized in the following steps.

1. Calculate the gene set enrichment score s_i and the associated (multiple testing corrected) p-value p_i for each of the gene sets S_1, ..., S_M, using the gene set analysis method $\mathcal{G}$: $(s_i, p_i) = \mathcal{G}(S_i, W, Y)$, $i = 1, ..., M$.
2. Rank the gene sets $S_{(1)}$, ..., $S_{(M)}$ descending by the gene set enrichment score, so that $p_{(1)} \leq ... \leq p_{(M)}$, i.e. the gene sets most associated with the target are at the top of the list. If none of the gene sets is significantly associated with the target, i.e. if $p_{(1)} > 0.05$ then STOP with the information that the algorithm fails to build the classifier based on pathway activation. Otherwise find the number of sets associated with the target $p_a = \max\{i : p_{(i)} \leq 0.05\}$ and select the features in the top most activated gene sets as $S_0 = S_{(1)} \cup ... \cup S_{(\min(k, p_a))}$, where k is a fixed parameter of the method [see Remark 3]. Move to step 3.

3. Repeat steps 4 through 8 for $i = 1, ..., n$.
4. Leave out sample i for model testing, i.e. remove column $W_{\bullet i}$ from W and element Y_i from Y. We denote the remaining matrix and vector as W^i and Y^i.
5. Using the training data (W^i, Y^i) calculate the enrichment score and the associated p-value for each of the gene sets $S_1, ..., S_M$ as $(s_j, p_j) = \mathcal{G}(S_j, W^i, Y^i)$, $j = 1, ..., M$. Order the gene sets by increasing p-values: $S_{(1)}, \ldots, S_{(M)}$.
6. Select rows from W^i related to the set of features $S = S_{(1)} \cup ... \cup S_{(\min(k, p_a))}$, and remove the remaining rows. Denote the resulting matrix as $X^i = (W^i_{j\bullet} : j \in S)$.
7. Using the training data (X^i, Y^i), build the predictive model g and classify the sample Y_i as $\hat{Y}_i = g(Y_i)$.
8. In the list of counters $c_1, ..., c_M$, corresponding to the M gene sets in $\mathbf{S}$, increment the counters which correspond to the gene sets selected in step 6.
9. Calculate the expected misclassification rate as

$$EPE = \frac{1}{n} \sum_{i=1}^{n} I(\hat{Y}_i \neq Y_i) \tag{5.22}$$

10. Select the rows from W related to the genes in S_0 and remove the remaining rows to obtain the matrix $X = (W_{i\bullet} : i \in S_0)$. Based on the training data (X, Y) build the predictive model g.
11. Based on the list of counters $c_1, ..., c_M$, calculate the stability measures defined in section 5.4, i.e. $N.SEL$ (defined by formula (5.23)), Q_α for $\alpha = 0.9, 0.75, 0.5$ (formula (5.24)), $FREQ_1$, ..., $FREQ_k$ (formula (5.25)). Calculate the mean p-values associated with the winning gene sets $PVAL_1$, ..., $PVAL_k$ as defined by formula (5.26).
12. Return g, the subset of features S_0, EPE and the stability measures as calculated in steps 9–11.

Note that in the proposed procedure we do not provide specific methods to be used for gene set ranking ($\mathcal{G}$), or for sample classification (model g); the procedure also relies on the parameter k. In the following remarks we want to clarify how these generic setting should be fixed in specific realizations of the algorithm.

Remarks:

1. As the gene set ranking method $\mathcal{G}$ (steps 1 and 5 of the algorithm) we propose to use the gene set analysis procedures based on models 1 or 4 (see sections 4.3.1 and 4.3.4), as these methods produce meaningful p-values and stay in line with the organization of the biological experiment which produced the dataset W, as explained in sections 4.3.5 and 4.5. Examples of such methods are the Globaltest, GSA, GSA2 (point 4 on page 52).
2. Regarding the classification model g (steps 7 and 10), the obvious choice is to use parametric models such as the SVM, logistic regression etc. Alternatively, we can use nonparametric methods such as the procedures proposed in section 5.1 (Equations (5.4), (5.8), (5.11)). Strictly, the classifier g returned by the latter methods does not contain a discriminant function of the inputs used to predict the target, but it rather contains the complete training data (or its representation in the form of distributions of the features, e.g. Equations (5.2) or (5.9)), like in-memory-based reasoning methods. In Chapter 6 we empirically compare selected parametric and nonparametric approaches.
3. The parameter k (steps 2 and 6) represents the number of top most activated pathways whose genes are taken as features in classification. This parameter should be tuned empirically, i.e. the value of k should be selected which minimizes the EPE. Therefore, the algorithm should be run for $k = 1, 2, ...$ up to some reasonable value not exceeding the number of activated pathways (denoted as p_a – step 2 of the algorithm). This way of tuning the parameter k is similar to tuning the size of the feature set selected with the classical univariate filter methods (see section 2.2).
4. Note that the algorithm proposed realizes (internal) cross-validation to estimate the EPE and the proposed feature stability measures. As indicated in Remark 3, we can use the estimate of generalization error for fine-tuning the parameters of the model, such as k, and, similarly, other parameters specific to the model g. Note however, that this is equivalent to splitting the available data into two (train and test) rather than three (train, validation and test) partitions (see Section 5.2.1). The consequence of this simplification may be some bias in estimation of the generalization error. To reduce the bias, we may in principle follow the train-validate-test scheme in cross-validation, as shown in (Maciejewski and Twaróg, 2009; Lai *et al.*, 2006). The idea is to use nested cross-validation loops, where the purpose of the outer loop is to leave-out samples for model testing, and in the inner

loop we leave-out samples for model fine-tuning (these samples play the role of the validation partition). Although this reduces the bias in EPE, we chose not to use this approach as (i) it enormously increases computational burden, and (ii) we can still use the proposed procedure to *relatively* evaluate different approaches to classification based on high dimensional data (i.e. the approach based on purely data driven feature selection vs. the approach based on the prior domain knowledge).

5. The final remark is technical and deals with the possible discrepancy between the definition of gene sets given is KEGG of Biocarta databases and the set of genes (transcripts) that label rows of the dataset W. Note that the latter is specific to the microarray or RT-PCR technology used to generate the data, where not necessarily all the genes referred to in e.g. KEGG or GO gene sets are measured with a particular microarray device. Additionally, prior to analysis, raw microarray data commonly undergoes *non-specific filtering*, which consists in deleting rows (i.e. genes/transcripts) due to data quality problems. As quality assurance of high-throughput data (i.e. normalization, non-specific filtering) is outside of scope of this work, throughout this document we assume that the dataset concerned (denoted here as W) has already undergone necessary quality-related transformations.

   Yet another issue is related to the fact that microarray technologies typically represent some selected genes by the collection of different transcripts scattered on the array, i.e. one gene may be represented by several rows in the resulting dataset. For all these reasons, creating the representation of the KEGG gene sets in the form of the sets of indices of rows of the matrix W, as required by the algorithm and denoted $\mathbf{S} = \{S_1, S_2, \ldots, S_p\}$ on page 100, is not a direct one-to-one mapping. Therefore, the data set $S_i \in \mathbf{S}$, representing a given pathway, is constructed by taking the indices of rows of W which correspond to a gene in that pathway; if no such indices are found, then the pathway is not represented in $\mathbf{S}$.

5.4. Measures of stability of feature selection

In the algorithm presented in section 5.3, we repeatedly perform feature selection in subsequent steps of the leave-one-out cross-validation procedure. This allows us to define $\mathrm{CV}_{\mathrm{loo}}$ stability measures of features selection which characterize variability/(in)stability of features under small changes in the training data.

Taken together with the EPE, these measures are supposed to provide stronger evidence pertaining to the quality of the final model g built from high dimensional data and returned in step 12 of the algorithm.

Note that in each step of cross-validation, we select k most activated gene sets (step 2 and 6). If, in a given study, features are insensitive to small changes in data, then we would expect to observe roughly the same subset of features selected throughout the cross-validation loop. To quantify this, we define the first measure which determines how many distinct features were selected at least once throughout the procedure:

$$N.SEL = \frac{1}{k} \sum_{i=1}^{M} I(c_i \neq 0) \tag{5.23}$$

where c_1, ..., c_M are the counters maintained in step 8 of the algorithm and I is the indicator function. Note that the number of distinct features ever selected actually equals $\sum_{i=1}^{M} I(c_i \neq 0)$, which, for stable features, should be close to k. The purpose of the standardization term $\frac{1}{k}$ used in formula (5.23) is to make $N.SEL$ directly comparable across analyses with different value of the parameter k. The values of $N.SEL$ close to 1 indicate that despite small changes in the training data, we tend to consistently select the subset of k core gene sets. Bigger values of $N.SEL$ mean that the set of top k gene sets includes instable features which are likely to change as a result of slight modifications in data. Whether (i) there are any stable feature sets in the group of top k winning gene sets, and if so, (ii) which of them are stable, can be determined using the measures defined next.

The following measure of stability refers to the question (i) whether there is a core of stable features in the group of top k winning gene sets:

$$Q_\alpha = \sum_{i=1}^{M} I(c_i \geq \alpha n) \tag{5.24}$$

Note that for some $\alpha \leq 1$ the measure Q_α indicates how many gene sets (out of $k \times N.SEL$ selected at least once) tend to be repeatedly selected at least αn times in n iterations of the cross-validation loop. This measure is interesting for $N.SEL$ exceeding 1, as for e.g. $\alpha = 0.9$ it shows how many (out of top k) winning gene sets actually form stable feature sets, roughly insensitive to changes in data.

When empirically comparing different methods, (Chapter 6), it will be interesting to observe specific measures such as $Q_{0.9}$, $Q_{0.75}$ or $Q_{0.5}$. Clearly, stable feature selection is characterized by e.g. $Q_{0.9}$ close to k.

Additionally, we want to introduce a related measure which indicates how many times (out of n) the top k winning gene sets were actually selected in subsequent steps of cross-validation. This addresses the above formulated question (ii) related to which of the top k winning gene sets tend to be stable. More specifically, if $c_{[1]}$, $c_{[2]}$, ..., $c_{[M]}$ denote the counters (calculated in step 6 of the algorithm) sorted *descending*, then the frequency of selection of the top winning gene sets is simply:

$$\begin{aligned} FREQ_1 &= \frac{c_{[1]}}{n} \\ &\dots \\ FREQ_k &= \frac{c_{[k]}}{n} \end{aligned} \tag{5.25}$$

Obviously, stable feature selection is indicated by (most of) the values $FREQ_1$ through $FREQ_k$ close to 1. On the other hand, if for some $l < k$ we observe consistently higher values of $FREQ_1$, ..., $FREQ_l$ as compared with $FREQ_{l+1}$, ..., $FREQ_k$, this may be an additional criterion for selection of the parameter k in the algorithm (i.e., based on this observation, l should be used as the preferable value of the parameter k in the algorithm). See also Remark 3 on selection of the parameter k (page 102).

Finally, it is informative to report the p-values associated with the top winning gene sets selected in the cross-validation loop. More specifically, referring to step 5 of the algorithm (page 101), in every iteration of the cross-validation loop, we rank the gene sets based on increasing p-values. Let us denote these p-values calculated for subsequent gene sets as p^i_1, p^i_2, ..., p^i_M, where the upper index represents the iteration number in the cross-validation loop, $i = 1$, ..., n. Based on the p-values sorted ascending, denoted $p^i_{(1)}$, $p^i_{(2)}$, ..., $p^i_{(M)}$, we calculate the average p-value associated with the winning gene sets:

$$\begin{aligned} PVAL_1 &= \text{mean}(p^1_{(1)}, p^2_{(1)}, ..., p^n_{(1)}) \\ &\dots \\ PVAL_k &= \text{mean}(p^1_{(k)}, p^2_{(k)}, ..., p^n_{(k)}) \end{aligned} \tag{5.26}$$

Note that the value of $PVAL_i$ is most interesting for the stable feature set i, i.e. in the case when $FREQ_i$ is close to 1. Then $PVAL_i$ indicates the strength of association of the selected gene sets with the target. For this interpretation to hold, it is essential that in the algorithm (step 1 and 5) we rank the gene sets using the method $\mathcal{G}$ whose p-value indicate the strength of association (i.e. for this

reason, methods of gene set analysis based on Models 2 and 3 should be avoided – see sections 4.3.2, 4.3.3, 4.3.5).

5.5. Classification using standard methods of feature selection

The approach proposed in section 5.5 employs prior domain knowledge on possible, known relationships among features. Classification is then based on most activated gene sets (pathways) where gene set activation is detected by the self-contained of competitive methods as described in Chapter 4. In Chapter 6, we will compare this approach with commonly used methods where feature selection is done in the data-driven way. In the comparison, we will used uni- and multivariate methods, including shrinkage-based techniques, as described in section 2.3. We will refer to these data-driven methods as the *standard* methods of feature selection.

In this section, we first present the algorithm for sample classification based on standard methods of feature selection. We introduce measures of stability of data-driven feature selection, analogous to the measures presented in section 5.4. Results obtained with the standard methods will form the baseline results for evaluation of the proposed, domain knowledge-based algorithm.

We use here the same notation as introduced in the previous chapter, where $(W_{d\times n}, Y_{1\times n})$ represent the training data and the vector of class labels of the samples, respectively (see section 5.3 for details).

The following procedure takes (W, Y) as input and produces on output the predictive model, the subset of features used by the model, generalization error of the model and measures of stability of the selected features under small changes in data – as defined in section 5.6.

The algorithm is realized in the following steps.

1. Repeat steps 2 through 6 for $i = 1, 2, ..., n$.
2. Remove sample i, $i = 1, 2, ..., n$ from the data set (W, Y), i.e. remove column i from W and element i from Y. The remaining matrix and vector are denoted W^i and Y^i.
3. Using the feature selection method $\mathcal{F}$ and the data (W^i, Y^i), select k features (rows of W^i) most strongly associated with the target. Denote indices of these features as $S_i \subset \{1, 2, ..., d\}$. Note that if the feature selection procedure identifies $l < k$ features as associated with the target and the remaining features as unassociated with the target, then the actual number of features selected in this step is l.

4. Select rows from W^i related to the features in S_i. Denote the resulting matrix as $X^i = (W^i_{j\bullet} : j \in S_i)$.
5. Using the training data (X^i, Y^i), fit the predictive model f.
6. Classify the sample i as $\hat{Y}_i = f(Y_i)$.
7. Calculate expected prediction error as $EPE = \frac{1}{n}\sum_{i=1}^{n} I(\hat{Y}_i \neq Y_i)$.
8. Using the feature selection method $\mathcal{F}$ select top k features from W; denote indices of the features as S_0, and rows from the training data related to the selected features as $X = (W_{i\bullet} : i \in S_0)$. Based on the training data (X, Y) build the predictive model f.
9. Based on S_1, ..., S_n calculate the stability measures *n.sel* (formula (5.27)) and q_α (formula (5.29)), as defined in section 5.6.
10. Return the model f built in step 8, the subset of features S_0, EPE and the stability measures calculated in step 9.

Note that this procedure realizes internal cross-validation, which allows us to reduce over-optimism in the estimation of the generalization error, as explained in section 5.2.3. The EPE derived from this procedure is attributed to the final predictive model built from the complete training data (step 8).

Similarly to the algorithm in section 5.3, this algorithm is a generic framework in which we do not provide specific methods to realize feature selection $\mathcal{F}$ or classification f. We also employ the parameter k to control the dimensionality of the feature vectors used in classification.

Regarding feature selection, $\mathcal{F}$ can be realized as any univariate or multivariate method, such as the methods described in sections 2.2 and 2.3, which performs ranking of features and allows us to select top k highest-ranked features. If $\mathcal{F}$ does not explicitly rank individual features but rather finds feature sets (such as the RFE – section 2.3.1 or greedy methods in – section 2.3.3), we can still select k 'best' features by observing association of individual features in the set with the target. By rerunning the algorithm with different values of k, we tune the parameter k, in as similar way as described in Remark 3 (page 102).

It should be noted that some methods $\mathcal{F}$ which use shrinkage techniques may return fewer features (denoted l in step 3 of the algorithm, $l < k$) then the requested number of k features. For instance, this effect is commonly observed in the lasso method (page 24). It is due to the way the lasso optimization task (Equation (2.7)) is solved, which tends to set a number of coefficients β_0, β_1, ..., β_d to zero, which effectively eliminates the corresponding variables. Similar effect may be observed with the elastic net algorithm (page 25).

5.6. Stability of features selected with standard methods

In subsequent iterations of the internal leave-one-out cross-validation procedure described in the previous section, we generate subsets of features S_1, ..., S_n (step 3 of the algorithm). Based on this, we can observe whether features selected from slightly different datasets (i.e. (W^1, Y^1), ..., (W^n, Y^n) where any two of them differ by only one sample) tend to be stable. To quantify this, we define several measures of $\mathrm{CV}_{\mathrm{loo}}$ stability of feature selection, similar to the measures introduced in section 5.4.

First we want to observe how many distinct features were selected in n steps of cross-validation, which can be calculated as $|\bigcup_{i=1}^{n} S_i|$. Based on this quantity, we define the first measure of stability of feature selection:

$$n.sel = \frac{1}{k^*} \left| \bigcup_{i=1}^{n} S_i \right| \tag{5.27}$$

where

$$k^* = \frac{1}{n} \sum_{i=1}^{n} |S_i| \tag{5.28}$$

is the average number of features actually selected in every iteration of cross-validation. Obviously, $k^* \equiv k$ for standard methods based on simple ranking of features, however, it is likely that $k^* < k$ for the lasso or similar shrinkage methods, which may generate fewer that the requested number of k features (see step 3 of the algorithm).

Note that if roughly the same set of features is repeatedly selected in subsequent steps of cross-validation, then $|\bigcup_{i=1}^{n} S_i|$ will be close to k^*. We use the term $\frac{1}{k^*}$ in order to make $n.sel$ directly comparable across studies with different values of k.

Values of $n.sel$ exceeding 1 indicate that slight changes in data can lead to selection of sets of features with little overlapping, which means that the process of feature selection in unstable. Note that $n.sel$ is the equivalent of the $N.SEL$ measure used with gene set-based feature selection (formula (5.23)).

It is also interesting to observe whether in the set $\bigcup_{i=1}^{n} S_i$ there is some core subset of features repeatedly selected despite changes in the training data. To analyze this, we define the measure:

$$q_\alpha = \sum_{j \in S} I\left(\left(\sum_{i=1}^{n} I_{ij} \right) \geq \alpha n \right) \tag{5.29}$$

where $S = \bigcup_{i=1}^{n} S_i$, and I_{ij} indicates whether a feature $j \in S$ was selected in the iteration i of the cross-validation procedure, i.e. $I_{ij} = 1$ for $j \in S_i$, and $I_{ij} = 0$ otherwise. This measure has similar interpretation as Q_α (formula (5.24)), e.g. for $\alpha = 0.9$, q_α equals the number of features which were selected at least 90% times in n iterations of the cross-validation procedure (i.e. $q_{0.9}$ may be regarded as the number of *stable* features). To get a better insight into stability of the competing feature selection procedures, we may observe $q_{0.9}$, $q_{0.75}$ or $q_{0.5}$. Clearly, stable feature selection is indicated by $q_{0.9}$ close to k^*, i.e. close to the actual number of features selected in an iteration of the cross-validation procedure.

Summarizing, in this chapter we defined the algorithm for classification of samples in high dimensional data (section 5.3). The key characteristic of this algorithm is the prior domain knowledge-based feature selection. We also provided the alternative algorithm where features selection relies entirely on the data-driven methods (section 5.5). In the following chapter, these two approaches are evaluated numerically in terms of generalization and stability.

Chapter 6

Numerical evaluation of the proposed methods

In this chapter, we empirically compare the proposed approach to classification employing prior domain knowledge-based feature selection with purely data-driven (*standard*) methods. These two approaches will be compared in terms of (i) the expected prediction error of the commonly used classifiers, and (ii) stability of feature selection from data undergoing small changes. Essentially, we will compare results obtained with the algorithms presented in sections 5.3 and 5.5, and stability measures defined in sections 5.4 and 5.6.

In this study, we will use simulated data, as we want to analyze predictive performance and stability as a function of known properties of data, such as correlation of features or signal-to-noise ratio.

In the next section, we describe organization of the numerical study and characteristics of the simulated data. Since in simulation we exactly know which features and gene sets are associated with the target (which we refer to as *relevant*), we define additional measures which allow us to track selection of relevant vs. irrelevant features. This is meant as illustration of the theoretical analysis presented in Chapter 3, where we evaluated the risk of selection of irrelevant features from $n \ll d$ data.

In section 6.2, we present results of sensitivity studies which illustrate properties of the standard and prior domain knowledge-based methods as a function of the varying signal-to-noise ratio, correlation among features and the sample size.

6.1. Organization of the numerical study

This study is based on the simulated dataset with n samples and d genes out of which the first $n.DE$ genes are associated with the target (i.e. differentially expressed) and possibly correlated. The target is defined as the binary variable with the value of 0 and 1 in the two groups of $n/2$ samples (we refer to these groups of samples as "group 0" and "group 1"). We assume that

expression of the $n.DE$ genes associated with the target comes from the multivariate normal distribution $MVN(0_{1\times n.DE}, \Sigma_{n.DE\times n.DE})$ in the first group and $MVN(\Delta_{1\times n.DE}, \Sigma_{n.DE\times n.DE})$ in the second group, where $0_{1\times n.DE}$ and $\Delta_{1\times n.DE}$ represent vectors of means with the elements equal 0 and Δ, respectively. Δ represents the strength of the differential expression effect between the groups. The covariance matrix Σ (the same in both groups) has the diagonal elements equal 1 and the remaining elements equal ρ. Note that since variances of the $n.DE$ variables equal 1, then ρ also represents mutual correlation of the variables.

We assume that the remaining $d-n.DE$ genes in the dataset are not correlated and not associated with the target, and we generate expression of these genes from the standard normal distribution for both groups of samples.

We also assume that the database of gene sets is given which represents prior domain knowledge on feature relationships, and that it includes M gene sets of size m each. The first gene set contains the first m genes, i.e. it also contains all the $n.DE$ genes associated with the target (as we assume that $n.DE \leq m$), while the remaining gene sets are generated randomly.

In the study, we will report results for $n = 30$ or 50 samples, with $d = 5000$ genes in the dataset, out of which $n.DE = 20$ are differentially expressed. We assume $M = 100$ genes sets each with $m = 40$ genes. We will vary the effect strength ($\Delta = 0.5,\ 1,\ 1.5$) and the correlation ($\rho = 0,\ 0.2,\ 0.4$). Note that since we keep the variance of the signal fixed, then by varying Δ we effectively change the signal-to-noise ratio in the data.

This data organization was inspired by the empirical simulation studies reported in literature, e.g., Ackermann and Strimmer (2009); Dinu *et al.* (2008).

We will analyze this simulated dataset using standard methods of feature selection (algorithm in section 5.5) and using feature selection based on prior domain knowledge (algorithm in section 5.3). We will generate the data 100 times and report the measures of stability and the generalization error of the classifiers as the average values calculated over replications of the experiment.

Note that in our experiment we know exactly which features and which gene sets are nominally associated with the target; we name these features and gene sets as *relevant*, and the remaining features and gene sets as *irrelevant*. Hence features 1 through $n.DE$ are relevant and the gene set 1 is also relevant. It will be interesting to observe which features (relevant or irrelevant) are actually selected by the different methods under varying characteristics of data (such as the signal-to-noise or correlation between features). This problem, related to high risk of selection or irrelevant features from $n \ll d$ data, is discussed from theoretical standpoint in

Chapter 3; while here we analyze this empirically. To observe this, we introduce the following additional measures.

For the case of standard methods of feature selection, we will analyze the sets S_i, $i = 1, ..., n$ (see step 3 of the algorithm in section 5.5) in terms of the proportion of relevant features included in these sets.

For this purpose we calculate the following measures. First we want to analyze the share of relevant features in the group of q_α (formula 5.29) features deemed stable (which means selected with frequency at least α in the cross-validation procedure). We report this using the following measure, similar to q_α but referring to the relevant features:

$$qr_\alpha = \sum_{j \in SR} I\left(\left(\sum_{i=1}^{n} IR_{ij}\right) \geq \alpha n\right) \tag{6.1}$$

where the set SR and the indicators IR_{ij} are related to the relevant features selected in the cross-validation procedure. More specifically, if we denote the set of the relevant features (their indices) as RF (in our study $RF = \{1, 2, ..., n.DE\}$), then the sets of relevant features selected in subsequent iterations are simply $SR_i = RF \cap S_i$, $i = 1, ..., n$. The set of all relevant features ever selected in the cross-validation procedure is $SR = \bigcup_{i=1}^{n} SR_i$. The indicators IR_{ij} show if the relevant feature $j \in SR$ was selected in iteration i (then $IR_{ij} = 1$), or not (then $IR_{ij} = 0$).

The second measure indicates the average number of relevant features (out of k^*) selected in subsequent iterations of cross-validation, i.e.

$$kr^* = \frac{1}{n} \sum_{i=1}^{n} |SR_i| \tag{6.2}$$

If mostly relevant features are being selected then kr^* should be close to k^*.

For the case of feature selection based on prior domain knowledge (algorithm in section 5.3, page 100), considering our simulation scenario, selection of the relevant gene set in subsequent iterations is recorded in the counter c_1 (see step 8 of the algorithm). We will observe the frequency of selection of this relevant gene set which we define as:

$$FREQ.R = \frac{c_1}{n} \tag{6.3}$$

It will be interesting to compare $FREQ.R$ with the frequency of selection of the top winning gene sets $FREQ_1 = \frac{c_{[1]}}{n}$, etc. (formula 5.25). Note that in section

5.4 we introduced the latter frequencies as indications of whether the gene sets selected are sensitive to changes in data. Now using $FREQ.R$ we will be able to tell whether the gene sets selected include the relevant gene set.

6.2. Results of the numerical study

In this section, we report results of the algorithm presented in section 5.3 and in section 5.5.

Regarding the algorithm in section 5.5, we use the following standard methods of feature selection, both univariate and multivariate:

- Feature ranking based on the Wilcoxon nonparametric test or on the t-test (section 2.2),
- Recursive Feature Elimination (section 2.3.1),
- Lasso (section 2.3.4),
- Elastic Net (section 2.3.4).

In the algorithm, we used the following predictive models (f):

- Support Vector Machine with the linear kernel (SVM),
- Penalized logistic regression (PLR), as proposed by Zhu and Hastie (2004),
- Nonparametric nearest neighbours (KNN).

We performed the analysis with the parameter k of the algorithm (see steps 3 and 8 of the algorithm, page 106) equal 5, 10, 15, 20.

Regarding the algorithm in section 5.3, we used the following methods of gene set analysis ($\mathcal{G}$):

- GT – Globaltest algorithm (Equation (4.1)),
- GSA algorithm (point 3 on page 51),
- GSA2 – the modified version of the GSA method with the p-value calculated according to Equation 4.7.

In this numerical study we used the following classifiers (g, step 7 and 10 of the algorithm in section 5.3):

- Support Vector Machine with the linear kernel (SVM),
- Nonparametric classifiers based on signatures of gene set activation in individual samples, proposed in sections 5.1.1, 5.1.2 and 5.1.3.

We performed the analysis with parameter k of the algorithm (see step 2 of the algorithm, page 100) equal 1, i.e. we deliberately want to select one gene set

since in the simulated dataset only the first gene set, out of $M = 100$, is assumed to be associated with the target. As the multiple testing adjustment (see step 1 of the algorithm) we use in this study the Holm procedure which controls the family-wise error rate (Holm, 1979; Dudoit *et al.*, 2003), see also Remark II on page 19.

6.2.1. Generalization error with standard feature selection

We first analyze expected prediction error of selected classifiers using standard methods of feature selection. Results for small, medium and strong effect (Δ) are given in Tables 6.1 through 6.3. In the tables, we show the relationship between the EPE (calculated in step 7 of the algorithm in section 5.5) and the sample size and correlation among the genes associated with the target. We make the following observations:

1. For the small effect ($\Delta = 0.5$, Table 6.1), none of the standard methods of feature selection is able to provide informative features, as the lowest classification error reported in Table 6.1 exceeds 45%. Note however that this data realizes very low signal to noise ratio, with the difference in signal

Table 6.1. Expected prediction error for standard methods of feature selection as a function of correlation among genes ρ, and the sample size n. Results for the small effect, $\Delta = 0.5$

Samples n	Cor ρ	Classifier	Feature selection				
			t-test	Wilcox	RFE	Lasso	Enet
30	0.0	SVM	0.52	0.526	0.511	0.5	0.485
		PLR	0.525	0.513	0.521	0.489	0.484
		KNN	0.512	0.501	0.523	0.485	0.475
	0.2	SVM	0.527	0.543	0.518	0.496	0.486
		PLR	0.52	0.519	0.513	0.487	0.482
		KNN	0.529	0.519	0.523	0.503	0.498
	0.4	SVM	0.525	0.537	0.527	0.5	0.491
		PLR	0.522	0.515	0.522	0.489	0.491
		KNN	0.497	0.52	0.526	0.515	0.495
50	0.0	SVM	0.468	0.473	0.455	0.46	0.465
		PLR	0.468	0.477	0.456	0.464	0.464
		KNN	0.46	0.469	0.469	0.476	0.494
	0.2	SVM	0.488	0.492	0.493	0.477	0.488
		PLR	0.494	0.505	0.488	0.487	0.484
		KNN	0.495	0.491	0.507	0.49	0.508
	0.4	SVM	0.487	0.486	0.498	0.484	0.486
		PLR	0.496	0.488	0.496	0.496	0.477
		KNN	0.493	0.484	0.511	0.502	0.503

Table 6.2. Expected prediction error for standard methods of feature selection as a function of correlation among genes ρ, and the sample size n. Results for the medium effect, $\Delta = 1$

Samples n	Cor ρ	Classifier	Feature selection				
			t-test	Wilcox	RFE	Lasso	Enet
30	0.0	SVM	0.299	0.305	0.235	0.288	0.284
		PLR	0.342	0.346	0.289	0.286	0.285
		KNN	0.284	0.286	0.277	0.293	0.29
	0.2	SVM	0.372	0.367	0.326	0.35	0.333
		PLR	0.417	0.419	0.384	0.35	0.332
		KNN	0.35	0.357	0.366	0.37	0.363
	0.4	SVM	0.412	0.417	0.368	0.388	0.388
		PLR	0.438	0.462	0.415	0.381	0.39
		KNN	0.386	0.39	0.393	0.411	0.378
50	0.0	SVM	0.128	0.146	0.115	0.138	0.151
		PLR	0.147	0.173	0.131	0.139	0.152
		KNN	0.144	0.159	0.144	0.164	0.179
	0.2	SVM	0.298	0.31	0.279	0.318	0.294
		PLR	0.34	0.357	0.318	0.332	0.304
		KNN	0.289	0.31	0.306	0.334	0.311
	0.4	SVM	0.358	0.355	0.347	0.371	0.321
		PLR	0.399	0.404	0.381	0.384	0.332
		KNN	0.341	0.344	0.337	0.364	0.337

between the groups equal half of standard deviation of the signal. This example empirically confirms results derived in Chapter 3 related to high risk of selecting irrelevant features in $d \gg n$ cases. Note that the following observations (points 2 through 5) do not apply to this lowest signal-to-noise case.

2. For the medium (Table 6.2) or strong effect (Table 6.3), EPE clearly depends on both the sample size and the correlation of genes. Interestingly, reduction of the EPE with growing sample size is most significant for uncorrelated features (e.g. in Table 6.2 we observe reduction of the EPE for $\rho = 0$ by roughly 50% if sample size is increased from 30 to 50). This effect tends to weaken if features are correlated. Generally, growing correlation between features leads to increasing EPE, which is in line with the conclusions drawn in section 4.6.
3. Regarding the worth of competing feature selection methods, we observe that Recursive Feature Elimination (RFE) generally outperforms other methods, especially for uncorrelated data – see e.g. Table 6.2 and 6.3, results for $\rho = 0$. We also observe that RFE coupled with the SVM classifier

Table 6.3. Expected prediction error for standard methods of feature selection as a function of correlation among genes ρ, and the sample size n. Results for the strong effect, $\Delta = 1.5$

Samples n	Cor ρ	Classifier	Feature selection				
			t-test	Wilcox	RFE	Lasso	Enet
30	0.0	SVM	0.031	0.034	0.014	0.081	0.102
		PLR	0.09	0.093	0.062	0.081	0.101
		KNN	0.03	0.026	0.014	0.099	0.098
	0.2	SVM	0.167	0.155	0.132	0.171	0.176
		PLR	0.297	0.286	0.248	0.173	0.171
		KNN	0.162	0.165	0.138	0.212	0.174
	0.4	SVM	0.239	0.251	0.215	0.258	0.251
		PLR	0.365	0.374	0.315	0.258	0.264
		KNN	0.215	0.23	0.212	0.245	0.248
50	0.0	SVM	0.004	0.004	0.002	0.031	0.051
		PLR	0.012	0.012	0.009	0.035	0.052
		KNN	0.005	0.006	0.004	0.041	0.066
	0.2	SVM	0.101	0.098	0.097	0.12	0.129
		PLR	0.167	0.163	0.159	0.127	0.131
		KNN	0.115	0.109	0.115	0.131	0.139
	0.4	SVM	0.178	0.176	0.168	0.208	0.172
		PLR	0.256	0.256	0.242	0.22	0.179
		KNN	0.192	0.204	0.191	0.217	0.201

consistently realizes the smallest prediction error observed over all other combinations of feature selectors and classifiers (see Tables 6.2 and 6.3). However, this result is not surprising considering the fact that the RFE procedure employs the SVM algorithm in the process of ranking features (see section 2.3.1). Clearly, features selected by the RFE perform best with the SVM (as compared with PLR or KNN), as similarly, the SVM classifier performs best using the RFE features (as compared with the t-test, Wilcoxon, Lasso or Elastic net features).

4. We also observe that for the strong signal simple univariate feature ranking procedures (like the t-test or the Wilcoxon test) perform remarkably well (see e.g. Table 6.3, sections for $\rho = 0$), and seem to outperform more sophisticated shrinkage-based methods. This effect is most clear if the methods are coupled with the KNN or SVM classifiers. However, for weaker signal (e.g. Table 6.2) and/or for correlated features, multivariate methods show similar, or better performance.
5. We observe that all the classification models minimize generalization error if they are coupled with a preferable, model-specific method of feature se-

lection. For uncorrelated data, all the models seem to prefer the RFE, while for correlated data, the PLR classifier prefers the regularization methods (the Lasso or the Elastic net), SVM prefers the RFE method, and the KNN prefers the RFE or the univariate methods.

6.2.2. Stability of standard feature selection

We now analyze stability of features used by the classifiers whose prediction error is summarized in Tables 6.1–6.3. In Figures 6.1 and 6.2 we report the *n.sel* measure (formula 5.27) as a function of the signal-to-noise ratio and correlation among features, for $n = 30$ or $n = 50$ samples, respectively. Values of *n.sel* close to 1 indicate that throughout the cross-validation procedure we repeatedly selected roughly the same set of k (or k^*) features, i.e. the feature selection was stable and insensitive to small changes in data. Clearly, the RFE method generally realizes the smallest *n.sel* (Figure 6.1), although for bigger samples sizes, where stability of feature selection generally improves, Wilcoxon method performs equally well (Figure 6.2). This observation coincides with the conclusions drawn from Tables 6.2 and 6.3 that RFE generally outperforms other methods of feature selection (see also note 3 (page 116) and note 4 (page 117)).

Interestingly, for the RFE and the univariate methods, *n.sel* is independent of the correlation among features, while for the shrinkage methods (the Lasso and the Elastic net), *n.sel* increases with the growing correlation.

The analyses summarized in Figures 6.1 and 6.2 were done for $k = 20$. Hence, for RFE and the univariate methods, where $k^* \equiv k$, all the classifiers reported in Tables 6.1–6.3 used 20 features. For the case of strong signal and 50 samples (Figure 6.2, top panel) we observe that these 20 features were actually selected from the set of ca. $1.3 \times k$ features (more specifically, for RFE the set of features included ca. 24 features, while for the Wilcoxon method – about 26 features).

However, for the Lasso and the Elastic net, we generally observed $k^* < 20$. For instance, for $n = 50$ samples and the strong signal, k^* changed between 8 and 10. This means that e.g. for $\Delta = 1.5$ and $n = 50$ (Figure 6.2, top panel) Lasso selected on average 8 features per iteration out of the set of 12 (for correlation 0), and about 10 out of ca. 27 for correlation 0.4.

In Figure 6.3, we show the values k^* compared with the number of relevant features kr^* among the k^* features actually selected per iteration of cross-validation. Note that in Figure 6.3, the three plots in a row are labelled by the signal strength ($\Delta = 0.5,\ 1,\ 1.5$), and the three plots in a column are labelled by the correlation in data ($\rho = 0,\ 0.2,\ 0.4$). In the figure, k^* is shown as the sum of kr^* relevant

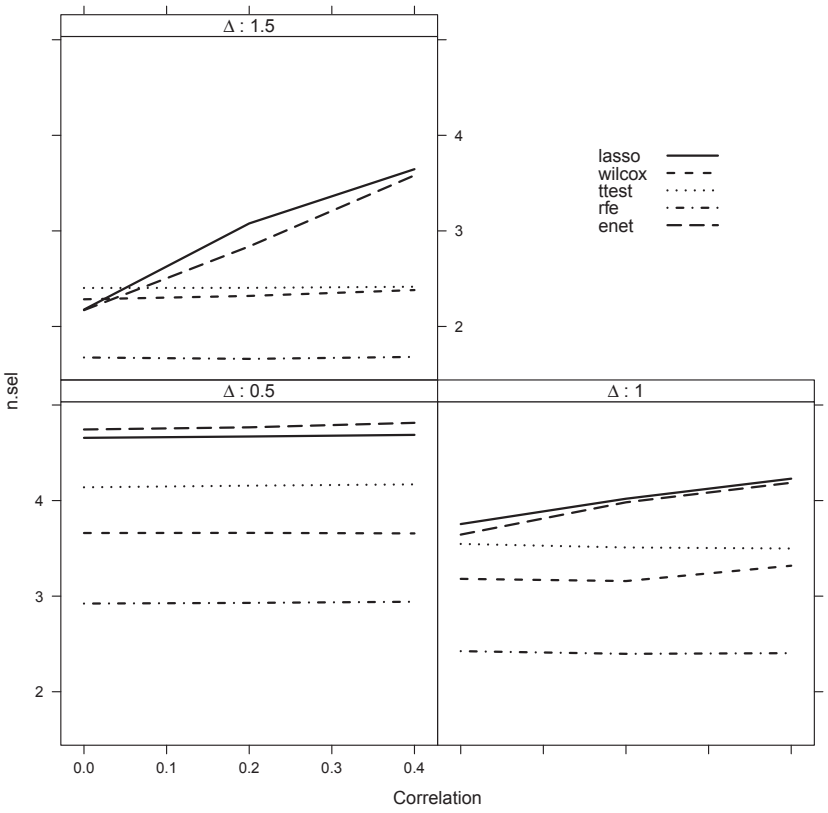

Fig. 6.1. Stability of standard feature selection: the *n.sel* measure (formula 5.27). Results for 30 samples

features (represented by the bright parts of the bars), and the remaining irrelevant features (represented by the dark parts of the bars).

We observe that for the small signal ($\Delta = 0.5$), standard methods of feature selection are virtually unable to select the informative, relevant features from the high-dimensional feature space. This explains the poor generalization error of classifiers reported in Tables 6.1–6.3. Note that the shrinkage methods (the Lasso and the Elastic net) do not perform better that simple univariate methods (t-test or Wilcoxon ranking) or the RFE. Clearly, with growing signal to noise ratio, the probability of selecting relevant features increases, which empirically illustrates the theoretical results in Chapter 3.

It is interesting to observe that performance of the univariate methods and the RFE algorithm is not affected by the correlation among features. However, the shrinkage methods are quite different in this respect: for the medium or strong

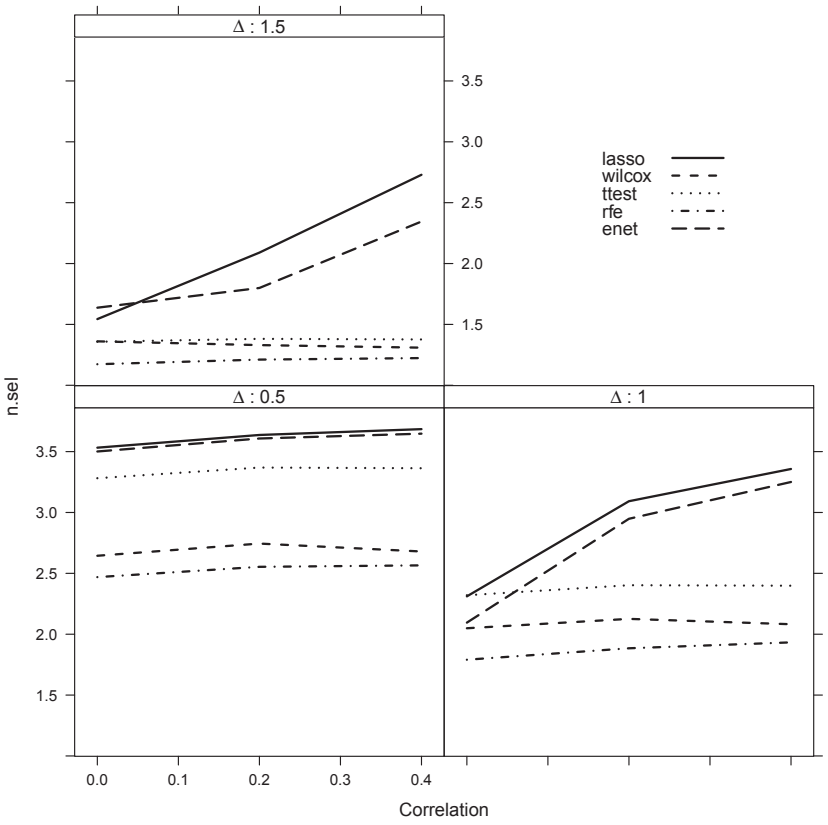

Fig. 6.2. Stability of standard feature selection: the *n.sel* measure (formula 5.27). Results for 50 samples

signal and uncorrelated data, they clearly outperform other methods in terms of the proportion of relevant features among the features selected, while for growing correlation in data, this proportion becomes remarkably poor as compared with the univariate methods.

In Figure 6.4, we analyze standard feature selection in terms of stability. We analyze whether there are some core subsets of stable features (among the total number of $n.sel \times k^*$ ever selected) which are insensitive to changes in data. In Figure 6.4, we report the number of stable features, $q_{0.9}$, compared with the number of relevant features, $qr_{0.9}$, among the $q_{0.9}$ stable features. The values of $q_{0.9}$ are represented by the total height of the bars, while the numbers of

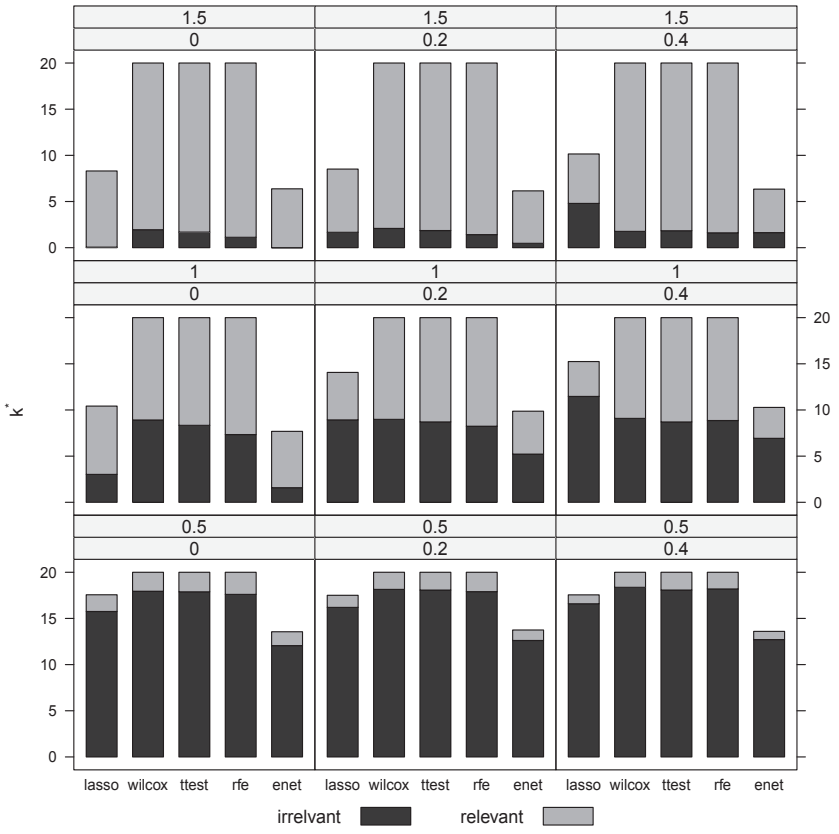

Fig. 6.3. Stability of standard feature selection: k^* and kr^* measures (formulae 5.28 and 6.2), as a function of correlation ($\rho = 0$, 0.2, 0.4) and signal to noise ($\Delta = 0.5$, 1, 1.5). Results for 50 samples. The number of features selected per iteration, k^*, is represented by the total height of a bar, while the number of relevant features, kr^*, is represented by the bright part of the bar. Dark parts represent $k^* - kr^*$ irrelevant features

relevant ($qr_{0.9}$) or irrelevant features are represented by the bright or dark parts, respectively.

First we observe that for the low signal ($\Delta = 0.5$) all the methods tend to find subsets of stable features, more specifically, roughly 50% of the features selected come out as stable features (compare Figures 6.3 and 6.4, bottom rows), However most of these stable features are irrelevant. This illustrates the major difficulties in identifying the *right* features from high-dimensional data, discussed from theoretical standpoint in Chapter 3.

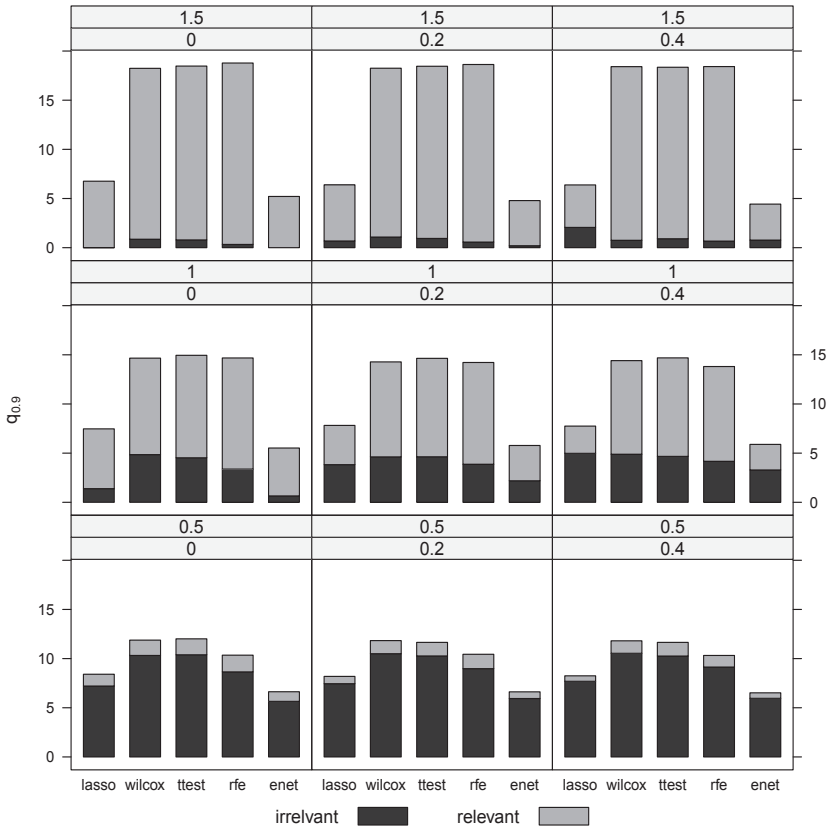

Fig. 6.4. Stability of standard feature selection: $q_{0.9}$ and $qr_{0.9}$ measures (formulae 5.29 and 6.1) shown as a function of correlation ($\rho = 0,\ 0.2,\ 0.4$) and signal to noise ($\Delta = 0.5,\ 1,\ 1.5$). Results for 50 samples. The number of stable features repeatedly selected in cross-validation, $q_{0.9}$, is represented by the total height of a bar, while the number of stable and relevant features is represented by the bright part. Dark parts represent the $q_{0.9} - qr_{0.9}$ stable but irrelevant features

We also observe that for the medium or strong signal, the proportion of relevant features among the stable features is generally higher than the proportion of relevant features among all the k^* features selected (compare Figures 6.3 and 6.4, middle rows: e.g. for univariate methods, roughly 50% of the $k^* = 20$ selected features are relevant, Figure 6.3, while among stable features the relevant features contribute about 2/3 (univariate) or 3/4 (RFE) features, Figure 6.4). This shows that ($\mathrm{CV}_{\mathrm{loo}}$) stability of features is indeed related to informativeness of the the features. We again observe that for uncorrelated data and the strong signal,

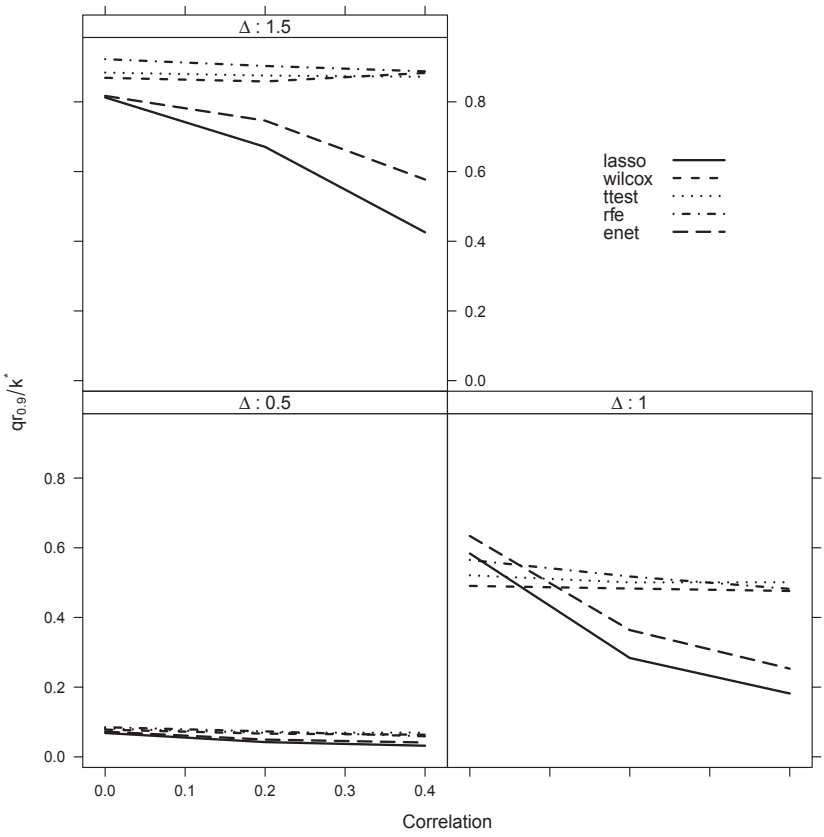

Fig. 6.5. Stability of standard feature selection: the $qr_{0.9}$ measure (formula 6.1) shown relative to the number of features actually selected k^*. Results for 50 samples

shrinkage methods outperform other methods in terms of the $qr_{0.9}/q_{0.9}$ ratio, we note that *all* the features selected from such data by Lasso or Elastic net are relevant. However, these methods are more affected by the correlation in data as compared with the univariate methods or with the RFE.

Finally, in Figure 6.5 we aggregate results presented in Figures 6.3 and 6.4 and show $qr_{0.9}$ relative to k^*, which gives the proportion of stable, relevant features among the features actually returned by feature selection methods. We observe that for the strong signal ($\Delta = 1.5$), roughly 85% of the features selected by RFE and the univariate methods are stable and relevant, with the RFE slightly outperforming the univariate methods. For uncorrelated data, the Lasso and the Elastic net show similar performance (or better, for $\Delta = 1$), however performance

of shrinkage methods decreases with growing correlation in data. We also observe that the Lasso is generally more affected by correlation in data than the Elasitc net (for instance, for $\Delta = 1.5$ and $\rho = 0.4$, $\frac{qr_{0.9}}{k^*} = 40\%$ of the features retured by the Lasso are the stable and relevant features, while for the Elastic net this proportion amounts 60%, Figure 6.5, top panel). This can be accounted for by observing the key difference between these methods in terms of how the methods treat correlated data: the Elastic net tends to keep all correlated variables in or out of the model, while the Lasso selects one variable out of the group of related variables and removes the remaining ones (section 2.3.4, page 24).

6.2.3. Generalization error with feature selection based on prior domain knowledge

In this experiment, we analyze the same data as we analyzed in the previous study using the algorithm presented in section 5.5, however now we use the algorithm presented in section 5.3. In this algorithm (step 2), we use as features for sample classification only the gene sets which are significantly associated with the target, based on the p-value ≤ 0.05 threshold, where the p-value is (Holm) multiple testing adjusted. Although in the original algorithm in section 5.3, the procedure stops is none of the gene sets is significantly associated with the target, in this study we report results irrespective of whether the "winning" gene set is significant. We report (i) the predictive performance and (ii) the p-values associated with the "winning" gene sets ($PVAL_1$, formula 5.26). We chose to do this, rather than strictly follow step 2 of the algorithm which stops on the $p_{(1)} > 0.05$ condition, in order to provide a comprehensive illustration of the relationship between the strength of association of a gene set with the target and its performance as the set of features in classification.

Expected prediction error observed using this approach is summarized in Tables 6.4–6.6, and the p-values associated with selection of the "winning" gene sets are reported in Tables 6.7 and 6.8 (where the former reports the multiple testing adjusted $PVAL_1$ and the latter reports raw, unadjusted p-values). Clearly, we do not expect good classification from insignificant gene sets, however we want to observe whether significant gene sets guarantee good classification. We discuss this further in section 6.3.

In this study, we compared one self-contained (Globaltest) and two competitive methods (GSA and its modified version, GSA2) of gene set analysis used for feature selection, together with different classifiers: one parametric algorithm (SVM) and three nonparametric methods proposed in sections 5.1.1 through 5.1.3. To

ease comparison of results of this study with results in Tables 6.1–6.3), we visually compare performance of the best standard methods (realized by the RFE+SVM pair) with performance of the two classifiers which use features selected by the GSA2, i.e. SVM and the nonparametric Method 1 – see Figure 6.6.

We make the following observations:

1. Comparing results for the low signal to noise ratio ($\Delta = 0.5$, Table 6.1 and 6.4), we observe that feature selection based on prior domain knowledge leads to reduction of the generalization error to about 25% for uncorrelated data (results for SVM or Method 1, 50 samples). Note that purely data driven methods are virtually unable to find informative features from such data and as such lead to classifiers with no generalization property. However, under correlated data, classifiers using most activated feature sets

Table 6.4. Expected prediction error for feature selection based on prior domain knowledge as a function of correlation among genes ρ, and the sample size n. Results for the small effect, $\Delta = 0.5$

Samples n	Cor ρ	Classifier	Feature selection		
			Globaltest	GSA	GSA2
30	0.0	Method 1	0.373	0.324	0.323
		Method 2	0.399	0.351	0.359
		Method 3	0.395	0.357	0.373
		SVM	0.352	0.295	0.304
	0.2	Method 1	0.457	0.451	0.454
		Method 2	0.457	0.456	0.454
		Method 3	0.454	0.464	0.456
		SVM	0.467	0.463	0.458
	0.4	Method 1	0.487	0.495	0.483
		Method 2	0.481	0.488	0.47
		Method 3	0.478	0.492	0.479
		SVM	0.492	0.484	0.485
50	0.0	Method 1	0.251	0.255	0.254
		Method 2	0.316	0.325	0.311
		Method 3	0.313	0.322	0.306
		SVM	0.259	0.248	0.242
	0.2	Method 1	0.378	0.405	0.387
		Method 2	0.391	0.405	0.393
		Method 3	0.391	0.4	0.394
		SVM	0.437	0.442	0.422
	0.4	Method 1	0.444	0.465	0.442
		Method 2	0.45	0.454	0.444
		Method 3	0.451	0.447	0.445
		SVM	0.472	0.477	0.456

Table 6.5. Expected prediction error for feature selection based on prior domain knowledge as a function of correlation among genes ρ, and the sample size n. Results for the medium effect, $\Delta = 1$

Samples n	Cor ρ	Classifier	Feature selection		
			Globaltest	GSA	GSA2
30	0.0	Method 1	0.07	0.058	0.07
		Method 2	0.128	0.125	0.126
		Method 3	0.128	0.121	0.123
		SVM	0.056	0.056	0.07
	0.2	Method 1	0.224	0.2	0.201
		Method 2	0.231	0.215	0.217
		Method 3	0.232	0.217	0.217
		SVM	0.266	0.249	0.265
	0.4	Method 1	0.317	0.279	0.278
		Method 2	0.321	0.283	0.282
		Method 3	0.322	0.283	0.282
		SVM	0.341	0.34	0.334
50	0.0	Method 1	0.03	0.028	0.031
		Method 2	0.108	0.098	0.095
		Method 3	0.102	0.092	0.096
		SVM	0.04	0.04	0.045
	0.2	Method 1	0.158	0.158	0.165
		Method 2	0.183	0.173	0.18
		Method 3	0.177	0.171	0.181
		SVM	0.221	0.225	0.227
	0.4	Method 1	0.245	0.224	0.236
		Method 2	0.247	0.223	0.245
		Method 3	0.251	0.221	0.242
		SVM	0.302	0.298	0.297

also loose generalization property. Note that for such data, all the gene set analysis methods fail to identify significant feature sets (Table 6.7, results for $n = 50$, $\Delta = 0.5$, $\rho > 0$). This may result from growing overlapping of features observed under correlation – effect discussed in section 4.6.

2. For the medium effect ($\Delta = 1$), the generalization error also improves if the classification is based on the activated feature sets. The most spectacular improvement is observed for low-correlation data (see Figure 6.6 and Tables 6.2 and 6.5). For instance, for the correlation = 0, we reduce the EPE from 12% to 3% (results for 50 samples), and from 24% to 7% (results for 30 samples). For correlation = 0.4, we observe reduction from 35% to ca 24%. Note that in the case of strong signal ($\Delta = 1.5$, Tables 6.3 and 6.6, Figure 6.6), where standard methods perform well, we still observe

Table 6.6. Expected prediction error for feature selection based on prior domain knowledge as a function of correlation among genes ρ, and the sample size n. Results for the strong effect, $\Delta = 1.5$

Samples n	Cor ρ	Classifier	Feature selection		
			Globaltest	GSA	GSA2
30	0.0	Method 1	0.005	0.004	0.004
		Method 2	0.021	0.024	0.023
		Method 3	0.019	0.022	0.02
		SVM	0.005	0.006	0.004
	0.2	Method 1	0.081	0.07	0.08
		Method 2	0.09	0.08	0.094
		Method 3	0.09	0.08	0.088
		SVM	0.119	0.112	0.112
	0.4	Method 1	0.145	0.15	0.144
		Method 2	0.155	0.153	0.149
		Method 3	0.155	0.153	0.15
		SVM	0.204	0.205	0.199
50	0.0	Method 1	0.002	0.001	0.003
		Method 2	0.016	0.013	0.014
		Method 3	0.014	0.012	0.014
		SVM	0.005	0.003	0.003
	0.2	Method 1	0.066	0.067	0.07
		Method 2	0.08	0.077	0.078
		Method 3	0.08	0.077	0.076
		SVM	0.105	0.1	0.102
	0.4	Method 1	0.13	0.128	0.133
		Method 2	0.134	0.134	0.135
		Method 3	0.135	0.135	0.133
		SVM	0.179	0.17	0.172

improved classification using prior domain knowledge. For instance, for 30 samples, EPE is reduced by roughly 1/2 to 1/3 as compared with the EPE for the standard methods, and for 50 samples, EPE is reduced by roughly 1/4).

3. Considering different classification methods used in this study, we observe that under the low correlation data, Method 1 and SVM significantly outperform the two remaining nonparametric classifiers (see Tables 6.4–6.6, results for $\rho = 0$). However, under correlation in data, Method 1 as well as Method 2 and Method 3 significantly outperform the SVM classifier (see Tables 6.5 and 6.6, results for $\rho = 0.2, 0.4$). Therefore, Method 1 can be regarded as the preferable approach to classification based on the activation of features sets (see also section 6.3 for a more detailed comment on this).

Table 6.7. Multiple testing corrected p-value related to selection of the "winning" gene set, $PVAL_1$ (formula 5.26) produced by the Globaltest (GT), GSA and GSA2 as a function of signal strength (Δ), correlation of features (ρ) and sample size (n)

Δ	Method	Correlation ρ					
		$n = 30$			$n = 50$		
		0	0.2	0.4	0	0.2	0.4
0.5	GT	0.23	0.45	0.53	0.024	0.23	0.41
	GSA	0.0098	0.11	0.16	1e−04	0.063	0.14
	GSA2	0.06	0.39	0.48	0.014	0.22	0.37
1	GT	3.1e−06	0.044	0.16	1.2e−13	0.00065	0.026
	GSA	0	0.01	0.058	0	0	0.0049
	GSA2	6.7e−05	0.028	0.11	0	0.0011	0.026
1.5	GT	7.3e−12	5.1e−05	0.0053	7.2e−22	6.1e−09	1.5e−05
	GSA	0	0	5e−04	0	0	0
	GSA2	0	0	0.0023	0	0	0

Table 6.8. Raw p-values (prior to multiple testing correction) related to selection of the "winning" gene set, produced by the Globaltest (GT), GSA and GSA2 as a function of signal strength (Δ), correlation of features (ρ) and sample size (n)

Δ	Method	Correlation ρ					
		$n = 30$			$n = 50$		
		0	0.2	0.4	0	0.2	0.4
0.5	GT	0.0027	0.0064	0.0079	0.00026	0.0027	0.0053
	GSA	9.8e−05	0.0012	0.0017	1e−06	0.00066	0.0015
	GSA2	0.00067	0.0051	0.0067	0.00014	0.0026	0.0049
1	GT	3.1e−08	0.00047	0.0018	1.2e−15	6.5e−06	3e−04
	GSA	0	0.00011	0.00066	0	0	4.9e−05
	GSA2	6.7e−07	0.00028	0.0011	0	1.1e−05	0.00027
1.5	GT	7.3e−14	5.1e−07	5.3e−05	7.2e−24	6.1e−11	1.5e−07
	GSA	0	0	5e−06	0	0	0
	GSA2	0	0	2.3e−05	0	0	0

6.2.4. Stability of prior domain knowledge-based feature selection

We provide analysis of stability of selection of features sets by the algorithm in section 5.3.

Note that in the simulation study we assumed that only one out of M gene sets is associated with the target (section 6.1), therefore here we report results of the algorithm executed with the parameter $k = 1$ (see Step 2 of the algorithm on page 100, and Remark 3, page 102).

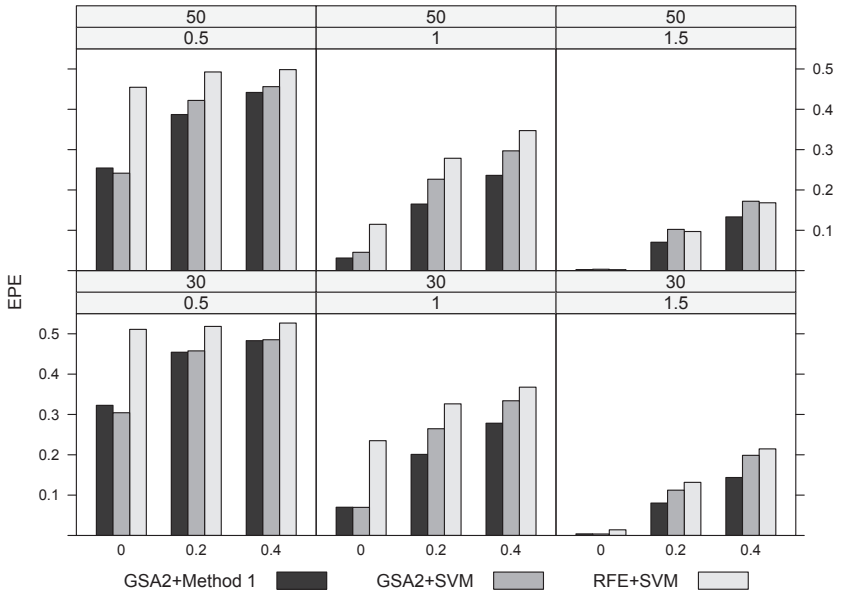

Fig. 6.6. Generalization error of the SVM using standard feature selection, RFE, (RFE+SVM), and of the two classifiers based on features selected with the GSA2: the SVM (GSA2+SVM) and the nonparametric classifier Method 1 (GSA2+Method 1), as a function of signal level ($\Delta = 0.5,\ 1,\ 1.5$), correlation in data ($\rho = 0,\ 0.2,\ 0.4$) and the sample size $n = 30,\ 50$

First we analyze the $N.SEL$ measure (formula 5.23) which indicates how many different gene sets were selected during the cross-validation procedure. Results are summarized in Figure 6.7. We observe that if enough samples are available ($n = 50$) then for the strong, medium or even weak signal (providing the features are uncorrelated), $N.SEL \approx 1$, i.e. despite changes in the training data, the algorithm tends to repeatedly select only one, the same gene set. Growing correlation among features leads to worse stability, especially for weaker signal and fewer samples, e.g. for $\rho = 0.4$, $N.SEL \approx 2.5$ for GSA2 and Globaltest, and $N.SEL \approx 5$ for GSA (Figure 6.7, top-left cell). Interestingly, the GSA is most affected by correlation among features, as compared with the GSA2 or the Globaltest.

Comparing these results with stability of standard feature selection ($n.sel$ measure, results in Figures 6.1 and 6.2), we observe that using prior domain knowledge generally leads to much better stability of features, e.g. for the medium signal ($\Delta = 1$) and 50 samples, $N.SEL = 1$, while $n.sel \approx 2$ for the best standard feature selection method – RFE (compare Figure 6.2, right panel, with Figure 6.7, top-middle cell). For the strong signal, prior domain knowledge feature selection

is always stable, even for smaller sample size ($n = 30$), whereas standard method require larger sample sizes to realize nearly stable features (see top panels in Figures 6.1 and 6.2). We notice however that all the gene set analysis methods tested (GSA, GSA2 and Globaltest) are affected by the correlation among features, especially for the case of weak signal and/or fewer samples. This effect is observed only with shrinkage methods, other standard methods are immune to correlation among features.

Next, we analyze the $FREQ_1$ (formula 5.25) and $FREQ.R$ (formula 6.3) measures. Obviously, for the cases where $N.SEL \approx 1$, $FREQ_1$ is also close to 1 (if only $N.SEL = 1$ feature is repeatedly selected, then the frequency of selection of the most frequently selected gene set, $FREQ_1$, must be 1). Hence $FREQ_1$ is most informative in the cases of less stable feature selection, as it then indicates whether among the $N.SEL$ feature sets ever selected some feature sets are stable. Results are presented in Figure 6.8. We observe that for the low signal GSA2 and Globaltest realize more stable selection then the GSA methods, e.g. for correlation $\rho = 0.4$, the "winning" gene set is selected ca. 80% times by GSA2

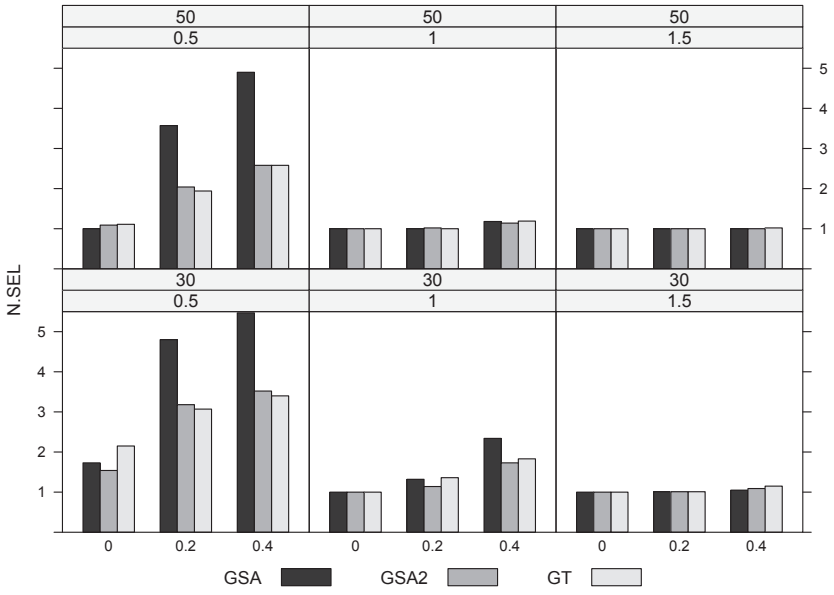

Fig. 6.7. Stability of feature selection based on prior domain knowledge: the $N.SEL$ measure (formula 5.23) as a function of the signal strength ($\Delta = 0.5, 1, 1.5$), correlation among features ($\rho = 0, 0.2, 0.4$) and the sample size ($n = 30, 50$)

or the Globaltest, but only ca. 60% times by the GSA (Figure 6.8, top-left and bottom-left cells).

It is interesting to investigate whether these stable gene sets include the relevant gene set. This is indicated by the $FREQ.R$ measure, with results summarized in Figure 6.9. We observe that for all the cases where feature selection brings stable feature sets (i.e. in the studies where $N.SEL \approx 1$ and $FREQ_1 \approx 1$, Figures 6.7 and 6.8), the stable feature set is actually the relevant gene set. However, if stability decreases, the chance of selecting the relevant feature sets tends to drop below the $FREQ_1$ level, e.g. in the worst case $FREQ.R \approx 0.2$ while $FREQ_1 \approx 0.75$ (see Figures 6.8 and 6.9, bottom-left cells, results for 30 samples, low signal and strong correlation). This means that under the low signal and correlated features, some non-relevant feature sets may come out as more stable than the relevant feature set. This effect is similar to the results shown in Figure 6.4, bottom panel, where we demonstrated that under the low signal, roughly half of the features selected come out as stable, however among these stable features very few features are relevant.

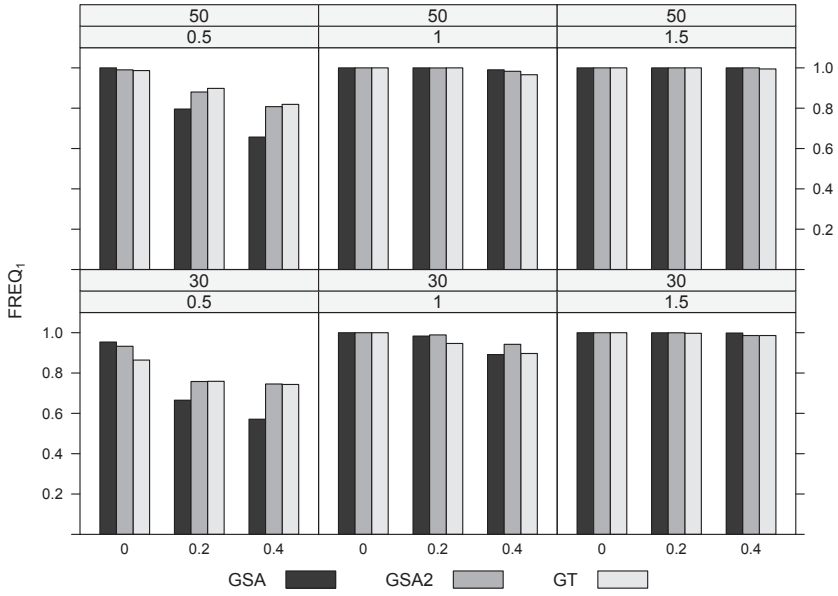

Fig. 6.8. Stability of feature selection based on prior domain knowledge: the $FREQ_1$ measure (formula 5.25) as a function of the signal strength ($\Delta = 0.5,\ 1,\ 1.5$), correlation among features ($\rho = 0,\ 0.2,\ 0.4$) and the sample size ($n = 30,\ 50$)

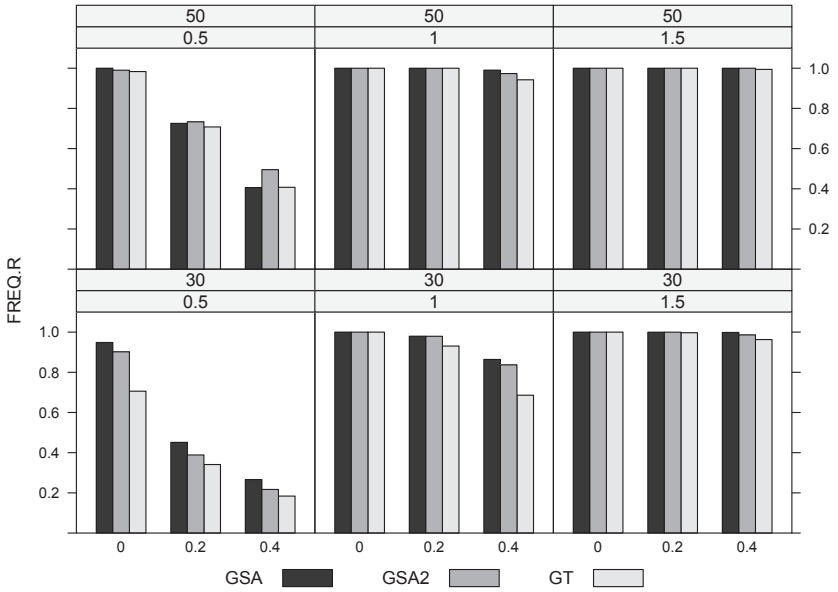

Fig. 6.9. Stability of feature selection based on prior domain knowledge: the $FREQ.R$ measure (formula 6.3) as a function of the signal strength ($\Delta = 0.5, 1, 1.5$), correlation among features ($\rho = 0, 0.2, 0.4$) and the sample size ($n = 30, 50$)

Interestingly, out of the three gene set analysis methods, the Globaltest consistently realizes the lowest $FREQ.R$ (Figure 6.9). This may be due to the fact that the Globaltest tends to be oversensitive if only single genes in a gene set realize high level of differential expression, as shown e.g. in Figures 4.2 or 4.3, bottom left panels, and discussed in section 4.4.2 on page 72. Therefore, an irrelevant gene set which by chance includes only 1-2 genes strongly associated with the target is more likely to be declared as significant by the Globaltest then by GSA or GSA2. The latter methods require that differentially expressed signal is demonstrated over a larger part of the gene set concerned.

6.3. Discussion and conclusions

Based on results of the numerical study, we draw the following conclusions regarding classification using standard feature selection and prior domain knowledge-based feature selection.

1. The general conclusion is that prior domain knowledge on relationships among features employed at the stage of features selection leads to significant improvement of predictive models in terms of the generalization error and stability of features. Obviously this conclusion is true providing that the database of pathways (which we use as representation of the prior domain knowledge) does include, at least partly, the *right* feature set which actually explains the difference between the groups of samples that we want to classify.
2. Considering different methods of gene set analysis employed as feature selectors, we conclude that the GSA2 should be recommended as the preferred method. This method is based on the popular GSA algorithm proposed by Efron and Tibshirani, however, it produces correct, interpretable p-values. This apparently leads to better stability of feature sets selected, when features are generated from *significant*, i.e. realizing p-value< 0.05, feature sets (pathways). On the other hand, self-contained gene set analysis methods such as the Globaltest seem to be too sensitive if only a few features in a gene set are differentially expressed, which leads to worse stability and generalization error.
3. Considering classification methods compared in this study, we recommend the nonparametric Method 1 as the preferable algorithm. This method is preferred as it (i) slightly outperforms the second best SVM classifier in terms of the generalization error, and (ii) provides user-interpretable signatures (formula 5.3) which represent similarity of the classified sample with profiles of gene expression in the classes compared.
4. Based on the comprehensive evaluation of the standard (data-driven) methods of feature selection, we conclude that the Recursive Feature Elimination (RFE) proposed by Guyon *et al.* (2002) outperforms univariate methods as well as shrinkage methods such as the Lasso or the Elastic net. The RFE offers best stability of feature selection which does not decrease with the correlation among features, unlike stability of shrinkage methods which remarkably worsens under correlation.
5. Analyzing significance of association of gene sets with the target, we observe that the p-values associated with the 'winning' gene set, returned by GSA, GT or GSA2 are *all* significant (i.e. p-value < 0.05). We report the raw p-values, prior to multiple testing adjustment, in Table 6.8, and the multiple testing corrected p-values in Table 6.7 (we used the Holm multiple testing adjustment). We observe that multiple testing corrected p-values

come out as insignificant for low signal and small sample sizes ($n = 30$) or for correlated data, which nicely coincides with the cases where we observe poor stability of feature selection. Hence our study clearly shows that significant *unadjusted* p-values are not indicative of the informative, stable feature sets. Therefore, it is essential that we use multiple testing adjusted p-values in the feature selection step of the algorithm (algorithm in section 5.3). This is especially important as the size of the databases containing a-priori given gene sets tends to grow which boost the risk that irrelevant feature sets are by chance declared as significant.

Chapter 7

Concluding remarks

In this monograph, we provided a comprehensive analysis of the problem of classification in high-dimensional data, where the number of samples in the training data is substantially smaller than the number of features. This problem arises in bioinformatics and is related to the analysis of data from high-throughput experiments in genomics or proteomics, such as data from gene expression studies. Although bioinformatics provides important context for the methods presented in this work, we envisage that the methods can be also applied in other areas where high-dimensional data plays an important role.

The main idea presented in this work is to employ prior domain knowledge in the process of building predictive models from high-throughput data. This idea is proposed as a solution to the fundamental problem of selection of relevant, informative and stable features from high-throughput data, if the analysis is done with purely data-driven methods. In this work, we proposed the methodology to use prior domain (e.g. biological) knowledge at the stage of feature selection, assuming that the domain knowledge is available as *a priori* defined sets of features which are expected to be functionally related.

Here we provide the summary of the most important results presented in this monograph.

- We provide analytical results related to the risk of selection of irrelevant features from high-throughput data when feature selection is realized with data-driven algorithms based on feature ranking. These results can be used to estimate the required sample size to guarantee that the relevant features are returned, rather than noisy, irrelevant and unstable features. Since in numerous applications (e.g. in gene expression studies) it is infeasible to gather enough samples, our results demonstrate the inherent limitations of the data-driven feature selection and thus motivate the proposed approach to include additional, *a priori* knowledge when dealing with high-throughput data.

- We provide a comprehensive methodological analysis of gene set analysis algorithms which we propose to use as the means for prior domain knowledge-based feature selection. In particular:
  - We analyzed the underlying models of the statistical experiment, (implicitly) assumed by the different algorithms. We grouped the algorithms into four distinct categories which differ in terms of the null hypothesis and interpretation of the p-values. We showed which of the algorithms produce meaningful results which indicate association of gene sets with the target. We also showed the algorithms whose significant p-values cannot be interpreted in terms of association of the gene sets with the target.
  - Since some of the algorithms, based on gene sampling, which fail to produce sound, interpretable results have been implemented in software tools and have gained popularity in bioinformatics, we propose a new interpretation of results of these algorithms. We showed that although the results cannot be interpreted in terms of statistical significance (p-value), they can have heuristic, biologically interpretable meaning.
  - We identified a flaw in the way the important gene set analysis algorithm GSA estimates significance. We proposed an improved version of this method with the modified procedure to estimate the p-value.
  - We provided a comprehensive empirical analysis of the effect of correlation in data on the size and power of different methods of gene set analysis.
- We proposed the algorithm for classification in high-dimensional data, where feature selection relies on prior domain knowledge about sets of related features. We propose the method to calculate per-sample signatures of gene set activation which are then used for classification of samples. We proposed nonparametric algorithms of classification of samples based on the signatures, which can be used as an (often more efficient) alternative to well known parametric classifiers.
- We proposed several measures which express stability (CV_{loo} stability) of data-driven and prior domain knowledge-based feature selection under small changes in data.
- We provide a comprehensive study in which we compared several data-driven methods and prior domain knowledge-based methods of feature selection in terms of the generalization error of classifiers as well as stability. We conclude that the proposed approach indeed results in improved generalization error and more stable features. Based on this study, we also formulate recommendations as to which of the data-driven and which of the prior domain knowledge-based methods proved most effective.

Finally, we want to discuss the limitations of this work as well as directions for extension of the proposed approach. Throughout this work, we assumed that the domain knowledge, which is required to stabilize feature selection, is available as a collection of *a priori* defined feature sets which group (functionally) related features. Examples of such collections could be signalling pathways or gene sets (in the context of bioinformatics), or sets of terms which are characteristic of some subject areas used for text categorization (in the context of text mining). As such, this representation is rather simple, or "flat", as it does not convey the information pertaining to the structure of the feature (gene) set, or relative importance of the members of the set. Although this assumption is justified by the current research and practice in the analysis of high-throughput studies in bioinformatics, a more structured approach to the organization of the domain knowledge would be interesting. For instance, the actual network of inter-gene relationships could be used (as shown in the structure of the signalling pathways), or hierarchical relationships between features could be considered (as done e.g. in (Meinshausen, 2008)).

Bibliography

Ackermann M., Strimmer K. (2009), *A general modular framework for gene set enrichment analysis*, BMC Bioinformatics, 10, 47.

Allison D.B., Cui X., Page G.P., Sabripour M. (2006), *Microarray data analysis: from disarray to consolidation and consensus*, Nature Reviews Genetics, 7(1), 55–65.

Al-Shahrour F., Diaz-Uriarte R., Dopazo J. (2005), *Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information*, Bioinformatics, 21(13), 2988–2993.

Al-Shahrour F., Arbiza L., Dopazo H., Huerta-Cepas J., Minguez P., Montaner D., Dopazo J. (2007), *From genes to functional classes in the study of biological systems*, BMC Bioinformatics, 8(1), 114.

Anders G., Maciejewski H., Jesus B., Remtulla F. (2006), *A comprehensive study of outage rates of air blast breakers*, IEEE Transactions on Power Systems, 21(1), 202–210.

Barry W.T., Nobel A.B., Wright F.A. (2005), *Significance analysis of functional categories in gene expression studies: a structured permutation approach*, Bioinformatics, 21(9), 1943–1949.

Basford K.E., McLachlan G.J., Rathnayake S.I. (2012), *On the classification of microarray gene-expression data*, Briefings in Bioinformatics, doi:10.1093/bib/bbs056, 1–9.

Benjamini Y., Hochberg Y. (1995), *Controlling the false discovery rate: A practical and powerful approach to multiple testing*, Journal of the Royal Statistical Society, Series B, 57, 289–300.

Benjamini Y., Yekutieli D. (2001), *The control of the false discovery rate in multiple testing under dependency*, Annals of Statistics, 29, 1165–1188.

Berenguel M., Klempous R., Maciejewski H., Nikodem J., Nikodem M., Valenzuela L. (2005a), *Explanatory analysis of data from a distributed solar collector field*, LNCS, 3643, 621–626.

Berenguel M., Cirre C.M., Klempous R., Maciejewski H., Nikodem M., Nikodem J., Rudas I., Valenzuela L. (2005b), *Hierarchical control of a distributed solar collector field*, LNCS, 3643, 614–620.

Bild A.H., Yao G., Chang J.T., Wang Q., Potti A., Chasse D., Joshi M.B., Harpole D., Lancaster J.M., Berchuck A., Olson J.A., Marks J.R., Dressman H.K., West M., Nevins J.R. (2005), *Oncogenic pathway signatures in human cancers as a guide to targeted therapies*, Nature, 439(7074), 353–357.

Bishop C. (1995), *Neural Networks for Pattern Recognition*. Claredon Press, Oxford.

Bittner M., Meltzer P., Chen Y., ... , Trent J. (2000), *Molecular classification of cutaneous malignant melanoma by gene expression profiling*, Nature, 406(6795), 536–540.

Bo T.H., Jonassen I. (2002), *New feature subset selection procedures for classification of expression profiles*, Genome Biology, 3(4).

Boorsma A., Foat B.C., Vis D., Klis F., Bussemaker H.J. (2005), *T-profiler: scoring the activity of predefined groups of genes using gene expression data*, Nucleic Acids Research, 33 (suppl 2), W592–W595.

Breslin T., Eden P., Krogh M. (2004), *Comparing functional annotation analyses with Catmap*, BMC Bioinformatics, 5(1), 193.

Chiaretti S., Li X., Gentleman R., Vitale A., Vignetti M., Mandelli F., Ritz J., Foa R. (2004), *Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival*, Blood, 103(7), 2771–2778.

Cho S.B., Won H.H. (2003), *Machine Learning in DNA Microarray Analysis for Cancer Classification*, In: Proceedings of the First Asia-Pacific Bioinformatics Conference, Adelaide, Australia, 2003.

Clarke R., Ressom H.W., Wang A., Xuan J., Liu M.C., Gehan E.A., Wang Y. (2008), *The properties of high-dimensional data spaces: implications for exploring gene protein and expression data*, Nature Reviews Cancer, 8(1), 37–49.

Davis C.A., Gerick F., Hintermair V., Friedel C.C., Fundel K., Küffner R., Zimmer R. (2006), *Reliable gene signatures for microarray classification: assessment of stability and performance*, Bioinformatics, 22(19), 2356–2363.

Devroye L., Györfi L., Lugosi G. (1996), *A probabilistic theory of pattern recognition*, Springer.

Dinu I., Potter J.D., Mueller T., Liu Q., Adewale A.J., Jhangri G.S., Einecke G., Famulski K.S., Halloran P., Yasui Y. (2007), *Improving gene set analysis of microarray data by SAM-GS*, BMC Bioinformatics, 8, 242.

Dinu I., Liu Q., Potter J.D., Adewale A.J., Jhangri G.S., Mueller T., Einecke G., Famulski K.S., Halloran P., Yasui Y. (2008), *A Biological Evaluation of Six Gene Set Analysis Methods for Identification of Differentially Expressed Pathways in Microarray Data*, Cancer Informatics, 6, 357–368.

Draminski M., Rada-Iglesias A., Enroth S., Wadelius C., Koronacki J., Komorowski J. (2008), *Monte Carlo feature selection for supervised classification*, Bioinformatics, 24(1), 110–117.

Draminski M., Kierczak M., Koronacki J., Komorowski J. (2010), *Monte Carlo feature selection and interdependency discovery in supervised classification*, In: *Advances in Machine Learning II*, 371–385, Springer Berlin Heidelberg.

Dudoit S., Fridlyand J., Speed T.P. (2002a), *Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data*, Journal of the American Statistical Association, 97(457), 77–87.

Dudoit S., Yang Y.H., Callow M.J., Speed T.P. (2002b), *Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments*, Statistica Sinica, 12, 111–139.

Dudoit S., Shaffer J.P., Boldrick J.C. (2003), *Multiple Hypothesis Testing in Microarray Experiments*, Statistical Science, 18(1), 71–103.

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.

Edelman E., Porrello A., Guinney J., Balakumaran B., Bild A., Febbo P.G., Mukherjee S. (2006), *Analysis of sample set enrichment scores: assaying the enrichment of sets of genes for individual samples in genome-wide expression profiles*, Bioinformatics, 22(14), e108–e116.

Efron B., Tibshirani R. (2006), *On testing the significance of sets of genes*, Stanford tech report rep 2006. http://www-stat.stanford.edu/ tibs/ftp/GSA.pdf.

Efron B., Tibshirani R. (2007), *On testing the significance of sets of genes*, The Annals of Applied Statistics, 1(1), 107–129.

Efron B. (2007), *Size, power and false discovery rates*, The Annals of Statistics, 35(4), 1351–1377.

Efron B. (2008), *Simultaneous inference: when should hypothesis testing problems be combined?*, The Annals of Statistics, 2(1), 197–223.

Ein-Dor L., Kela I., Getz G., Givol D., Domany E. (2005), *Outcome signature genes in breast cancer: is there a unique set?*, Bioinformatics, 21(2), 171–178.

Ein-Dor L., Zuk O., Domany E. (2006), *Thousands of samples are needed to generate a robust gene list for predicting outcome of cancer*, Proceedings of the National Academy of Sciences, 103(15), 5923–5928.

Eisen M.B., Spellman P.T., Brown P.O., Botstein D. (1998), *Cluster analysis and display of genome-wide expression patterns*, Proceedings of the National Academy of Sciences, 95(25), 14863–14868.

Fisher R.A. (1915), *Frequency distribution of the values of correlation coefficient in samples from an indefinitely large population*, Biometrica, 10(4), 507–521.

Fisher R.A. (1921), *On the "probable error" of a coefficient of correlation deduced from a small sample*, Metron, 1, 3–32.

Franiak-Pietryga I., Maciejewski H., Wołowiec D., Sałagacka A., Błoński J.Z., Janus A., Kotkowska A., Wawrzyniak E., Ghia P., Mirowski M., Robak T., Korycka-Wołowiec A. (2012a), *Changes in the apoptotic gene expression profile in CLL patients treated with rituximab combined with cladribine and cyclophosphamide – preliminary results*, Leukemia Research, 36(9), 1134–1140.

Franiak-Pietryga I., Sałagacka A., Maciejewski H., Błoński J.Z., Borowiec M., Mirowski M., Robak T., Korycka-Wołowiec A. (2012b), *Apoptotic gene expression under influence of fludarabine and cladribine in chronic lymphocytic leukemia – microarray study*, Pharmacological Reports, 64(2), 412–420.

Fridley B.L., Jenkins G.D., Biernacka J.M. (2010), *Self-Contained Gene-Set Analysis of Expression Data: An Evaluation of Existing and Novel Methods*, PLoS One, 5(9), e12693.

Fujarewicz K., Wiench M. (2003), *Selecting differentially expressed genes for colon tumor classification*, International Journal of Applied Mathematics and Computer Science, 13(3), 327–336.

Fujarewicz K., Kimmel M., Rzeszowska-Wolny J., Świerniak A. (2003), *A note on classification of gene expression data using support vector machines*, Journal of Biological Systems, 11(01), 43–56.

Fujarewicz K., Jarząb M., Eszlinger M., Krohn K., Paschke R., Oczko-Wojciechowska M., Wiench M., Kukulska A., Jarząb B., Świerniak A. (2007), *A multi-gene approach to differentiate papillary thyroid carcinoma from benign lesions: gene selection using support vector machines with bootstrapping*, Endocrine-related Cancer, 14(3), 809–826.

Geman D., d'Avignon C., Naiman D., Winslow R. (2004), *Classifying Gene Expression Profiles from Pairwise mRNA Comparisons*, Statistical Applications in Genetics and Molecular Biology, 3(1).

Gentleman R. (2012), *Using categories to model genomic data*, Vignette of Category Bioconductor package, http://www.bioconductor.org/packages/release/bioc/vignettes/Category/inst/doc/Category.pdf.

Goeman J.J., van de Geer S.A., de Kort F., van Houwelingen H.C. (2004), *A global test for groups of genes: testing association with clinical outcome*, Bioinformatics, 20(1), 93–99.

Goeman J.J., Oosting J., Cleton-Jansen A.M., Anninga J.K., van Houwelingen H.C. (2005), *Testing association of a pathway with survival using gene expression data*, Bioinformatics, 21(9), 1950–1957.

Goeman J.J., Bühlmann P. (2007), *Analyzing gene expression data in terms of gene sets: methodological issues*, Bioinformatics, 23(8), 980–987.

Golub T.R., Slonim D.K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J.P., Coller H., Loh M.L., Downing J.R., Caligiuri M.A., Bloomfield C.D., Lander E.S. (1999), *Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring*, Science, 286, 531–537.

Greblicki W. (1978), *Asymptotically optimal pattern recognition procedures with density estimates (Corresp.)*, Information Theory, IEEE Transactions on, 24(2), 250–251.

Greblicki W. (1980), *Learning to recognize patterns with a probabilistic teacher*, Pattern Recognition, 12(3), 159–164.

Greblicki W., Pawlak M. (1987), *Necessary and sufficient conditions for Bayes risk consistency of a recursive kernel classification rule (Corresp.)*, Information Theory, IEEE Transactions on, 33(3), 408–412.

Guyon I., Weston J., Barnhill S. (1999), *Gene selection for cancer classification using Support Vector Machines*, Machine Learning, 46, 389–422.

Guyon I., Elisseeff A. (2003), *An Introduction to Variable and Feature Selection*, Journal of Machine Learning Research, 3, 1157–1182.

Hastie T., Tibshirani R., Friedman J. (2001), *The Elements of Statistical Learning*, Springer.

Hochberg Y. (1988), *A sharper Bonferroni procedure for multiple tests of significance*, Biometrika, 75, 800–802.

Holm S. (1979), *A simple sequentially rejective multiple test procedure*, Scand. J. Statist., 6 65–70.

Hung J.H., Yang T.H., Hu Z., Weng Z., DeLisiet Ch. (2012), *Gene set enrichment analysis: performance evaluation and usage guidelines*, Briefings in Bioinformatics, 13(3), 281–291.

Irizarry R.A., Wang C., Zhou Y., Speed T.P. (2009), *Gene Set Enrichment Analysis Made Simple*, Statistical Methods in Medical Research, 18(6), 565–575.

Jarząb B., Wiench M., Fujarewicz K., ... , Świerniak A. (2005), *Gene expression profile of papillary thyroid cancer: sources of variability and diagnostic implications*, Cancer research, 65(4), 1587–1597.

Jiang Y., Gentleman R. (2007), *Extensions to gene set enrichment*, Bioinformatics, 23(3), 306–313.

Joshi A., Van de Peer Y., Michoel T. (2008), *Analysis of a Gibbs sampler method for model-based clustering of gene expression data*, Bioinformatics, 24(2), 176–183.

Kalousis A., Prados J., Hilario M. (2007), *Stability of feature selection algorithms: a study on high-dimensional spaces*, Knowledge and Information Systems, 12(1), 95–116.

Kim S.Y., Volsky D.J. (2005), *PAGE: Parametric Analysis of Gene Set Enrichment*, BMC Bioinformatics, 6, 144.

Kohavi R., John G.H. (1997), *Wrappers for feature subset selection*, Artificial Intelligence, 97, 273–324.

Kong S.W., Pu W.T., Park P.J. (2006), *A multivariate approach for integrating genome-wide expression data and biological knowledge*, Bioinformatics, 22(19), 2373–2380.

Król M., Pawłowski K.M., Szyszko K., Maciejewski H., Dolka I., Manuali E., Jank M., Motyl T. (2012), *The gene expression profiles of canine mammary cancer cells grown with carcinoma-associated fibroblasts (CAFs) as a co-culture in vitro*, BMC Veterinary Research, 8, 35.

Lai C., Reinders M.J.T., van't Veer L.J., Wessels L.F.A. (2006), *A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets*, BMC Bioinformatics, 7, 235.

Li L., Weinberg C.R., Darden T.A., Pedersen L.G. (2001), *Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method*, Bioinformatics, 17(12), 1131–1142.

Li L., Umbach D.M., Terry P., Taylor J.A. (2004), *Application of the GA/KNN method to SELDI proteomics data*, Bioinformatics, 20(10), 1638–1640.

Li T., Zhang C., Ogihara M. (2004), *A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression*, Bioinformatics, 20(15), 2429–2437.

Liu Q., Dinu I., Adewale A.J., Potter J.D., Yasui Z. (2007), *Comparative evaluation of gene set analysis methods*, BMC Bioinformatics, 8, 431.

Lossos I.S., Czerwinski D.K., Alizadeh A.A., Wechser M.A., Tibshirani R., Botstein D., Levy R. (2004), *Prediction of survival in diffuse large-B-cell lymphoma based on the expression of six genes*, New England Journal of Medicine, 350(18), 1828–1837.

Łabaj P.P., Leparc G.G., Linggi B.E., Markillie L.M., Wiley H.S., Kreil D.P. (2011), *Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling*, Bioinformatics, 27(13), i383–i391.

Maciejewski H., Jasińska A. (2005), *Clustering DNA microarray data*, In: *Computer Recognition Systems, Springer Advances in Soft Computing*, 595–601.

Maciejewski H., Konarski Ł., Jasińska A., Drath M. (2005), *Analysis of DNA Microarray Data: Methods and Tools*, [in Polish], Bio-Algorithms and Med-Systems, 1(1/2), 129–131.

Maciejewski H. (2007), *Adaptive selection of feature set dimensionality for classification of DNA microarray samples*, In: *Computer Recognition Systems, Springer Advances in Soft Computing*, 831–837.

Maciejewski H. (2008a), *Quality of feature selection based on microarray gene expression data*, LNCS, 5103, 140–147.

Maciejewski H. (2008b), *Predictive performance of top differentially expressed genes in microarray gene expression studies*, In: *Information Technologies in Biomedicine, Springer Advances in Soft Computing*, 395–402.

Maciejewski H., Valenzuela L., Berenguel M., Fernandez-Reche J., Adamus K., Jarnicki M. (2008), *Analyzing solar power plant performance through data mining*, Journal of Solar Energy Engineering – Transaction of the ASME, 130(4), 44503-1–44503-3.

Maciejewski H., Caban D. (2008), *Estimation of repairable system availability within fixed time horizon*, Reliability Engineering and Systems Safety, 93(1), 100–106.

Maciejewski H., Twaróg P. (2009), *Model instability in microarray gene expression class prediction studies*, LNCS, 5717, 745–752.

Maciejewski H. (2011a), *Competitive and self-contained gene set analysis methods applied for class prediction*, In: Proc. of the Federated Conference on Computer Science and Information Systems, Sept. 18–21, 2011, IEEE Computer Society Press, 55–61.

Maciejewski H. (2011b), *Class prediction in microarray studies based on activation of pathways*, LNCS, LNAI, 6678, 321–328.

Maciejewski H. (2012), *Feature selection based on activation of signaling pathways applied for classification of samples in microarray studies*, LNCS, LNAI, 7268, 284–292.

Maciejewski H. (2013), *Gene set analysis methods: statistical models and methodological differences*, Briefings in Bioinformatics, doi:10.1093/bib/bbt002, 1–15.

MAQC Consortium (2006), *The MicroArray Quality Control (MAQC) project shows inter-and intraplatform reproducibility of gene expression measurements*, Nature Biotechnology, 24(9), 1151–1161.

Meinshausen N. (2008), *Hierarchical testing of variable importance*, Biometrika, 95(2), 265–278.

Mansmann U., Meister R. (2005), *Testing differential gene expression in functional groups – Goeman's global test versus an ANCOVA approach*, Methods Inf. Med., 44, 449–453.

Marincevic M., Mansouri M., Kanduri M., Isaksson A., Göransson H., Smedby K.E., Jurlander J., Juliusson G., Davi F., Stamatopoulos K., Rosenquist R. (2010), *Distinct gene expression profiles in subsets of chronic lymphocytic leukemia expressing stereotyped IGHV4-34 B-cell receptors*, Haematologica, 95(12), 2072–2079.

Markowetz F., Spang R. (2005), *Molecular diagnosis. Classification, Model Selection and Performance Evaluation*, Methods Inf. Med., 44, 438–443.

Miklos G.L.G., Maleszka R. (2004), *Microarray reality checks in the context of a complex disease*, Nature Biotechnology, 22(5), 615–621.

Mootha V.K., Lindgren C.M., Eriksson K.F., *et al.* (2003), *PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes*, Nature Genetics, 34(3), 267–273.

Mukherjee S., Niyogi P., Poggio T., Rifkin R. (2002), *Statistical learning: Stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization*, AI Memo 2002-024, Massachusetts Institute of Technology, Revised 2003.

Mukherjee S., Niyogi P., Poggio T., Rifkin R. (2006), *Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization*, Advances in Computational Mathematics, 25(1), 161–193.

Nam D., Kim S.Y. (2008), *Gene-set approach for expression pattern analysis*, Briefings in Bioinformatics, 9(3), 189–197.

Nelson P.S. (2004), *Predicting prostate cancer behavior using transcript profiles*, Journal of Urology, 172, S28–S33.

Newton M.A., Quintana F.A., den Boon J.A., Sengupta S., Ahlquist P. (2007), *Random set methods identify distinct aspects of the enrichment signal in gene-set analysis*, Annals of Applied Statistics, 1(1), 85–106.

Ooi C.H., Tan P. (2003), *Genetic algorithms applied to multi-class prediction for the analysis of gene expression data*, Bioinformatics, 19(1), 37–44.

Ostrzeszewicz M., Maciejewski H., Sapierzyński R., Micuń J., Majewska A., Lechowski R., Motyl T., Jank M. (2012), *Mismatch between transcriptomic and histopathologic picture of canine lymphomas*, Polish Journal of Veterinary Sciences, 15(4), 781–790.

Patterson T.A., Lobenhofer E.K., Fulmer-Smentek S.B., ... , Wolfinger R.D. (2006), *Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project*, Nature biotechnology, 24(9), 1140–1150.

Pavlidis P., Lewis D.P., Noble W.S. (2002), *Exploring gene expression data with class scores*, Pac Symp Biocomput, 474–485.

Pavlidis P., Qin J., Arango V., Mann J.J., Sibille E. (2004), *Using the gene ontology for microarray data mining: a comparison of methods and application to age effect in human prefrontal cortex*, Neurochemical Research, 29(6), 1213–1222.

Pawłowski K.M., Maciejewski H., Dolka I., Mol J.A., Motyl T., Król M. (2013a), *Gene expression profiles in canine mammary carcinomas of various grades of malignancy*, BMC Veterinary Research, 9:78, doi:10.1186/1746-6148-9-78.

Pawłowski K.M., Maciejewski H., Majchrzak K., Dolka I., Mol J.A., Motyl T., Król M. (2013b), *Five markers useful for the distinction of canine mammary malignancy*, BMC Veterinary Research, 9(1), 138, doi:10.1186/1746-6148-9-138.

Peng S., Xu Q., Ling X.B., Peng X., Du W., Chen L. (2003), *Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines*, FEBS letters, 555, 358–362.

Poggio T., Smale S. (2003), *The mathematics of learning: Dealing with data*, Notices of the AMS, 50(5), 537–544.

Poggio T., Rifkin R., Mukherjee S., Niyogi P. (2004), *General conditions for predictivity in learning*, Nature, 428(6981), 419–422.

Qin Z.S. (2006), *Clustering microarray gene expression data using weighted Chinese restaurant process*, Bioinformatics, 22(16), 1988–1997.

Ramaswamy S., Ross K.N., Lander E.S., Golub T.R. (2002), *A molecular signature of metastasis in primary solid tumors*, Nature Genetics, 33(1), 49–54.

Robbins K.R., Zhang W., Bertrand J.K., Rekaya R. (2007), *The ant colony algorithm for feature selection in high-dimension gene expression data for disease classification*, Mathematical Medicine and Biology, 24(4), 413–426.

Rogalińska M., Franiak-Pietryga I., Błoński J.Z., Góralski P., Maciejewski H., Janus A., Robak P., Mirowski M., Piekarski H., Robak T., Kiliańska Z.M. (2013), *Towards personalized therapy for chronic lymphocytic leukemia: DSC and cDNA microarray assessment of two cases*, Cancer Biology & Therapy, 14(1), 6–12.

Saeys Y., Inza I., Larranaga P. (2007), *A review of feature selection techniques in bioinformatics*, Bioinformatics, 23(19), 2507–2517.

Scheer M., Klawonn F., Münch R., Grote A., Hiller K., Choi C., Koch I., Schobert M., Härtig E., Klages U., Jahn D. (2006), *JProGO: a novel tool for the functional interpretation of prokaryotic microarray data using Gene Ontology information*, Nucleic acids research, 34 (suppl 2), W510–W515.

Scott D.W., Thompson J.R. (1983), *Probability density estimation in higher dimensions*, In: *Computer Science and Statistics: Proceedings of the Fifteenth Symposium on the Interface*, vol. 528, 173–179, North-Holland, Amsterdam.

Shen Y., Lin Z., Zhu J. (2011), *Shrinkage-based regularization tests for high-dimensionality data with application to gene set analysis*, Computational Statistics and Data Analysis, 55, 2221–2233.

Silverman B.W. (1986), *Density Estimation*, Chapman and Hall, London.

Simek K., Fujarewicz K., Świerniak A., Kimmel M., Jarząb B., Wiench M., Rzeszowska J. (2004), *Using SVD and SVM methods for selection, classification, clustering and modeling of DNA microarray data*, Engineering Applications of Artificial Intelligence, 17(4), 417–427.

Simon R., Radmacher M.D., Dobbin K., McShane L.M. (2003), *Pitfalls in the Use of DNA Microarray Data for Diagnostic and Prognostic Classification*, Journal of the National Cancer Institute, 95(1), 14–18.

Simon R. (2003), *Supervised analysis when the number of candidate features (p) greatly exceeds the number of cases (n)*, ACM SIGKDD Explorations Newsletter, 5(2), 31–36.

Singh D., Febbo P.G., Jackson D.G., Manola J., Ladd C., Tamayo P., Renshaw A.A., D'Amico A.V., Richie J.P., Lander E.S., Loda M., Kantoff P.W., Golub T.R., Sellers W.R. (2002), *Gene expression correlates of clinical prostate cancer behavior*, Cancer Cell, 1(2), 203–209.

Sobczak W., Malina W. (1985), *Methods of data selection*, Warsaw, WNT [in Polish].

Sobieszczańska B., Kasprzykowska U., Turniak M., Maciejewski H., Franiczek R., Duda-Madej A. (2012), *Virulence genes profiles and phylogenetic origin of Escherichia coli from acute and chronic intestinal diseases revealed by comparative genomic hybridization microarray*, Polish Journal of Microbiology, 61(2), 105–110.

Song S., Black M.A. (2008), *Microarray-based gene set analysis: a comparison of current methods*, BMC Bioinformatics, 9, 502.

Sørlie T., Perou C.M., Tibshirani R., ... , Borresen-Dale A.L. (2001), *Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications*, Proceedings of the National Academy of Sciences, 98(19), 10869–10874.

Sørlie T., Tibshirani R., Parker J., ... , Botstein D. (2003), *Repeated observation of breast tumor subtypes in independent gene expression data sets*, Proceedings of the National Academy of Sciences, 100(14), 8418–8423.

Statnikov A., Wang L., Aliferis C.F. (2008), *A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification*, BMC Bioinformatics, 9, 319.

Sturn A., Quackenbush J., Trajanoski Z. (2002), *Genesis: cluster analysis of microarray data*, Bioinformatics, 18(1), 207–208.

Subramanian A., Tamayo P., Mootha V.K., Mukherjee S., Ebert B.L., Gillette M.A., Paulovich A., Pomeroy S.L., Golub T.R., Lander E.S., Mesirov J.P. (2005), *Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles*, Proceedings of the National Academy of Sciences, 102(43), 15545–15550.

Subramanian J., Simon R. (2010), *Gene Expression–Based Prognostic Signatures in Lung Cancer: Ready for Clinical Use?*, Journal of the National Cancer Institute, 102(7), 464–474.

Szmit S., Jank M., Maciejewski H., Grabowski M., Główczyńska R., Majewska A., Filipiak K., Motyl T., Opolski G. (2010), *Gene expression profiling in peripheral blood nuclear cells in patients with refractory ischaemic end-stage heart failure*, Journal of Applied Genetics, 51(3), 353–368.

Szmit S., Jank M., Maciejewski H., Balsam P., Majewska A., Loj M., Grabowski M., Filipiak K., Motyl T., Opolski G. (2012), *White blood cell transcriptome correlates with renal function in acute heart failure*, International Heart Journal, 53(2), 117–124.

Tamayo P., Slonim D., Mesirov J., Zhu Q., Kitareewan S., Dmitrovsky E., Lander E.S., Golub T.R. (1999), *Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation*, Proceedings of the National Academy of Sciences, 96(6), 2907–2912.

Tamayo P., Steinhardt G., Liberzon A., Mesirov J.P. (2011), *Gene Set Enrichment Analysis Made Right*, Preprint arXiv:1110.4128, DOI: 10.1016/j.jbi.2011.12.002.

Theodoridis S., Koutroumbas K. (2006), *Pattern Recognition, Third Edition*, Elsevier.

Tian L., Greenberg S.A., Kong S.W., Altschuler J., Kohane I.S., Park P.J. (2005), *Discovering statistically significant pathways in expression profiling studies*, Proceedings of the National Academy of Sciences, 102(38), 13544–13549.

Tibshirani R. (1996), *Regression Shrinkage and Selection via the Lasso*, Journal of the Royal Statistical Society. Series B, 58(1), 267–288.

Troyanskaya O., Garber M.E., Brown P.O., Botstein D., Altman R.B. (2002), *Nonparametric methods for identifying differentially expressed genes in microarray data*, Bioinformatics, 18(11), 1454–1461.

Tsai C.A., Chen J.J. (2009), *Multivariate analysis of variance test for gene set analysis*, Bioinformatics, 25(7), 897–903.

Tusher V.G., Tibshirani R., Chu G. (2001), *Significance analysis of microarrays applied to the ionizing radiation response*, Proceedings of the National Academy of Sciences, 98(9), 5116–5121.

van't Veer L.J., Dai H., van de Vijver M.J., *et al.* (2002), *Gene expression profiling predicts clinical outcome of breast cancer*, Nature, 415(31), 530–536.

Vapnik V.N., Chervonenkis A.J. (1991), *The necessary and sufficient conditions for consistency of the method of empirical risk*, Pattern Recognition and Image Analysis, 1(3), 284–305.

Vapnik V.N. (1999), *An overview of statistical learning theory*, IEEE Transactions on Neural Networks, 10(5), 988–999.

Volinia S., Evangelisti R., Francioso F., Arcelli D., Carella M., Gasparini P. (2004), *GOAL: automated Gene Ontology analysis of expression profiles*, Nucleic Acids Research, 32 (suppl 2), W492–W499.

Walkowicz E., Unold O., Maciejewski H., Skrobanek P. (2011), *Zoometric indices in Silesian horses in the years 1945–2005*, Annals of Animal Science, 11(4), 555–565.

Walkowicz E., Unold O., Maciejewski H., Skrobanek P. (2013), *Influence of Schweres Warmblut breed stallions on Silesian breed horses exterior*, Journal of Animal and Feed Sciences, in press.

West M., Blanchette C., Dressman H., Huang E., Ishida S., Spang R., Zuzan H., Olson Jr. J.A., Marks J.R., Nevins J.R. (2001), *Predicting the clinical status of human breast cancer using gene expression profiles*, Proceedings of the National Academy of Sciences, 98, 11462–11467.

Wibisono A., Rosasco L., Poggio T. (2009), *Sufficient Conditions for Uniform Stability of Regularization Algorithms*, Computer Science and Artificial Intelligence Laboratory Technical Report, MIT-CSAIL-TR-2009-060.

Wieteska-Skrzeczyńska W., Grzelkowska-Kowalczyk K., Jank M., Maciejewski H. (2009), *Transcriptional dysregulation of skeletal muscle protein metabolism in streptozotocin-diabetic mice*, Journal of Physiology and Pharmacology, 60(1), 29–36.

Wu M.C., Lin X. (2009), *Prior biological knowledge-based approaches for the analysis of genome-wide expression profiles using gene sets and pathways*, Statistical Methods in Medical Research, 18(6), 577–593.

Xu L., Tan A., Naiman D., Geman D., Winslow R. (2005), *Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data*, Bioinformatics, 21(20), 3905–3911.

Yang Y., Pedersen J.O. (1997), *A comparative study on feature selection in text categorization*, In: ICML, 97, 412–420.

Zhu J., Hastie T. (2004), *Classification of gene microarrays by penalized logistic regression*, Biostatistics, 5(3), 427–443.

Zou H., Hastie T. (2005), *Regularization and variable selection via the elastic net*, Journal of the Royal Statistical Society. Series B, 67(2), 301–320.