A Multivariate Framework for Variable Selection and Identification of Biomarkers in High-Dimensional Omics Data

Von der Fakultät für Mathematik und Informatik
der Universität Leipzig
angenommene

DISSERTATION

zur Erlangung des akademischen Grades

DOCTOR RERUM NATURALIUM

(Dr. rer. nat.)

im Fachgebiet
Informatik

vorgelegt
von M.Sc. Verena Zuber
geboren am 6. Juni 1983 in Donauwörth

Die Annahme der Dissertation wurde empfohlen von

1. Professor Dr. Jörg Rahnenführer (Dortmund)

2. Professor Dr. Peter F. Stadler (Leipzig)

Die Verleihung des akademischen Grades erfolgt mit Bestehen
der Verteidigung am 27.06.2012 mit dem Gesamtprädikat magna cum laude.

Acknowledgments

Mein größter Dank geht an Korbinian Strimmer für die ausgezeichnete Betreuung. Im Besonderen möchte ich mich vielmals bedanken, dass seine Tür jederzeit für jegliche erdenkliche Frage offen stand. Nicht vergessen möchte ich zudem die äußerst hilfreichen Anregungen und Förderung zum kreativen und selbstständigen Arbeiten.

Darüber hinaus geht mein Dank an Professor Peter Stadler und an Professor Jörg Rahnenführer für die Korrektur der Arbeit.

Moreover, a huge "dankeschön" goes to my research group and its visitors for support, discussion, and advice. Thank you Bernd, Carsten, Miika, and Sebastian. Des Weiteren möchte ich meinen Kollegen am IMISE danken. Insbesondere geht mein Dank an Professor Markus Löffler für die Unterstützung, Cornelia Will für die aufbauenden Worte, Dirk Hasenclever für seine statistische Expertise, Holger Kirsten und Peter Ahnert für den biologischen Rat und Katja Rösch für den mathematischen Rat.

Meinen Eltern möchte ich danken für ihre anhaltende Unterstützung. Finally, big hugs to Deb and Yasmine for their support, advice, good vibes, and enthusiasm for science.

Und zu guter Letzt ein großes Dankeschön an den wunderschönen Kater Justus, der so lieb war, sein Fell für den Umschlag photographieren zu lassen, und natürlich an Ben, den stolzen Katzen-Besitzer.

Abstract

In this thesis, we address the identification of biomarkers in high-dimensional omics data. The identification of valid biomarkers is especially relevant for personalized medicine that depends on accurate prediction rules. Moreover, biomarkers elucidate the provenance of disease, or molecular changes related to disease. From a statistical point of view the identification of biomarkers is best cast as variable selection. In particular, we refer to variables as the molecular attributes under investigation, e.g. genes, genetic variation, or metabolites; and we refer to observations as the specific samples whose attributes we investigate, e.g. patients and controls. Variable selection in high-dimensional omics data is a complicated challenge due to the characteristic structure of omics data. For one, omics data is high-dimensional, comprising cellular information in unprecedented details. Moreover, there is an intricate correlation structure among the variables due to e.g internal cellular regulation, or external, latent factors. Variable selection for uncorrelated data is well established. In contrast, there is no consensus on how to approach variable selection under correlation.

Here, we introduce a multivariate framework for variable selection that explicitly accounts for the correlation among markers. In particular, we present two novel quantities for variable importance: the correlation-adjusted t (CAT) score for classification, and the correlation-adjusted (marginal) correlation (CAR) score for regression. The CAT score is defined as the Mahalanobis-decorrelated t-score vector, and the CAR score as the Mahalanobis-decorrelated correlation between the predictor variables and the outcome. We derive the CAT and CAR score from a predictive point of view in linear discriminant analysis and regression; both quantities assess the weight of a decorrelated and standardized variable on the prediction rule. Furthermore, we discuss properties of both scores and relations to established quantities. Above all, the CAT score decomposes Hotelling's T^2 and the CAR score the proportion of variance explained. Notably, the decomposition of total variance into explained and unexplained variance in the linear model can be rewritten in terms of CAR scores.

To render our approach applicable on high-dimensional omics data we devise an efficient algorithm for shrinkage estimates of the CAT and CAR score. Subsequently, we conduct extensive simulation studies to investigate the performance of our novel approaches in ranking and prediction under correlation. Here, CAT and CAR scores consistently improve over marginal approaches in terms of more true positives selected and a lower model error. Finally, we illustrate the application of CAT and CAR score on real omics data. In particular, we analyze genomics, transcriptomics, and metabolomics data. We ascertain that CAT and CAR score are competitive or outperform state of the art techniques in terms of true positives detected and prediction error.

Contents

Chapter 1

Introduction

1.1 Biomarkers and personalized medicine

Personalized medicine is one of the great promises of modern clinical medicine (Hamburg and Collins, 2010). The aim is to tailor therapies to an individual patient, finding "the right drug for the right person" (Allison, 2008). Many diseases, including types of cancers, exhibit heterogeneous characteristics in clinical outcome or responsiveness to drug therapy. A correct diagnosis of such specific subtypes of cancer aids to administer the ideal targeted therapy. For example, scientists identified a molecular pattern in women suffering from breast cancer that can be used to quite accurately predict recurrence of cancer after surgery. This diagnostic test helps to decide whether the patient needs to undergo continuing chemotherapy (Paik et al., 2004). Personalized medicine reduces costs by improving the clinical success rate of the therapy prescribed (Woodcock, 2007) and thus can be beneficial for the patient as well as for the health care system. Moreover, pharmaceutical companies are able to increase the efficiency of their drugs if they can exactly predict which patients respond to the drug. This has prompted some pharmaceutical companies to develop drug/diagnostic pairs (Allison, 2008).

The success of personalized medicine decisively depends on the accuracy of the diagnosis (Hamburg and Collins, 2010). Thus, the discovery of precise *biomarkers for disease* is essential. Over the last two decades biotechnological inventions, like the microarray, sequencing technologies, or mass-spectrometry, have provided unprecedented information on biological and molecular processes. Such techniques enable comprehensive views on all constituents of the cell, ranging from the genetic code, to protein synthesis, and the metabolism. Derived from the Greek word for “all-encompassing” *omics* has been coined as a general term for this emerging data since it can literally comprise all constituents. For example, modern microarrays can measure the activity of all $25,000$ genes known in human. In particular, the microarray technology has stimulated the search for molecular biomarkers. See e.g the reference publications by Golub et al. (1999), Ramaswamy et al.

(2001), or Veer et al. (2002) that propagate molecular cancer diagnosis. Apart from prediction, biomarkers can also provide insight into the heterogeneity and molecular changes of disease states (Schilsky, 2010).

1.2 The search for biomarkers as statistical problem

From a statistical point of view the discovery of biomarkers is best cast as *variable selection*. In particular, we refer to variables as the molecular attributes under investigation, e.g. genes, genetic variation, or metabolites; and we refer to observations as the specific samples whose attributes we investigate, e.g. patients and controls.

Variable selection in omics data poses an intricate challenge due to the characteristic data structure of omics data. First, omics data is *high-dimensional*, literally comprising all constituents. Unfortunately, the limitation in omics data is the sample size that has not expanded with the same speed as the dimension of variables. Second, certain processes and elements of the cell are interconnected in complex patterns due to e.g internal cellular regulation or external, latent factors influencing the cell. This results in an *intricate correlation structure* among the variables. Variable selection for uncorrelated data is well established. In contrast, there is no consensus on how to approach *variable selection under correlation*.

1.3 Contributions

This thesis illustrates our attempt to incorporate knowledge on the correlation structure into the selection of variables. Since omics data exhibit an intrinsic correlation structure among variables, we argue that it is beneficial to incorporate this information in the selection of variables. In particular, we propose two novel quantities for variable selection that explicitly model the correlation structure, the correlation-adjusted t (CAT) score in classification and the correlation-adjusted (marginal) correlation (CAR) score in linear regression.

To allow application of CAT and CAR scores in high-dimensional omics data we devise an efficient algorithm to derive estimates of CAT and CAR scores. Hence, we provide two highly competitive approaches for variable selection and biomarker identification in high-dimensional omics data. The CAT and CAR score are implemented in the publicly available packages `st` and `care` in the free statistical programming language `R` (R Development Core Team, 2012).

We compare our approaches with other state-of-the-art techniques in extensive simulation studies, where both the CAT and the CAR score are on par with or even outperform their competitors in terms of true positives selected and prediction error. Moreover, we illustrate the application of CAT and CAR score in high-dimensional omics data. We analyze the performance of CAT and CAR score in genomics, transcriptomics, and metabolomics data.

This thesis is based on the following publications:

- V. Zuber and K. Strimmer. 2009. *Gene ranking and biomarker discovery under correlation.* Bioinformatics 25 (20): 2700-2707

- V. Zuber and K. Strimmer. 2009. *Correlation-adjusted t-scores in application to functional magnetic resonance imaging data.* Proceedings of the 6th International Workshop on Computational Systems Biology, WCSB 2009 (June 10-12, 2009, Aarhus, Denmark). pp. 163-166.

- V. Zuber and K. Strimmer. 2011. *High-Dimensional Regression and Variable Selection Using CAR Scores.* Statistical Applications in Genetics and Molecular Biology 10: 34

- V. Zuber, P. Duarte Silva, and K. Strimmer. 2012. *A novel algorithm for simultaneous SNP selection in high-dimensional genome-wide association studies.* BMC Bioinformatics 13:284

1.4 Outline

This thesis is organized as follows. First, we illustrate background information on the scope of this thesis. Section **2.1** describes the biological background of omics data, while Section **2.2** sketches the statistical background of variable selection. In particular, we distinguish between variable selection with respect to prediction or ranking. Then, Section **3** and Section **4** provide detailed information on strategies for variable selection in classification and linear regression, respectively. Both chapters share the same structure; in the beginning existing approaches to variable selection in prediction and ranking are discussed. Then, we present our novel quantities for variable selection, the CAT score is introduced in Section **3.3**, and the CAR score in Section **4.4**. Subsequently, we discuss the derivation, properties, connection to different established quantities, and strategies for estimation. To conclude, both chapters report results on extensive simulation studies.

In Section **5** we illustrate algorithmic details on the estimation of CAT and CAR scores. In particular, we highlight an efficient algorithm that allows to use our approaches even in the case of large dimensional omics data. Finally, Section **6** reports comprehensive studies of real omics data. This includes genomics in Section **6.2**, transcriptomics in Section **6.3**, and metabolomics in Section **6.4**.

Chapter 2

Background on omics data and variable selection

This chapter provides introductory information on the scope of this thesis. First the biological background of omics data is provided to motivate the study of variable selection under correlation in statistics. Then, the basics of variable selection are established to introduce the elementary concepts of prediction and ranking.

2.1 The biological background of omics data

The last decade of biological science witnessed revolutionary biological discoveries and biotechnological inventions, that allow to investigate processes in the cell on new levels of accuracy.

It has been in the year 1953 that James D. Watson and Francis Crick described a structural model of the genetic code encoded in the deoxyribonucleic acid (DNA) by a helix model (Watson and Crick, 1953). This discovery was awarded the Nobel Prize in Physiology or Medicine. In 1958 Francis Crick reported the synthesis of proteins from DNA in two distinct phases: Transcription and translation (Crick (1958) and Crick (1970)). The DNA sequence, given by a specific sequence of nucleotides, is the most elementary part of life and the starting point of the genetic information flow. Genes constitute certain regions of the DNA. In the *transcription phase* DNA is recoded into complementary ribonucleic acid (RNA) that includes messenger RNA (mRNA), ribosomal RNA (rRNA), transfer RNA (tRNA) and micro RNA. The transcription of genes that encode proteins results into mRNA; thus, mRNA is the carrier of protein information that is synthesized in the *translation phase* under influence of rRNA and tRNA into proteins, the product of cells. See e.g. Pollard and Earnshaw (2007) for more detailed information of the protein synthesis. With the advent of blotting technologies in the 1970's it became possible to actually measure the mRNA level, also referred to as expression, of single genes in cells. Twenty years later, microarrays have turned-over the perspective on how detailed processes in the cell can be observed as they allow to measure the expression of several thousands of genes at once. The development of high-throughput microarrays was revolutionary since it captured not only the expression of single genes but the expression of the vast majority of all known genes. Thus, it became practicable to actually map the transcriptome, that is all mRNA constituents of the transcription phase.

Moreover, biotechnological inventions have changed the perspective on further components of the genetic information flow. Sequencing techniques allow to explore the genome, the entire genetic code. So far genomes of several species have been sequenced; most spectacular has been the sequencing of the human genome by the International Human Genome Sequencing Consortium (2001). Today also products of the translation phase are explored in great multiplicity and detail. Mass spectrometry and nuclear magnetic resonance spectroscopy provide insights into the proteome and metabolome, that is all proteins, respectively all metabolites in a cell.

Derived from the greek word "ome" referring to all constituents, "omics" describes the study of all constituents. Thus, *omics* has become a synonym for the datasets produced by modern high-throughput technologies. Characteristic for omics data is for one the large size of observed variables, literally comprising all constituents, and only moderate numbers of observations. Hence, such data sets are described as "small n, large d", where n refers to the number of observations and d to the number of attributes investigated. Furthermore, there is an intrinsic correlation or dependence structure due to unobserved biological processes that influence the observed data, e.g internal cellular regulation, or external, latent factors. More detail on the respective dependence structures is provided in the following sections.

These decisive changes in data structure demanded new strategies for analysis. While standard statistic tools require that the number of observations is larger than the dimension of variables, i.e. $n > d$, the new small n, large d, setting of omics data triggered new innovations in statistics, in particular, regularized regression, the false discovery rate, and a rediscovery of Stein's shrinkage estimates.

Before delving deeper into statistical modeling the next sections discuss in greater detail the most prominent examples for omics data that are used to discover biomarkers. Section **6** illustrates the analysis of high-dimensional omics data from the following levels in the protein-synthesis of a cell and beyond:

- Genomics: DNA sequence level
- Transcriptomics: (m)RNA level
- Metabolomics: Metabolite level

2.1.1 Genomics

Genomics refers to the study of the genome of organisms which is encoded in the DNA. In humans, DNA segments are incorporated in two homologous chromosomes, each inherited from a parent. The DNA information is represented by the four DNA bases or nucleotides: Adenin, thymin, guanine, and cytosine. Genotype is the term for DNA base pairs observed at a specific location. Specific DNA segments code genes or constitute noncoding-regions.

Genetic association studies focus on genetic variation that is captured by so called single nucleotid polymorphisms (SNPs). A SNP is defined as "a single base pair change that is variable across a certain fraction of the general population" (Foulkes, 2009). Association refers to the relationship of the observed genotype to a phenotype of interest. It is studied either in an hypothesis-driven way in candidate gene or fine mapping studies, where only small parts of the genome are considered, or in an exploratory style in genome-wide association studies (GWAS). GWAS have become feasible by the development of next-generation sequencing techniques that parallelize the sequencing process and thus allow to measure vast parts of the genome (Schuster, 2008). The most prominent sequencing technologies are the 454 pyrosequencer, the Ilumina, and SOLiD sequencing. For detailed information of theses high-throughput sequencing techniques see Shendure and Hanlee (2008).

The aim of GWAS is to detect causal variants that are responsible for the occurrence of certain phenotypes. For example, phenotypes of interest can be categorical, like healthy versus affected, or quantitative, like the body mass index, blood serum measurements, or survival time. For the design of genetic association studies it is essential that the phenotype of interest is heritable. Heritability refers to the variance of the phenotype explained by the genotype relative to the overall variance of the phenotype. Whereas the phenotype structure is well-studied and relatively easy to handle the genotype data exhibits an intricate structure to the analyst. In the raw data each observed SNP is coded by the two alleles that are observed at the corresponding diploid location of the genome. The alleles are referred to as major or minor depending on their frequency in the population under study. Thus, the minor allele frequency (MAF) of a SNP can range in an open interval from 0 to 0.5. Often SNPs are divided into common and rare variants with respect to their MAF.

In statistical analysis it is complicated to consider the categorical nature of SNPs. The use of cross-tables becomes prohibitively complex in high dimensions and is so far only recommended for small sets of SNPs. Therefore, the genetic information on SNPs is often recoded into (pseudo-)metric quantities. For recoding SNPs the analyst usually focuses on a specific allele of interest. As an example, this can be the allele associated with risk or the minor allele with the smaller frequency. Following coding schemes are used to recode SNP x_i (Lewis, 2002).

- Additive model:

$$x_i = \begin{cases} 2 & \text{if there are two alleles of interest} \\ 1 & \text{if there is one allele of interest} \\ 0 & \text{if there is no allele of interest} \end{cases}$$

- Recessive model:

$$x_i = \begin{cases} 1 & \text{if there are two alleles of interest} \\ 0 & \text{if there is at most one allele of interest} \end{cases}$$

- Dominant model:

$$x_i = \begin{cases} 1 & \text{if there is at least one allele of interest} \\ 0 & \text{if there is no allele of interest} \end{cases}$$

- Heterozygous model:

$$x_i = \begin{cases} 1 & \text{if there are two identic alleles} \\ 0 & \text{if the two alleles differ} \end{cases}$$

In GWAS, studies that simultaneously analyze all SNPs mostly adopt the additive model, see e.g. Ayers and Cordell (2010) or Hoggart et al. (2008).

Furthermore, there is an intrinsic dependence structure among SNPs, that linkage disequilibrium that describes a non-random association between SNPs. This association is mostly due to mutation and recombination, but various other factors like genetic drift, population growth, migration, population structure, variable recombination rates, variable mutation rates or gene conversions are supposed to influence the association between alleles (Ardlie et al., 2002). Common measures of linkage disequilibrium between two SNPs are the statistics D' and r^2 (Foulkes, 2009). D' is the standardized deviation based on the differences in a 3×3 cross-table under independence and the observed distribution of alleles. Pearson's squared correlation coefficient r^2 is only an ad hoc, but computationally efficient measure for association.

Altogether, SNP-selection in association studies aims at finding genetic variation that is associated with a trait of interest. To accommodate for the dependence structure multivariate models that simultaneously analyze all SNPs are appealing. However, the development of such models is hindered by the dimension of the data.

2.1.2 Transcriptomics

Transcriptomics refers to the study of all RNA molecules. Of special interest is messenger RNA (mRNA) that is the carrier of information from the DNA sequence in the transcription phase and initiates the synthesis of proteins. While most DNA studies focusing on sequence variations consider the amount of DNA constant, the magnitude of mRNA varies considerably and depends on inherited as well as environmental influences. From the amount of mRNA in a cell conclusions can be made regarding to the expression of genes.

There exist several techniques to quantify the expression of genes in a cell. RNA microarrays are the most wide-spread high-throughput technology so far that allows to capture the expression of several thousand genes at once. A microarray chip is equipped with gene-specific probes designed from complementary DNA (cDNA). Being single-stranded cDNA binds to complementary build nucleotides. Since the binding of two complementary DNA strands is due to hydrogen it is called hybridization.

There are two dominant designs of microarrays. The first design is based on synthetic probes in a single-channel system that is structured in 16-20 pairs of perfect match and mismatch probes of 25 bases length. Preprocessed mRNA from the cell studied is given on the array and binds to the corresponding probes. Afterwards the measurements are read out by laser technologies and finally the 16-20 pairs need to be combined to one single intensity value. Another approach is competitive hybridization, a two-channel system. Here, two samples are prepared, one with red-fluorescent dye, the other one with green-fluorescent dye. Then, the two samples are brought together on a microarray chip and hybridization takes place. Relative intensities are finally determined by scanning the differing wavelength of the fluorescence.

Both techniques are prone to systematic errors and inept design of the microarray chips. Especially, a careful preprocessing of the gene expression data is essential. The first step is calibration or normalization to account for differing levels of intensity in the data. Furthermore variance-stabilizing transformations are applied. For an extensive discussion on preprocessing microarray data see e.g. Huber et al. (2002) and Huber et al. (2003), and for application the Bioconductor package `VSN`.

In the context of transcriptomics two more techniques are worth mentioning. Quantitative polymerase chain reaction (qPCR) is a different technique to quantify accurately the abundance of mRNA in a sample. In contrast to RNA microarrays where several thousands of gene expression profile can be captured, qPCR is a cheap (per experiment, not per gene) low-throughput technique measuring the expression of only up to 20 genes in a single reaction.

RNA-Seq is a brand new tool based on deep-sequencing that has the potential to replace the RNA microarray in future (Wang et al., 2009). Still under development, it promises several advantages over the microarray technology. Whereas microarrays are only able to capture pre-specified DNA sequences, RNA-Seq can record unexpected genomic sequences and -at the same time - identifies these sequences. Moreover, in RNA-Seq there is no upper limit to measure mRNA, and thus it provides a more dynamic range of expression profiling. In contrast, microarray chips are only able to capture mRNA abundance up to a certain threshold due to the design of the microarray chips. Finally, using qPCR it has been illustrated that RNA-Seq can provide more accurate measurements than microarrays.

Nevertheless, data from DNA microarrays is the most established way to quantify the transcriptome today. There are two striking characteristics of microarray data. First, it is high-dimensional, capturing the expression of ten thousands of genes (e.g. Affymetrix GeneChip with almost 30 000 gene probes). But the number of observations is only a minor fraction of the variables. Thus, this kind of data is described as "small n, large d". Second, there is a complicated correlation structure among genes that is partly due to common regulation in a joint pathway. There exist several methods to elucidate the correlation structure among genes, e.g. singular value decomposition and eigengenes (e.g. Alter et al., 2000), gene-networks (e.g. Schäfer and Strimmer, 2005), clustering techniques (e.g. Eisen et al., 1998), or independent component analysis (Liebermeister, 2002). A collection of pathways is provided in the Kyoto Encyclopedia of Genes and Genomes (KEGG) *http://www.genome.jp/kegg/*.

2.1.3 Metabolomics

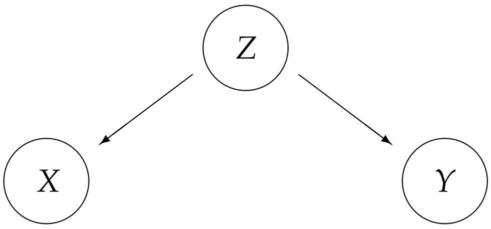
Metabolomics refers to the analysis of all endogenous low-molecular-weight-components in a biological sample (Holmes et al., 2008). It describes a snapshot of end products of chemical processes in the cell. In contrast to the genome the metabolome is not fixed, but depends on various components. The primary metabolome is controlled by the host genome, while the co-metabolome is controlled by different microorganisms, like bacteria, protozoa, or fungi that populate the host. Additionally, the metabolome depends on environmental influences and physical demands on the host. For example, Stella et al. (2006) show that the personal diet leads to differing patterns in the metabolome. The authors report characteristic signatures in the metabolome depending on meat consumption. Due to these latent external factors and interactions among the cell components there is an intrinsic correlation structure among metabolites.

Quantification of the metabolome is mostly performed using nuclear magnetic resonance spectroscopy or mass spectrometry-based techniques, resulting in high-throughput data in form of metabolite levels or digitized spectra. The Human Metabolome Database (Wishart et al., 2007) is the first effort to inventory knowledge on existing metabolites.

The aim of metabolomic studies comprises a wide range. It includes toxicological tests, studies of environmental effects on the cell process, or the detection of molecular patterns of disease. For example, Sreekumar et al. (2009) investigate the metabolic signature inherent in the progression of prostate cancer.

2.2 The statistical background on variable selection

Variable selection is ubiquitous in applied statistics. Although most methods promise good performance in selecting variables, the analyst must consider that methods for variable selection are designed with respect to different aims: Prediction or ranking according to importance. Depending on the aim, this results in possibly different optimal subsets of variables to select; in machine learning these subsets are referred to as "minimal-optimal" and "all-relevant" (Nilsson et al., 2007). Especially in highly correlated data, the "minimal-optimal" and "all-relevant" sets include different variables. For example, in microarray data, there are genes with a highly correlated expression profile due to the regulation in a joint pathway. To predict the outcome it is enough to choose one gene as a representative for this group, whereas in a ranking the whole group should be included on similar positions in the gene-list. Another example is the intricate case of spurious correlation, when one variable is useful to predict the outcome, though it is only indirectly related with the outcome through another variable (Strobl et al., 2008), like in the following illustration, where X and Y are only indirectly linked by a third variable Z.



In case of uncorrelated predictor variables, optimal criteria exist and the "minimal-optimal" set is equal to the top of the "all-relevant" list. It is the correlation among predictors that complicates the analysis. Thus, correlation is often disregarded as e.g in the naive Bayes classifier (Bickel and Levina, 2004) or sure independence screening (Fan and Lv, 2008). *The aim of this thesis is to show how correlation among predictors can be incorporated to improve variable selection.*

This chapter introduces general aspects of prediction and ranking, especially motivation, structure of the problem and criteria to assess the performance. Particular strategies are presented in the following Section **3** and Section **4**.

The following notation is used throughout this thesis:

- Y denotes the one-dimensional variable of interest, also referred to as outcome, output, response, or dependent variable. Depending on the level of measurement different analysis schemes are used: If Y is categorical this is a classification task (Section **3**), while with Y metric this falls into the regression set-up (Section **4**).

- $\boldsymbol{X}$ denotes the d-dimensional explanatory variables, also referred to as input, features, predictors, or independent variables. In machine learning features are "variables constructed for input variables" (Guyon and Elisseeff, 2003). Here, feature and variable are interchangeable.

2.2.1 Prediction

Variable selection for prediction aims at finding a "minimal-optimal" subset of variables. "Minimal" denotes the size of the variable-set; "optimal" refers to accuracy of prediction. The pivotal element of prediction is the *prediction rule* that is derived from training data, where $\boldsymbol{X}$ as well as Y are known. Generally, the prediction rule is a function of predictor variables $\hat{f}(x)$. For example, in linear regression the prediction rule is given by a linear combination of the estimated regression coefficients $\hat{\beta}$ and the explanatory variables $\boldsymbol{x}$

$$\hat{f}(\boldsymbol{x}) = \hat{\beta}\boldsymbol{x}\,.$$

Applying the prediction rule it is possible to assign a prediction $\hat{y}_l = \hat{f}(\boldsymbol{x}_l)$ for the observation $l \in 1, ..., n$ of explanatory variables $\boldsymbol{x}_l$.

Furthermore, a loss function is needed to quantify the divergence between the true value Y and the prediction $\hat{Y} = \hat{f}(\boldsymbol{x})$. The squared error loss function is most commonly used in regression

$$L(Y, \hat{f}(\boldsymbol{x})) = (Y - \hat{f}(\boldsymbol{x}))^2$$

and the zero-one loss function in classification

$$L(G, \hat{G}(\boldsymbol{x})) = I(G \neq \hat{G}(\boldsymbol{x}))$$

where G represents the true class and $\hat{G}(\boldsymbol{x})$ the class estimate inferred by the prediction rule $\hat{f}(\boldsymbol{x})$. I is an indicator-function that penalizes each misclassification with one unit.

It is essential in prediction to derive a prediction rule that does not fit the training data too closely so that it can *generalize* to new data. Here, the selection of variables with good predictive effects plays a vital part. Intuitively, the performance of a predictor is evaluated by the prediction error that is quantified by a prespecified loss function. In estimating the prediction error it is essential to distinguish on which data the prediction error is observed. Optimally the data can be divided into three parts (Hastie et al., 2009):

- Training data to fit the models.
- Validation data to estimate extra parameters of the prediction rule.
- Test data to assess the generalization properties.

The *training error* is defined as the average loss over the specific training sample $\boldsymbol{x}_{\text{tr}}$ of size n that is also used to fit the prediction rule

$$\overline{\text{err}} = \frac{1}{n} \sum_{l \in \boldsymbol{x}_{\text{tr}}} [L(y_l, \hat{f}(\boldsymbol{x}_l))] \,. \tag{2.1}$$

But the training error is not adequate to assess the performance of a prediction rule since it is overoptimistic in how good the prediction rule can generalize to new data. In particular, prediction rules with a low training error tend to overfit, i.e. when the bias is minimized by a too complex model that includes too many redundant variables. The *test error*, also known as generalization error, over an independent test sample, denoted by Y^0 and $\boldsymbol{X}^0$, is defined as the expected loss of an independent test sample with respect to the specific training set $\boldsymbol{x}_{\text{tr}}$ that was used to generate the prediction rule $\hat{f}(\boldsymbol{x})$

$$\text{Err}_{\boldsymbol{x}_{\text{tr}}} = \text{E}_{Y^0,\boldsymbol{X}^0}[L(Y^0, \hat{f}(\boldsymbol{x}^0)) \mid \boldsymbol{x}_{\text{tr}}] \,.$$

Then, the *expected test error* eliminates the randomness of the training set by taking the expectation

$$\text{Err} = \text{E}_{\boldsymbol{x}_{\text{tr}}}\text{E}_{Y^0,\boldsymbol{X}^0}[L(Y^0, \hat{f}(\boldsymbol{x}^0)) \mid \boldsymbol{x}_{\text{tr}}] = \text{E}_{\boldsymbol{x}_{\text{tr}}}[\text{Err}_{\boldsymbol{x}_{\text{tr}}}] \,. \tag{2.2}$$

An idealized illustration of the test and training error with respect to increasing model complexity is given in Figure **2.1**. To assess the quality of a prediction rule the expected test error is the decisive quantity since it reflects best the performance of the prediction rule on future observations. In the following, we continue referring to the test error simply as *prediction error*. In practice it is often not possible to split the observations given into different data sets since there are too few observations. Then the prediction error can be estimated by cross-validation as we will discuss in Section **5.3.3**.

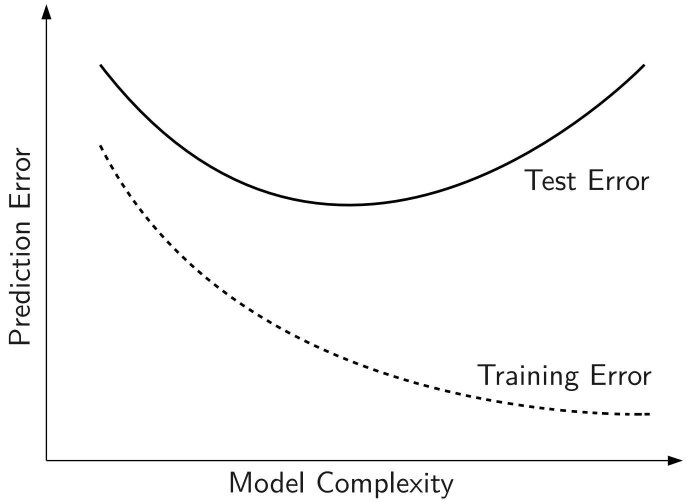


Figure 2.1: Illustration of a model test and training error depending on the model complexity.

According to Hastie et al. (2009), under the weak assumption that

$$Y = f(x) + \epsilon$$

where ϵ, with $\mathrm{E}(\epsilon) = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2$, represents the error term, residuum, or noise, the expected prediction error of a prediction rule $\hat{f}(x)$ at an input point $X = x_0$ can be decomposed into

$$\begin{aligned} \mathrm{Err}(x_0) &= \mathrm{E}[(Y - \hat{f}(x_0))^2 \mid X = x_0] \\ &= \sigma^2 + [\mathrm{E}\hat{f}(x_0) - f(x_0)]^2 + \mathrm{E}[\hat{f}(x_0) - \mathrm{E}\hat{f}(x_0)]^2 \\ &= \sigma^2 + \quad \mathrm{Bias}^2(\hat{f}(x_0)) \quad + \quad \mathrm{Var}(\hat{f}(x_0)) \, . \end{aligned}$$

σ^2 is called the irreducible error since it is the variance around the true mean $f(x_0)$ and thus independent of the prediction rule. To minimize the expected prediction error bias and variance of $\hat{f}(x_0)$ need to be traded-off:

- $\mathrm{Bias}^2(\hat{f}(x_0)) = [\mathrm{E}\hat{f}(x_0) - f(x_0)]^2$:
  The bias can be reduced including more variables.

- $\mathrm{Var}(\hat{f}(x_0)) = \mathrm{E}[\hat{f}(x_0) - \mathrm{E}\hat{f}(x_0)]^2$:
  The variance is increased if more variables are added, since more parameters need to be estimated and each estimation contributes more variance to the overall variance.

As mentioned before the main aim of prediction is to select a set of variables that minimize the expected prediction error. The variables selected are not necessary optimal for interpretation. First, the "minimal-optimal" set does not include all variables related to the outcome. When two variables are highly correlated and have equal effects on the outcome, a good prediction rule needs to include only one of them and discard the other, because the second one does not add any new information. Furthermore, due to spurious correlation, a variable might be useful for prediction, though it is not related to the outcome. Ranking procedures that provide an "all-relevant" subset of variables are more suitable for interpretation.

Accurate prediction rules are especially important in the classification of cancer subtypes based on gene expression signatures. In clinical practice, some subtypes of cancer are difficult to discriminate in standard histology. Still, exact classification is vital to design and to employ the best fitting therapy. For example, Khan et al. (2001) discuss classification of the small, round blue cell tumors of childhood which can be divided in four subtypes, including neuroblastoma, rhabdomyosarcoma, non-Hodgkin lymphoma, and the Ewing family of tumors. The four subtypes are hard to distinguish in light microscopy. Other techniques for diagnosis, like immunohistochemistry for the analysis of proteins or molecular markers based on PCR, fail to provide secure classification due to technical difficulties or too variable measurements (Khan et al., 2001). Using an artificial neural network, the authors are able to construct a classification rule based on gene-expression data of 63 training samples that correctly classifies 25 test samples. In Section **6.3** we illustrate in more detail that there exist clear genetic signatures that allow to discriminate between the subtypes with high precision.

Prediction can also be used to identify subtypes of cancer that exhibit distinct expression signatures. Diffuse large B-cell lymphoma are characterized by heterogeneous clinical outcomes. Less than every second patient responds well to the given therapy and exhibits durable remission. More than half of the patients die of the lymphoma. Alizadeh et al. (2000) illustrate different gene expression patterns depending on the malignancy of the tumor. Additionally, the authors suggest the definition of prognostic groups on the gene expression patterns and important clinical indicators. Depending on the prognostic group the course of therapeutic actions should be adapted. While standard treatment starts with chemotherapy and in case of failed complete remission takes up bone marrow transplantation, a specific prognostic group of patients should receive early bone marrow transplantation.

To sum up, prediction rules constructed on omics data may provide a valuable tool in clinical diagnostics if they can generalize to new data and predict accurately future cases. The central quantity to assess the performance of a prediction rule is the prediction error estimated on an independent test data set or by cross-validation. Variable selection in prediction aims at finding variables with good predictive effects that minimize the prediction error.

2.2.2 Ranking

In contrast, ranking procedures aim at selecting all relevant variables, the "all-relevant" set. These rankings are mainly used for interpretation and provide a "short-list" of interesting or *important* features. Such a short list is especially relevant if the analysis is not hypothesis driven. This is the case when there is only scarce or no a priori information on the true effects and an abundance of possible variables is to be taken into account. Then, a ranking provides an explanatory tool for interpretation.

First, it is essential to define when a variable is important and how this notion of importance can be quantified by a score. For example, in the two group case, where Y is binary, predictor variables are important that discriminate well between the two groups. Thus, importance is often defined as the standardized mean difference which is quantified by the well-known t-score. An extensive discussion on the notion of importance can be found in Section **3.2** and Section **4.3**. Then, all variables are ordered according to decreasing importance. Finally, a cut-off is set to distinguish between variables of interest or *non-null* or *nonzero variables* at the top of the list and uninteresting variables or *null* or *zero variables* at the bottom of the list. It is recommended to fix the cut-off in a way to control the number of false positives included in the variables selected. If the number of variables is small standard hypothesis tests with significance levels adjusted for multiple testing can be used. For high-dimensional settings control of the false-discovery rate is wide-spread (Benjamini and Hochberg, 1995). More details on how to determine a cut-off are given in Section **5.3**. In practice, it is desirable to validate the top-listed variables. Constraints in financial resources or constraints of the techniques used in the follow-up experiments lead to ad-hoc determinations of the cut-offs. For example, qPCR is often used as cheap low-throughput technique for validation of high-throughput gene-expression experiments. Since qPCR captures the expression of only few genes, e.g. 20 genes, the top 20 genes of the ranking are considered for follow-up experiments.

To assess the performance of a ranking quantities are used that are based on:

- True positives (TP):
  Non-null variables correctly identified as non-null variables.

- False positives (FP):
  Null variables incorrectly labeled as non-null variables.

- True negatives (TN):
  Null variables correctly identified as null variables.

- False negatives (FN):
  Non-null variables incorrectly labeled as null variables.

The use of the true positive rate (Sensitivity: $\frac{\text{TP}}{\text{TP+FN}}$) and the true negative rate (Specificity: $\frac{\text{TN}}{\text{TN+FP}}$) is widespread. A combination of those is represented in the receiver operating characteristic (ROC). In case of skewed classes, i.e. when there is only a small fraction of either non-null or null variables, the use of ROC-curves is discouraged in favor of precision (True discovery rate) and recall (True positive rate, sensitivity, power):

- Precision $= \frac{\text{TP}}{\text{TP+FP}}$
- Recall $= \frac{\text{TP}}{\text{TP+FN}}$

Ranking is especially popular in the analysis of transcriptomics, where gene-lists often present the genes highly associated with the outcome of interest. For example, Pomeroy et al. (2002) analyze the differences in gene expression depending on the outcome of central nervous system embryonal tumor. Since there is little knowledge on the molecular basis of the tumor, the analysis is conducted in an explanatory fashion to elucidate markers related to the outcome. Thus, the authors compile two lists of genes, one with genetic markers for survival and one with genetic markers for treatment failure. In a similar fashion Singh et al. (2002) investigate gene expression profiles of patients suffering from prostate cancer. The authors compile a list of 29 genes correlated to the Gleason score that quantifies the degree of tumor cell differentiation.

To conclude, ranking procedures are valuable tools in explanatory studies that facilitate interpretation. Variable selection in ranking aims at finding all variables related to the variable of interest, the true positive effects, while controlling the number of false positives contained in the list.

Chapter 3

Variable selection in classification

Classification is the general term for methods that model a categorical outcome Y that represents the membership to one of K classes or groups. It is important that the number of possible classes K is limited, classes are disjoint, and the membership to a class is unambiguous. For illustration, this thesis focuses on the two-group case where $K = 2$. i.e. Y is a factor variable with only two possible realizations:

$$Y = \begin{cases} 1 & \text{if the observation belongs to group 1} \\ 2 & \text{if the observation belongs to group 2} \end{cases}$$

A generalization to $K > 2$ is straightforward (Ahdesmäki and Strimmer, 2010), thus this thesis gives only a short sketch of multi-class classification. First, strategies for prediction are discussed. These include discriminant analysis and logistic regression. Then, quantities for variable ranking are presented. Here, variants of the t-score are widespread. After presenting established methods we introduce a novel approach to variable selection under correlation, correlation-adjusted t (CAT) scores. We show that CAT scores are the natural approach to ranking variables under correlation since they are motivated from linear discriminant analysis which is the classical approach to prediction in case of correlated predictors. Furthermore, we discuss the most important properties of the CAT score, point at relations to other quantities, and present strategies to estimate them in high-dimensional data. Finally, the performance of prediction and ranking strategies is examined in simulations. For application of CAT scores on real omics data we refer to Section **6.3** and Section **6.4**.

3.1 Prediction

For prediction this thesis focuses on linear models for classification, (linear) discriminant analysis and logistic regression. Linear refers to the decision boundary that is modeled to discriminate between the groups. There exist various techniques that provide more flexible solutions, like nearest neighbor classifiers, support vector machines, or neural networks. For an extensive overview see Hastie et al. (2009). Although linear models might seem to be of dusted quality compared to modern computer-intensive techniques, they are highly popular in practice. The popularity of linear models is due to the well-established statistical framework, interpretable results, and efficient performance (Hand, 2006).

3.1.1 Discriminant analysis

Discriminant analysis is based on the assumption that the d predictor or explaining variables $\boldsymbol{X}$ follow the gaussian distribution with differing parameters depending on the class-membership $Y = k$. Each class $k \in 1, ..., K$ is represented by a multivariate normal distribution

$$\begin{aligned} f(\boldsymbol{x}|k) = (2\pi)^{-d/2}|\boldsymbol{\Sigma}_k|^{-1/2} \times \\ \exp\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k)\} \end{aligned}$$

with

- the vector of expectations $\boldsymbol{\mu}_k$ of length d, and
- the covariance matrix $\boldsymbol{\Sigma}_k$ of size $d \times d$ that has the variances σ_i^2, with $i \in 1, ..., d$, on its diagonal.

The multivariate normal distribution is solely determined by these two parameters. There are three different types of discriminant analysis depending on the restrictions on the covariance matrix $\boldsymbol{\Sigma}_k$:

- Diagonal discriminant analysis (DDA): $\boldsymbol{\Sigma}_k = \text{diag}(\sigma_1^2, ..., \sigma_d^2)$
  assumes no correlation among the predictor variables $\boldsymbol{X}$, but differing variances σ^2. In machine learning DDA is often referred to as 'independence rule' or 'naive Bayes' (Bickel and Levina, 2004).
- Linear discriminant analysis (LDA): $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$
  relaxes the assumption of DDA to the presence of correlation among the predictor variables $\boldsymbol{X}$. But there is *no difference in covariance structure between the groups*. That is all K groups have the same covariance $\boldsymbol{\Sigma}$.
- Quadratic discriminant analysis:
  further eases the restriction of equal covariances and allows different covariances $\boldsymbol{\Sigma}_k$ in all K groups. Nonetheless, quadratic discriminant analysis is seldom used in practice since K covariance matrices of size $d \times d$ need to be estimated. The remainder of this thesis neglects

quadratic discriminant analysis since it is impracticable in the analysis of high-dimensional data.

For classification the probability of belonging to class k conditional on the predictors $\boldsymbol{X}$ is decisive. In discriminant analysis this conditional probability for class k given the data is derived using Bayes' theorem. The joint distribution of $\boldsymbol{X}$ is given by a mixture of all K conditional distributions with a priori mixing weights π_k

$$f(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k f(\boldsymbol{x}|k)\,.$$

Using the conditional distribution $f(\boldsymbol{x}|k)$ and the joint mixture density $f(\boldsymbol{x})$ Bayes' theorem gives the posteriori probability of group k given the data,

$$\Pr(k|\boldsymbol{x}) = \frac{\pi_k f(\boldsymbol{x}|k)}{f(\boldsymbol{x})}.$$

In turn the conditional probability defines the discriminant score $d_k(\boldsymbol{x}) = \log\{\Pr(k|\boldsymbol{x})\}$. The *discriminant score* is the pivotal element of discriminant analysis as it is the basis for classification. In LDA it is possible to drop terms constant across groups to simplify the discriminant score $d_k^{\text{LDA}}(\boldsymbol{x})$ to

$$d_k^{\text{LDA}}(\boldsymbol{x}) = \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \log(\pi_k). \tag{3.1}$$

Due to the common covariance $d_k^{\text{LDA}}(\boldsymbol{x})$ is linear in $\boldsymbol{x}$, hence the name of the procedure. For DDA the discriminant score reduces to

$$d_k^{\text{DDA}}(\boldsymbol{x}) = \boldsymbol{\mu}_k^T \boldsymbol{V}^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_k^T \boldsymbol{V}^{-1}\boldsymbol{\mu}_k + \log(\pi_k) \tag{3.2}$$

where $\boldsymbol{V} = \text{diag}(\sigma_1^2, ..., \sigma_d^2)$ is a $d \times d$ matrix with the variances of $\boldsymbol{X}$ on its diagonal.

In the two group case ($K = 2$) the prediction rule is given by the difference between the two discriminant scores

$$\Delta(\boldsymbol{x}) = d_1(\boldsymbol{x}) - d_2(\boldsymbol{x}) = \log\{\Pr(k=1|\boldsymbol{x})\} - \log\{\Pr(k=2|\boldsymbol{x})\}\,. \tag{3.3}$$

The assignment to class 1 or 2 depends on the sign of the prediction rule; if $\Delta(\boldsymbol{x}) \geq 0$ the observation is classified to group 1, for $\Delta(\boldsymbol{x}) < 0$ it is classified to group 2. In classification of more than two classes an observation is assigned to the class k with the highest a posteriori probability $\Pr(k|\boldsymbol{x})$ or discriminant score $d_k(\boldsymbol{x})$.

A different representation for multi-class classification is given by the pooled centroid formulation. Here, the pooled mean over all K classes is given by

$$\boldsymbol{\mu}_{pool} = \sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_k \,.$$

Following Equation **3.1** the pooled discriminant score is given as

$$d_{pool}^{\mathrm{LDA}}(\boldsymbol{x}) = \boldsymbol{\mu}_{pool}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x} - \frac{1}{2} \boldsymbol{\mu}_{pool}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_{pool}.$$

Then, the centered prediction score for class k is

$$\Delta_k(\boldsymbol{x}) = d_k(\boldsymbol{x}) - d_{pool}(\boldsymbol{x})\,.$$

For prediction, this centered prediction score is equivalent to $d_k^{\mathrm{LDA}}(\boldsymbol{x})$ and analogously an observation is assigned to the class k with the highest centered prediction score.

In practice, the mixing probabilities π_k, the expectations μ_k, and the covariance $\boldsymbol{\Sigma}$ need to be estimated and plugged into the discriminant score.

- The mixing weights π_k represent the a priori probability of belonging to group k. It is estimated by the relative frequency of group k in the sample. In the two group case $1 = \pi_1 + \pi_2$ holds.

- If there are more observations than variables ($n > d$) empirical estimates can be used, like the arithmetic mean for the expectation and the covariance estimate

$$\begin{aligned} \hat{\mu}(\boldsymbol{x}) = \quad \bar{\boldsymbol{x}} = \quad & \frac{1}{n} \sum_{l=1}^{n} \boldsymbol{x}_l \,, \\ \widehat{\mathrm{Cov}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \quad \tfrac{1}{n-1} \quad & \sum_{l=1}^{n} (\boldsymbol{x}_{li} - \bar{x}_i)(\boldsymbol{x}_{lj} - \bar{x}_j)\,. \end{aligned}$$

- In settings with small n, large d, the estimation of the covariance matrix is intricate since the dimension of the $d \times d$ matrix becomes prohibitively large. There are two strategies, the first is to apply regularized estimates on the covariance matrix, like shrinkage estimates (Schäfer and Strimmer, 2005) or $L1$ penalties (Witten and Tibshirani, 2011). Otherwise it is possible to discard the correlation among the predictor variables which is equal to a diagonal covariance matrix, i.e. restrict to DDA (Bickel and Levina, 2004).

- The widely-used prediction analysis for microarrays (PAM) algorithm (Tibshirani et al., 2002) is based on the independence assumption from DDA. Moreover, it applies shrunken centroids to perform an inherent selection of variables by a soft threshold. The standard centroid definition $\boldsymbol{c}_k$ with regularization for the standard deviation for class k with expectation $\boldsymbol{\mu}_k$ is given as the standardized difference

$$c_k = \frac{\boldsymbol{\mu}_k - \boldsymbol{\mu}_{pool}}{\sigma + s_0}$$

where s_0 is a positive constant to stabilize the estimation of the variance since too small values in the denominator can lead to large values in $\boldsymbol{c}_k$. This is equivalent to

$$\boldsymbol{\mu}_k = \boldsymbol{\mu}_{pool} + (\sigma + s_0)\boldsymbol{c}_k\,.$$

PAM replaces the centroids $\boldsymbol{c}_k$ by shrunken centroids $\boldsymbol{c}_k^\lambda$ that are defined by a soft threshold with a constant λ

$$\boldsymbol{c}_k^\lambda = sign(\boldsymbol{c}_k)(\mid \boldsymbol{c}_k \mid -\lambda)_+$$

where + denotes the positive part and zero otherwise. The penalization parameter λ is determined by cross-validation. Using these shrunken centroids PAM applies the following expectations for class k

$$\boldsymbol{\mu}_k^\lambda = \boldsymbol{\mu}_{pool} + (\sigma + s_0)\boldsymbol{c}_k^\lambda\,.$$

Thus, any variable X_i with $(\mid c_{ik} \mid -\lambda) < 0$ for all k does not contribute to the prediction, since $\boldsymbol{\mu}_k^\lambda = \boldsymbol{\mu}_{pool}$. Finally, the discriminant score of PAM is defined as

$$d_k^{\mathrm{PAM}}(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{\mu}_k^\lambda)^T \boldsymbol{V}_{PAM}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k^\lambda) - \log(\pi_k)$$

where $\boldsymbol{V}_{PAM}$ is a diagonal matrix with $(\sigma + s_0)^2$ on its diagonal. PAM is implemented in the `R`-package `pamr`.

3.1.2 Logistic regression

Generalized linear models allow to model responses that follow distributions different from the gaussian distribution. If Y follows a distribution that can be expressed as an exponential family, it is possible to define a link function g that allows to model a linear relationship between the response Y and the linear predictor $\eta = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{X}$ (Fahrmeir and Tutz, 2001). In particular, for classification in the two group case the logit link is the most popular one. Using the logit link the probability of membership to class 1 given the data yields the linear predictor η

$$g\left(\Pr(k = 1 \mid \boldsymbol{x})\right) = \log\{\frac{\Pr(k = 1 \mid \boldsymbol{x})}{1 - \Pr(k = 1 \mid \boldsymbol{x})}\} = \eta. \tag{3.4}$$

In turn $\Pr(k = 1 \mid \boldsymbol{x})$ can be expressed by the inverse transformation

$$\Pr(k = 1 \mid \boldsymbol{x}) = g^{-1}(\eta) = \frac{\exp(\eta)}{1 + \exp(\eta)}.$$

Using the notation of the regression coefficients the probabilities of belonging to class 1, respectively class 2, depending on $\boldsymbol{X}$ are given as

$$\Pr(k = 1|\boldsymbol{x}) = \frac{\exp(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x})}{1 + \exp(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x})},$$

$$\Pr(k = 2|\boldsymbol{x}) = \frac{1}{1 + \exp(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x})}.$$

Like in the linear discriminant analysis, the d predictor variables $\boldsymbol{X}$ follow a mixture of two normal distributions with different expectations μ_1 and μ_2 but common covariance matrix $\boldsymbol{\Sigma}$. The a priori mixing weights are given as π_1 and $\pi_2 = 1 - \pi_1$, respectively. Then, the intercept β_0 and regression coefficients $\boldsymbol{\beta}$ can be expressed as (Efron, 1975)

$$\beta_0 = \log \frac{\pi_1}{\pi_2} - \frac{1}{2} \left(\boldsymbol{\mu}_1^T \Sigma^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^T \Sigma^{-1} \boldsymbol{\mu}_2 \right),$$

$$\boldsymbol{\beta}^T = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \Sigma^{-1}. \tag{3.5}$$

There is a close connection between LDA in the two group case and logistic regression: The discriminant function in LDA in Equation **3.3** corresponds to the logit link model in Equation **3.4**

$$\begin{aligned}\Delta(\boldsymbol{x}) &\overset{(\,\mathbf{3.3})}{=} d_1(\boldsymbol{x}) - d_2(\boldsymbol{x}) = \log\{\Pr(k = 1|\boldsymbol{x})\} - \log\{\Pr(k = 2|\boldsymbol{x})\} = \\ &= \log\{\frac{\Pr(k = 1 \mid \boldsymbol{x})}{1 - \Pr(k = 1 \mid \boldsymbol{x})}\} \overset{(\,\mathbf{3.4})}{=} g(\Pr(k = 1 \mid \boldsymbol{x})).\end{aligned}$$

Still, the estimation of coefficients differs in the approach and its efficiency (Efron, 1975). In contrast to LDA, where only estimates of covariance and expectation are plugged in to derive estimates for the discriminant function, in logistic regression the estimates b_0 and $\boldsymbol{b}$ of the regression coefficients are obtained by maximizing the conditional likelihood with respect to b_0 and $\boldsymbol{b}$

$$f_{b_0,\boldsymbol{b}}(y_1, ..., y_n \mid \boldsymbol{x}_1, ..., \boldsymbol{x}_n) = \prod_{l=1}^{n} \frac{\exp(b_0 + \boldsymbol{b}^T \boldsymbol{x}_l) y_l}{1 + \exp(b_0 + \boldsymbol{b}^T \boldsymbol{x}_l)}.$$

Maximum likelihood estimation is only possible if there are more observations than variables. Otherwise regularized approaches are needed. A detailed overview of regularized regression is presented in Section **4.2**.

3.2 Variable ranking

The ranking of variables in classification is based on measures of variable importance. A variable is considered as important if *it helps to discriminate between the K classes*. As in the previous chapter the focus is on the two-group case. First, the t-score is introduced as a standard criterion for variable importance to compare two groups. Furthermore, the relation of t-score and diagonal discriminant analysis is discussed. Since the t-score suffers from unstable estimates in the analysis of high-dimensional microarray data regularization strategies are presented in the following section, including the well-established strategies SAM (Tusher et al., 2001), shrinkage t (Opgen-Rhein and Strimmer, 2007c), and moderated t (Smyth, 2004).

3.2.1 Quantities for variable importance

In classification a (explanatory) variable X_i in $\boldsymbol{X}$ is considered as important if it discriminates between the groups defined by the variable of interest Y. The dominant quantity of importance in the two-group case is the t-score. The empirical t-score for variable $\boldsymbol{x}_i$ is defined as the difference between the mean $\bar{x}_1(i)$ in group 1 and the mean $\bar{x}_2(i)$ in group 2, standardized with the sample size n_1 in group 1, respectively the sample size n_2 in group 2, and the sample variance $\boldsymbol{s}^2(i)$

$$t(i) = \left((\frac{1}{n_1} + \frac{1}{n_2}) s^2(i) \right)^{-1/2} (\bar{x}_1(i) - \bar{x}_2(i)) \,. \qquad (3.6)$$

Depending on the assumptions on the variance different estimates are applied. For equal variances in the groups the following estimate is used

$$s^2(i) = \frac{(n_1 - 1)s_1^2(i) + (n_2 - 1)s_2^2(i)}{n_1 + n_2 - 2}$$

with

$$s_1^2(i) = \frac{1}{n_1 - 1} \sum_{l=1}^{n_1} (x_l(i) - \bar{x}_1(i))^2 \quad \text{and} \quad s_2(i) = \frac{1}{n_2 - 1} \sum_{l=1}^{n_2} (x_l(i) - \bar{x}_2(i))^2 \,.$$

The t-score is the test-statistic from Student's t-test (Student, 1908) that was derived by W.S. Gosset, better known by his pseudonym Student, in 1908 to assess the yield of different varieties of barley (Box, 1987). Student's t-test checks if there is a difference in means of one metric variable X in two groups defined by a binary factor Y. The (two-sided) null hypothesis is given as $\mu_1 = \mu_2$, i.e. there is no difference in means between the two groups. Under the null hypothesis the t-score follows the t-distribution with $n_1 + n_2 - 2$ degrees of freedom.

A special case of the t-score is the assumption that all variances are equal, that is $\sigma_1^2 = ... = \sigma_d^2$. Then, the t-score is proportional to the fold change defined as the mean differences between the two groups. A different definition for the fold change is the ratio of the two mean differences; in this thesis, the definition of mean differences is used. Since microarray data is often log-transformed in the preprocessing step it holds that $\log(\frac{\mu_1}{\mu_2}) = \log \boldsymbol{\mu}_1 - \log \boldsymbol{\mu}_2$.

Interestingly, there is a close connection between the t-score and diagonal discriminant analysis. Let $\boldsymbol{\tau}$ be a multivariate representation of the d-dimensional t-score vector, where the estimated variance
$\boldsymbol{S} = \text{diag}(s^2(1), ..., s^2(d))$ is replaced by the population variance
$\boldsymbol{V} = \text{diag}(\sigma_1^2, ..., \sigma_d^2)$, $(\frac{1}{n_1} + \frac{1}{n_2})$ by a constant c_n and the means by the expectations $\boldsymbol{\mu}_1$, respectively $\boldsymbol{\mu}_2$

$$\boldsymbol{\tau} = \{c_n \boldsymbol{V}\}^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\,. \tag{3.7}$$

The constant c_n is a scale factor inherent in the t-score that depends only on the sample size in the two groups and is constant for all d variables. Thus, it is negligible in ranking the d variables. In the following, this thesis discriminates between $\boldsymbol{\tau}$, the t-score on the *population level*, and $\boldsymbol{t}$, the definition of the *empirical* t-score.

Recalling the prediction rule in DDA as difference in discriminant scores (Equation **3.2**) we decompose the prediction rule into a weight vector $\boldsymbol{\omega}^{\text{DDA}}$, independent of $\boldsymbol{X}$, and a vector-valued distance function $\delta(\boldsymbol{x})$

$$\begin{aligned}
\Delta^{\text{DDA}}(\boldsymbol{x}) &= d_1^{\text{DDA}}(\boldsymbol{x}) - d_2^{\text{DDA}}(\boldsymbol{x}) \\
&= \left(\boldsymbol{\mu}_1^T \boldsymbol{V}^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_1^T \boldsymbol{V}^{-1}\boldsymbol{\mu}_1 + \log(\pi_1)\right) - \\
&\quad \left(\boldsymbol{\mu}_2^T \boldsymbol{V}^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_2^T \boldsymbol{V}^{-1}\boldsymbol{\mu}_2 + \log(\pi_2)\right) \\
&= (\boldsymbol{\mu}_1^T - \boldsymbol{\mu}_2^T)\boldsymbol{V}^{-1}\boldsymbol{x} - \frac{1}{2}\left((\boldsymbol{\mu}_1^T - \boldsymbol{\mu}_2^T)\boldsymbol{V}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right) + \log(\frac{\pi_1}{\pi_2}) \\
&= \underbrace{(\boldsymbol{\mu}_1^T - \boldsymbol{\mu}_2^T)}_{1\times d}\underbrace{\boldsymbol{V}^{-1}}_{d\times d}\underbrace{\left(\boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right)}_{d\times 1} + \log(\frac{\pi_1}{\pi_2}) \\
&= \underbrace{(\boldsymbol{\mu}_1^T - \boldsymbol{\mu}_2^T)\boldsymbol{V}^{-1/2}}_{\boldsymbol{\omega}^{\text{DDA}^T}} \cdot \underbrace{\boldsymbol{V}^{-1/2}\left(\boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right)}_{\delta(\boldsymbol{x})^{\text{DDA}}} + \log(\frac{\pi_1}{\pi_2}) \\
&= \underbrace{\boldsymbol{\omega}^{\text{DDA}^T}}_{1\times d}\underbrace{\delta(\boldsymbol{x})^{\text{DDA}}}_{d\times 1} + \log(\frac{\pi_1}{\pi_2})\,.
\end{aligned} \tag{3.8}$$

This decomposition illustrates the main constituents of the DDA prediction rule:

- A constant governed by the a priori mixing properties, $\log(\frac{\pi_1}{\pi_2})$,
- the vector-valued distance function $\delta(\boldsymbol{x})^{\text{DDA}}$ that quantifies the distance of the variables $\boldsymbol{X}$ to the overall mean $\frac{\mu_1+\mu_2}{2}$ corrected for the variances
$$\delta(\boldsymbol{x})^{\text{DDA}} = \boldsymbol{V}^{-1/2}(\boldsymbol{x} - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2}),$$
- a weight vector, independent of $\boldsymbol{X}$
$$\boldsymbol{\omega}^{\text{DDA}} = \boldsymbol{V}^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

Thus, the weight vector $\boldsymbol{\omega}^{\text{DDA}}$ controls the influence of a standardized variable $\delta(\boldsymbol{x})$ on the prediction rule. Here, the weight vector $\boldsymbol{\omega}^{\text{DDA}}$ is proportional up to the constant c_n to the t-score vector $\boldsymbol{\tau}$ as defined on the population level in Equation **3.7**. In the following, $\boldsymbol{\omega}^{\text{DDA}}$ is interpreted as an unscaled version of the population t-score $\boldsymbol{\tau}$. Moreover, this decomposition shows that the t-score is the natural quantity to measure the importance of a variable on the prediction rule in case of uncorrelated explanatory variables. Hence, the t-score is the *optimal criterion* to select variables for prediction if there is *no correlation* among predictors (Fan and Fan, 2008).

3.2.2 Regularized estimates

The previous section has introduced the standard definition of the t-score, that may suffer from a severe defect in practice due to the presence of small variances near zero. Then, the standard deviation tends even more to zero and the t-score from Equation **3.6** as ratio of mean difference to standard deviation tends to infinity even for small to moderate differences in mean. Thus, the standard t-score is prone to high values due to small variances that are usually considered as unimportant for biological interpretation. For interpretation the mean difference is much more important. To overcome this instability in estimation regularized versions of the t-score have been introduced. The first publication on a regularized t-score is by Kerr et al. (2001). Since then several versions of t-scores have been introduced. Regularized t-scores modify Equation **3.6** by adding a small constant s_0 to the denominator

$$t_{\text{reg}}(i) = \frac{(\bar{x}_1(i) - \bar{x}_2(i))}{\sqrt{c_n}(s(i) + s_0)}. \tag{3.9}$$

All regularized t-scores share the same generalized definition from Equation **3.9** but they differ in how to derive the denominator, especially the constant s_0. The following section gives a short introduction to three established regularized t-scores: SAM (Tusher et al., 2001), moderated t (Smyth, 2004) and shrinkage t (Opgen-Rhein and Strimmer, 2007c).

- SAM is the abbreviation for significance analysis of microarrays and is implemented in the `R`-package `samr`. Here, the constant s_0 in Equation **3.9** is set to minimize the coefficient of variation.

- The moderated t is derived from a hierarchical model, where prior information on the variance $s(i)$ is modeled by an inverse χ^2-distribution with d_0 degrees of freedom

  $$\frac{1}{s^2(i)} \sim \frac{1}{d_0 s_0^2}\chi^2(d_0)\,.$$

  This prior is incorporated into the posterior variance $s^2(i)_{mod}$ that is represented by a mixture of the sample variance $s(i)$ and the prior variance s_0

  $$s^2(i)_{\text{mod}} = \frac{d_g}{d_g + d_0}s^2(i) + \frac{d_0}{d_g + d_0}s_0^2\,.$$

  The amount of mixture is governed by the a priori degrees of freedom d_0 and the sample degrees of freedom d_g. Finally, this posteriori variance is plugged into the denominator of Equation **3.9**. Thus, the moderated t is proportional to an ordinary t-score by the constant c_n but instead of using the sample variance a mixture of sample and prior variance is used to stabilize the t-score. The prior parameters d_0 and s_0 are estimated from the data in an empirical Bayes approach. Moderated t is implemented in the `R`-package `limma`.

- In contrast to the moderated t the shrinkage t makes no assumptions on the distribution of the variance. The shrinkage t is derived from the James-Stein rule. A generalized illustration of a James-Stein ensemble estimate $\boldsymbol{\theta}^{shrink}$ for an unknown parameter vector $\boldsymbol{\theta}$ of length d is based only on three components

  $$\boldsymbol{\theta}^{\text{shrink}} = (1-\lambda)\hat{\boldsymbol{\theta}} + \lambda\boldsymbol{\theta}^{\text{target}} \qquad (3.10)$$

  with

  - $\hat{\boldsymbol{\theta}}$ as an unregularized estimate for $\boldsymbol{\theta}$,
  - $\boldsymbol{\theta}^{\text{target}}$ as the target estimate, and
  - λ as the parameter to govern the amount of shrinkage.

  The target estimate $\boldsymbol{\theta}^{\text{target}}$ is specified using a priori information. E.g. for estimating the variance vector of a set of variables the target can be set as the median variance of the variables, or for estimating the correlation matrix the diagonal identity matrix can be used as target representing the a priori information of no correlation. The shrinkage parameter λ is set to minimize the expected loss that is quantified by a loss function. Using the quadratic loss function is equivalent to the minimization of the mean squared error (MSE) of the shrinkage estimate

$$\begin{aligned}\mathrm{MSE}(\boldsymbol{\theta}^{\text{shrink}}) &= \mathrm{MSE}(\hat{\boldsymbol{\theta}}) + \lambda^2 \overbrace{\sum_{i=1}^{d} \mathrm{E}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i^{\text{target}})^2}^{b} \\ &\quad - 2\lambda \underbrace{\sum_{i=1}^{d} \{\mathrm{Var}(\hat{\boldsymbol{\theta}}_i) - \mathrm{Cov}(\hat{\boldsymbol{\theta}}_i, \boldsymbol{\theta}_i^{\text{target}}) + \mathrm{Bias}(\hat{\boldsymbol{\theta}}_i)\mathrm{E}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i^{\text{target}})\}}_{a} \\ &= \mathrm{MSE}(\hat{\boldsymbol{\theta}}) + \lambda^2 b - 2\lambda a\,.\end{aligned}$$

Thus, the MSE curve depending on λ equals a parabola with parameters b and a. Hence, the optimal shrinkage rule to minimze the MSE is given as $\lambda = \frac{a}{b}$. If λ equals zero, no shrinkage is applied and an unregularized estimate is used; if λ equals one, the target is used and no estimation influences the shrinkage estimate. Small values of λ may arise if b is large due to an inept specification of the target. In contrast, high values of λ may arise if a is large due to high variance or bias of the unregularized estimate.

Shrinkage t applies a James-Stein estimate for the d-dimensional variance vector $\boldsymbol{s}^2$ of the t-score. As target the median of all d empirical variances $s^2(1), ..., s^2(d)$ is employed. Then, the actual shrinkage estimate $s(i)_{\text{shrink}}$ for variable $\boldsymbol{x}_i$ is a mixture of empirical sample variance $s(i)$ and the median variance

$$s(i)_{\text{shrink}} = (1 - \lambda)s(i) + \lambda s_{\text{median}}\,. \tag{3.11}$$

A slight modification of the optimal shrinkage estimate is used, a so called positive part estimate that truncates λ at 1

$$\lambda = \min\left(1, \frac{\sum_{i=1}^{d} \widehat{\mathrm{Var}}(s(i))}{\sum_{i=1}^{d}(s(i) - s_{\text{median}})}\right)\,.$$

The larger the variance of the empirical estimate, the more shrinkage is applied to the James-Stein estimate. Otherwise, if the target is specified inadequately the James-Stein estimate equals more the empirical sample variance. Shrinkage t is implemented in the R-package `st`. Interestingly, the shrinkage t is an intermediate between t-score and fold change.

$$\text{If } \lambda = \begin{cases} 1, & \text{the median variance is used as estimate for all variables,} \\ & \text{then } t_{shrink} \text{ is proportional to the fold change.} \\ 0, & \text{no shrinkage is applied;} \\ & t_{\text{shrink}} \text{ equals the unrestricted } t\text{-score.} \end{cases}$$

All of the presented adaptions of the t-score are essentially *univariate*. Information on the variance of a variable is shared or borrowed across variables, like in the shrinkage t, where the median variance is used as target, or in the moderated t, where the prior variance is estimated in an empirical Bayes approach on all variables. But no correlation or covariance across variables is considered.

3.3 Decorrelation: The correlation-adjusted t-score

In the following chapter we demonstrate how variable selection in case of correlated variables can be improved by explicitly modeling the correlation structure among variables. We introduce a novel approach to variable selection under correlation, correlation-adjusted t-scores abbreviated as CAT scores. First, the definition of CAT scores is presented. Then, we derive the CAT score from the predictive point of view as analogon to the t-score in case of correlated variables from LDA. Furthermore, we show how CAT scores decompose Hotelling's T^2 and interpret them as intermediate between the t-score and logistic-regression coefficients $\boldsymbol{\beta}$. The CAT score is a population quantity, hence any suitable estimate can be used. Here, we concentrate on empirical estimates for scenarios with more samples than variables and the shrinkage approach for large d, small n settings. Finally, results from extensive simulation studies are presented.

3.3.1 Definition of the CAT score

We define correlation-adjusted t-scores (CAT score) as the inverse square root of the correlation among $\boldsymbol{X}$ times the d-dimensional t-score vector

$$\begin{aligned} \boldsymbol{\tau}^{adj} &\equiv \boldsymbol{P}^{-1/2} \times \{c_n \boldsymbol{V}\}^{-1/2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ &= \boldsymbol{P}^{-1/2} \boldsymbol{\tau} \end{aligned} \tag{3.12}$$

where

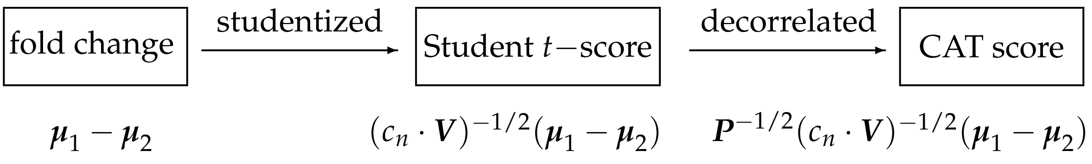
- μ_1, μ_2 are the expectations of $\boldsymbol{X}$ in group 1, respectively group 2,
- $\boldsymbol{P}$ is the $d \times d$ correlation matrix of $\boldsymbol{X}$,
- $\boldsymbol{V}$ is the $d \times d$ variance matrix with the variances of $\boldsymbol{X}$ on its diagonal,
- $c_n = \frac{1}{n_1} + \frac{1}{n_2}$ is a scale factor inherent in the t-score,
- $\boldsymbol{\tau}$ is the d-dimensional t-score vector as described in Equation **3.7**.

The scale factor c_n ensures that the empirical version of the CAT score matches the scale of the empirical t-score. In turn, the empirical CAT score is denoted as

$$\boldsymbol{t}^{adj} = \boldsymbol{R}^{-1/2}\boldsymbol{t} \tag{3.13}$$

where $\boldsymbol{R}$ is the estimated correlation matrix. Decorrelation is performed by the Mahalanobis transform. More details on this special transformation is provided in Section **5.1**. The CAT score is a natural and intuitive extension of both the fold change and t-score, as illustrated in Figure **3.1**. While the t-score is the standardized mean difference $\mu_1 - \mu_2$, the CAT score is the standardized as well as *decorrelated* mean difference.

Figure 3.1: The CAT score as generalization of the fold change and t-score. It is the mean difference, studentized as well as decorrelated.


fold change
studentized
Student t−score
decorrelated
CAT score
$\mu_1 - \mu_2$
$(c_n \cdot \boldsymbol{V})^{-1/2}(\mu_1 - \mu_2)$
$\boldsymbol{P}^{-1/2}(c_n \cdot \boldsymbol{V})^{-1/2}(\mu_1 - \mu_2)$

3.3.2 Derivation from linear discriminant analysis

The CAT score is derived from LDA, the natural approach to classification under correlation. If there is no correlation among variables, DDA is considered as the optimal classification technique. It is well known that in the DDA setting the t-score is the natural and optimal ranking criterion for discovering variables that best differentiate the two classes (Fan and Fan, 2008). Previously, Equation **3.8** has demonstrated how the quantity $\boldsymbol{\omega}^{\text{DDA}}$ that is proportional to the t-score, governs the prediction rule in DDA, as $\boldsymbol{\omega}^{\text{DDA}}$ represents the weights of the standardized variables on the prediction rule. We follow an analogous decomposition of the prediction rule (Equation **3.3**) of LDA with discriminant scores as described in Equation **3.1**. Note, the common covariance matrix $\boldsymbol{\Sigma}$ is decomposed into variance and correlation matrix $\boldsymbol{\Sigma} = \boldsymbol{V}^{1/2}\boldsymbol{P}\boldsymbol{V}^{1/2}$.

For $K = 2$, the prediction rule in LDA is represented by

$$\begin{aligned}
\Delta^{\text{LDA}}(\boldsymbol{x}) &= d_1^{\text{LDA}}(\boldsymbol{x}) - d_2^{\text{LDA}}(\boldsymbol{x}) \\
&= \left(\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \log(\pi_1)\right) - \\
&\quad \left(\boldsymbol{\mu}_2^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_2^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 + \log(\pi_2)\right) \\
&= (\boldsymbol{\mu}_1^T - \boldsymbol{\mu}_2^T)\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}\left((\boldsymbol{\mu}_1^T - \boldsymbol{\mu}_2^T)\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right) + \log\left(\frac{\pi_1}{\pi_2}\right) \\
&= (\boldsymbol{\mu}_1^T - \boldsymbol{\mu}_2^T)\underbrace{\boldsymbol{\Sigma}^{-1}}_{\boldsymbol{V}^{-1/2}\boldsymbol{P}^{-1}\boldsymbol{V}^{-1/2}}\left(\boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right) + \log\left(\frac{\pi_1}{\pi_2}\right) \\
&= \underbrace{(\boldsymbol{\mu}_1^T - \boldsymbol{\mu}_2^T)\boldsymbol{V}^{-1/2}\boldsymbol{P}^{-1/2}}_{\boldsymbol{\omega}^T} \cdot \\
&\quad \underbrace{\boldsymbol{P}^{-1/2}\boldsymbol{V}^{-1/2}\left(\boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right)}_{\delta(\boldsymbol{x})} + \log\left(\frac{\pi_1}{\pi_2}\right) \\
&= \underbrace{\boldsymbol{\omega}^T}_{1\times d}\underbrace{\delta(\boldsymbol{x})}_{d\times 1} + \log(\frac{\pi_1}{\pi_2}) \ . \qquad (3.14)
\end{aligned}$$

Thus, the prediction rule $\Delta^{\text{LDA}}(\boldsymbol{x})$ can be decomposed into the weight vector

$$\underbrace{\boldsymbol{\omega}}_{d\times 1} = \boldsymbol{P}^{-1/2}\boldsymbol{V}^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \qquad (3.15)$$

and the vector-valued distance function

$$\underbrace{\delta(\boldsymbol{x})}_{d\times 1} = \boldsymbol{P}^{-1/2}\boldsymbol{V}^{-1/2}(\boldsymbol{x} - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2}). \qquad (3.16)$$

The benefit of expressing two-class LDA in this fashion is that it clarifies the underlying mechanism. In particular, the difference score $\Delta^{\text{LDA}}(\boldsymbol{x})$ is governed solely by three factors:

- The log-ratio of the mixing proportions π_1 and π_2,
- $\delta(\boldsymbol{x})$, the standardized and decorrelated distance of the data $\boldsymbol{X}$ to the average centroid, and
- the variable-specific feature weights $\boldsymbol{\omega}$, proportional to the CAT score.

A special benefit of decomposing the prediction rule accordingly is that the weight vector $\boldsymbol{\omega}$ is not a function of $\boldsymbol{x}$ and that it carries no units of measurements. Its components ω_i directly control how much each particular variable X_i contributes to the overall score Δ^{LDA}. Thus, $\boldsymbol{\omega}$ *defines the importance of a variable on prediction in two-class linear discriminant analysis*. In the absence of correlation the weights $\boldsymbol{\omega}$ directly reduce to $\boldsymbol{\omega}^{\text{DDA}} = \boldsymbol{V}^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ (see Equation **3.8**) which is (apart from the constant c_n) the usual vector of two-sample t-scores.

3.3.3 Properties of the CAT score

- **Intermediate between t-score and β-coefficient**

First, the CAT score is an intermediate between the unscaled t-score $\boldsymbol{V}^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \boldsymbol{\omega}^{\text{DDA}}$ and the standardized regression coefficients $\boldsymbol{\beta}_{std} = \boldsymbol{V}^{1/2}\boldsymbol{\beta}$ from logistic regression (Equation **3.5**). As illustrated in Figure **3.2** the weight vector $\boldsymbol{\omega}$ equals the $\boldsymbol{P}^{-1/2}$ times the unscaled t-score vector $\boldsymbol{\omega}^{\text{DDA}}$. And $\boldsymbol{\beta}_{std}$ equals $\boldsymbol{P}^{-1/2}$ times the unscaled CAT score vector. Thus, the CAT score is an intermediate between unscaled t-score and standardized $\boldsymbol{\beta}$-coefficient with respect to how correlation is integrated.

Figure 3.2: The CAT score as an intermediate between t-score and standardized regression coefficients.

$$\begin{tikzcd}
\boxed{\boldsymbol{\omega}^{\text{DDA}}} \arrow[r, "\boldsymbol{P}^{-1/2}"] & \boxed{\boldsymbol{\omega}} \arrow[r, "\boldsymbol{P}^{-1/2}"] & \boxed{\boldsymbol{\beta}_{std}} \\
\boldsymbol{V}^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) & \boldsymbol{P}^{-1/2}\boldsymbol{V}^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) & \boldsymbol{P}^{-1}\boldsymbol{V}^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)
\end{tikzcd}$$

- **Decomposition of Hotelling's T^2**

Second, the sum of squared CAT scores adds up to Hotelling's T^2 (Hotelling, 1931), a standard criterion to test if there is any difference in means in a set of variables. It can be considered as the multivariate generalization of the t-score. For d variables it is defined as

$$\begin{aligned}
T^2 &= \{(\frac{1}{n_1} + \frac{1}{n_2})\}^{-1}(\bar{x}_1 - \bar{x}_2)^T \boldsymbol{S}^{-1/2}\boldsymbol{R}^{-1}\boldsymbol{S}^{-1/2}(\bar{x}_1 - \bar{x}_2) \\
&= \boldsymbol{t}^T\boldsymbol{R}^{-1}\boldsymbol{t} \\
&= (\boldsymbol{t}^{\text{adj}})^T\boldsymbol{t}^{\text{adj}}
\end{aligned} \tag{3.17}$$

where $\boldsymbol{R}$ is the estimated correlation matrix, $\boldsymbol{S}$ is the estimated variance matrix, $\boldsymbol{t}$ the vector containing the Student t-statistics, and $\boldsymbol{t}^{\text{adj}}$ the empirical CAT score vector. Thus, the inner product of CAT scores or the sum of squared CAT scores adds up to Hotelling's T^2.

- **Grouped CAT score**

  For evaluating the total effect of a set of features on group separation we exploit the close connection of CAT scores to Hotelling's T^2 statistic, a standard criterion in the multivariate analysis of sets of variables. We define the grouped CAT score for variable $\boldsymbol{x}_i$ belonging to a given set as the signed square root of the sum over the squared CAT scores of all variables in the given set,

  $$\tau_i^{\text{adj,grouped}} = \text{sign}(\tau_i^{\text{adj}}) \sqrt{\sum_{g \in \text{set}} (\tau_g^{\text{adj}})^2} \,.$$

  Note that any normalization with regard to the size of the set is implicit in the factor $\boldsymbol{R}^{-1}$. There are two main cases when it is important to consider sets of variables rather than individual variables:

  - First, if predefined sets of variables exist. For example, prior knowledge is used in a gene set enrichment analysis where prespecified pathways or functional units rather than individual genes are being investigated (cf. Ackermann and Strimmer (2009)).
  - Second, if variables are highly correlated and thus provide the same information on group separation. To accommodate for this collinearity we suggest constructing a suitable correlation neighborhood around each variable, e.g., by the rule $|r| \geq 0.85$. Typically, the resulting sets are rather small and most sets comprise only the variable itself – see Tibshirani and Wasserman (2006) and Läuter et al. (2009) for similar procedures.

  We note that using the grouped CAT score provides to a simple procedure for high-dimensional feature selection where whole sets of variables are simultaneously included or excluded, in contrast to the classical view of feature selection where only one of those features is retained.

- **Grouping property**

  Additionally, the CAT score exhibits an intrinsic grouping property for highly correlated variables. Using the definition $\boldsymbol{\omega} = \boldsymbol{P}^{1/2}\boldsymbol{\beta}_{\text{std}}$ for two predictors X_1 and X_2 and correlation $\text{Cor}(X_1, X_2) = \rho$ a simple algebraic calculation shows that the difference between the two squared

CAT scores equals

$$\omega_1^2 - \omega_2^2 = \left((\beta_{\text{std}})_1^2 - (\beta_{\text{std}})_2^2 \right) \sqrt{1-\rho^2} \, .$$

Thus, highly correlated variables tend to have identical CAT scores. Hence, both variables are located on adjacent positions in a ranking.

3.3.4 Estimation

The original CAT score $\boldsymbol{\tau}^{adj}$ is a population quantity and not tied to any estimation scheme. Any suitable estimates can be used to substitute means, variances, and correlations in Equation **3.12**. In settings with $n \gg d$ using empirical estimates for means, variances, and correlations provides a simple recipe.

For small-sample yet high-dimensional settings we suggest to employ James-Stein-type shrinkage estimators of correlation (Schäfer and Strimmer, 2005) and of variances (Opgen-Rhein and Strimmer, 2007a). Plugging these two James-Stein-type estimators into Equation **3.12** yields a shrinkage version of the CAT score

$$\boldsymbol{t}^{\text{adj}}_{\text{shrink}} = (\boldsymbol{R}^{\text{shrink}})^{-1/2} \, \boldsymbol{t}^{\text{shrink}} \tag{3.18}$$

where $\boldsymbol{t}^{\text{shrink}}$ is the shrinkage t-statistic as presented in Section **3.2.2**. The shrinkage correlation matrix $\boldsymbol{R}^{\text{shrink}}$ is given as a mixture between the empirical correlation matrix $\boldsymbol{R}^{\text{empirical}}$ and the target $\boldsymbol{\theta}$ that is governed by the shrinkage parameter λ

$$\boldsymbol{R}^{\text{shrink}} = (1-\lambda)\boldsymbol{R}^{\text{empirical}} + \lambda\boldsymbol{\theta} \, . \tag{3.19}$$

The empirical correlation matrix $\boldsymbol{R}^{\text{empirical}}$ for $i, j \in 1, ..., d$ is defined as

$$r_{ij} = \frac{1}{n-1} \sum_{l=1}^{n} (x_{il} - \bar{x}_i)(x_{jl} - \bar{x}_j) / (\hat{\sigma}_i \hat{\sigma}_j) \, .$$

As target matrix $\boldsymbol{\theta}$ we employ the $d \times d$ identity matrix $\boldsymbol{I}_d$. If there is structural information, e.g. an autoregressive structure among the data, a structured target can be used. The shrinkage parameter λ is set to minimize the mean squared error. Using the identity matrix as target the optimal parameter $\lambda^\star$ simplifies to

$$\lambda^\star = \sum_{i \neq j} \widehat{\text{Var}}(r_{ij}) / \sum_{i \neq j} r_{ij}^2 \, .$$

For more details see Schäfer and Strimmer (2005). This shrinkage estimate of the CAT score represents an intermediate between shrinkage t-score and full empirical CAT score. If $\lambda = 1$ then $\boldsymbol{R}^{\text{shrink}} = I$ and no correlation among $\boldsymbol{X}$ is incorporated and if $\lambda = 0$ no shrinkage is applied and the correlation among $\boldsymbol{X}$ is fully integrated. Any $0 < \lambda < 1$ provides an intermediate. Note that for highly variable, and thus insecure, estimates of the correlations the shrinkage parameter λ tends to one. This results in an estimated CAT score equal to the shrinkage t, where no correlation among $\boldsymbol{X}$ is incorporated.

A major obstacle in the application of Equation **3.18** is the problem of efficiently computing the matrix power $(\boldsymbol{R}^{\text{shrink}})^{-1/2}$. To this end we introduce an efficient algorithm to derive $(\boldsymbol{R}^{\text{shrink}})^{-1/2}$ in Section **5.2**.

3.4 Simulation studies

In order to study the performance of the CAT score for ranking variables, we conducted an extensive simulation study. Specifically, we investigated six different correlation scenarios, three synthetic models and three correlation structures derived from real data. To mimic dependency structures in omics data we estimated empirical correlation matrices from three different gene expression data sets. First, we sketch the data generation and list competing versions of regularized t-scores. Finally, we report the true discovery rates and precision-recall curves.

3.4.1 Correlation scenarios

For the correlation structure, we considered a variety of scenarios. Specifically, we employed six different correlation patterns (cf. Figure **3.3**):

A: First, as a negative control we assumed a diagonal correlation matrix $\boldsymbol{P} = \boldsymbol{I}$ of size 1000×1000.

B: Next, we employed an autoregressive block-diagonal correlation matrix (Guo et al., 2007). We used 10 blocks of size 100×100 variables. Within each block, the correlation between two variables $i, j, = 1, \ldots, 100$ equals $\rho(i, j) = \rho^{\text{abs}(i-j)}$. We set $\rho = 0.99$ with alternating sign in each block. This correlation matrix is sparse with most entries being very small, nevertheless it also contains some highly correlated variables.

C: Third, we employed a correlation block structure where the first 100 variables have pairwise correlation of 0.7 and the remaining 900 variables have pairwise correlation of 0.3. Between the two groups there is no correlation. The block with the larger correlation corresponds to the variables with differences in means.

D: In addition to the three artificial correlation structures, we also employed shrinkage estimators of correlations matrices from three trancriptomics data sets, using a sample of 1000 genes. Structure D is obtained from gene expression data of colon cancer (Alon et al., 1999).

E: As D, but using gene expression data on breast cancer (Hedenfalk et al., 2001).

F: As D, but using gene expression data from a spike-in experiment (Choe et al., 2005).

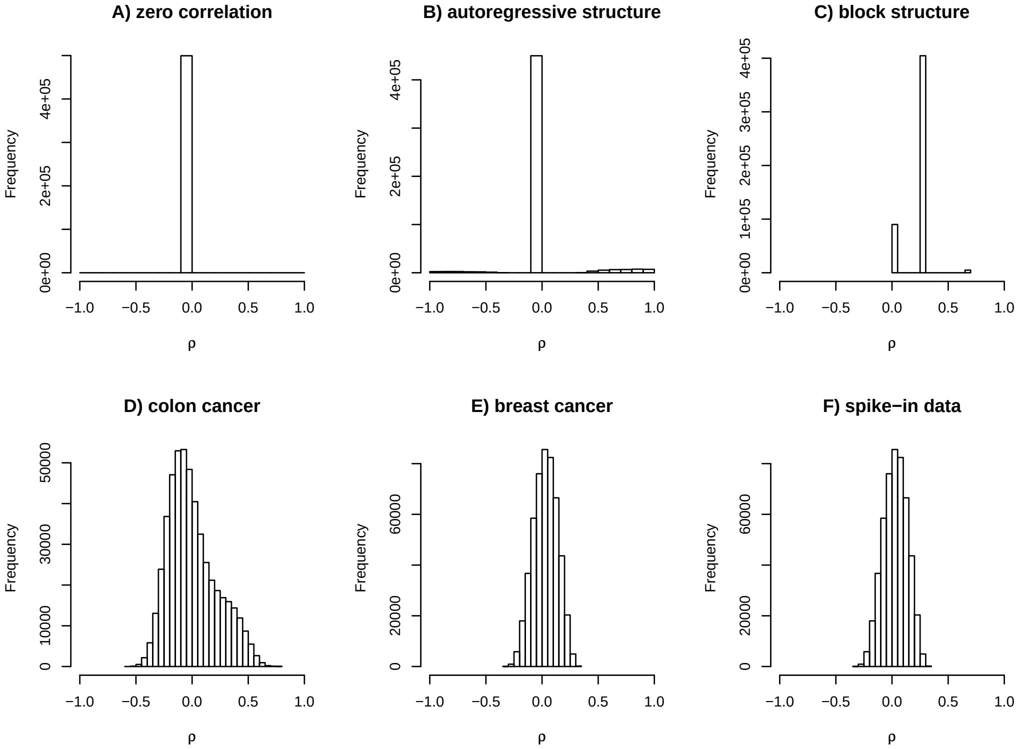


Figure 3.3: The six correlation scenarios investigated in our study. All correlation matrices have size 1000 × 1000 and thus contain 499500 correlation values. *Top row:* Histograms of the correlations of three synthetic correlation patterns (A–C). *Bottom row:* Histograms of the three shrinkage correlation structures (D–F).

3.4.2 Data generation

In our data generation procedure we followed closely the setup in Smyth (2004) and Opgen-Rhein and Strimmer (2007a), with the additional specification of a correlation structure among variables. In detail, the simulations were conducted as follows:

- The number of variables was fixed at $d = 1000$, where the first 100 variables were designated to have differences in means unequal to zero. These variables characterized by differences in means between the two groups are referred to as nonzero variables.

- The variances of the variables were drawn from a scale-inverse-chi-square distribution Scale-inv-$\chi^2(d_0, s_0^2)$. We used $s_0^2 = 4$ and $d_0 = 4$ which corresponds to the "balanced" variance case in Smyth (2004). Thus, the variances vary moderately from variable to variable.

- The difference of means for the nonzero variables (1–100) were drawn from a normal distribution with mean zero and the variable-specific variance. For the zero variables (101–1000) the difference was set to zero.

- The data were generated by drawing from group-specific multivariate normal distributions with the given variances and means. The correlation matrix assumed one of the above structures A–F.

- We also varied the sample sizes n_1 and n_2 in each group, from very small $n_1 = n_2 = 3$ to fairly large $n_1 = n_2 = 50$. Here, we report results for $n_1 = n_2 = 8$.

3.4.3 Competing test statistics

In our comparison we included the following statistics: Fold change, empirical t statistic, SAM (Tusher et al., 2001), moderated t (Smyth, 2004), and shrinkage t (Opgen-Rhein and Strimmer, 2007a). As in Opgen-Rhein and Strimmer (2007a) the latter three regularized t-scores gave nearly identical estimates and always outperformed Student t, so we report here only the results for shrinkage t. As baseline reference we also included random ordering in the analysis.

For the CAT score we investigated two variants: The shrinkage CAT score (Equation **3.18**) and an oracle version which uses the true underlying correlation matrix rather than estimating the correlation structure. For the two structures with high correlations (B and C) we employed the grouped CAT score using a correlation neighborhood threshold of 0.85. Note that the suggested threshold of 0.85 is rather conservative. It defines a priori which pairs of variables are assumed to be collinear.

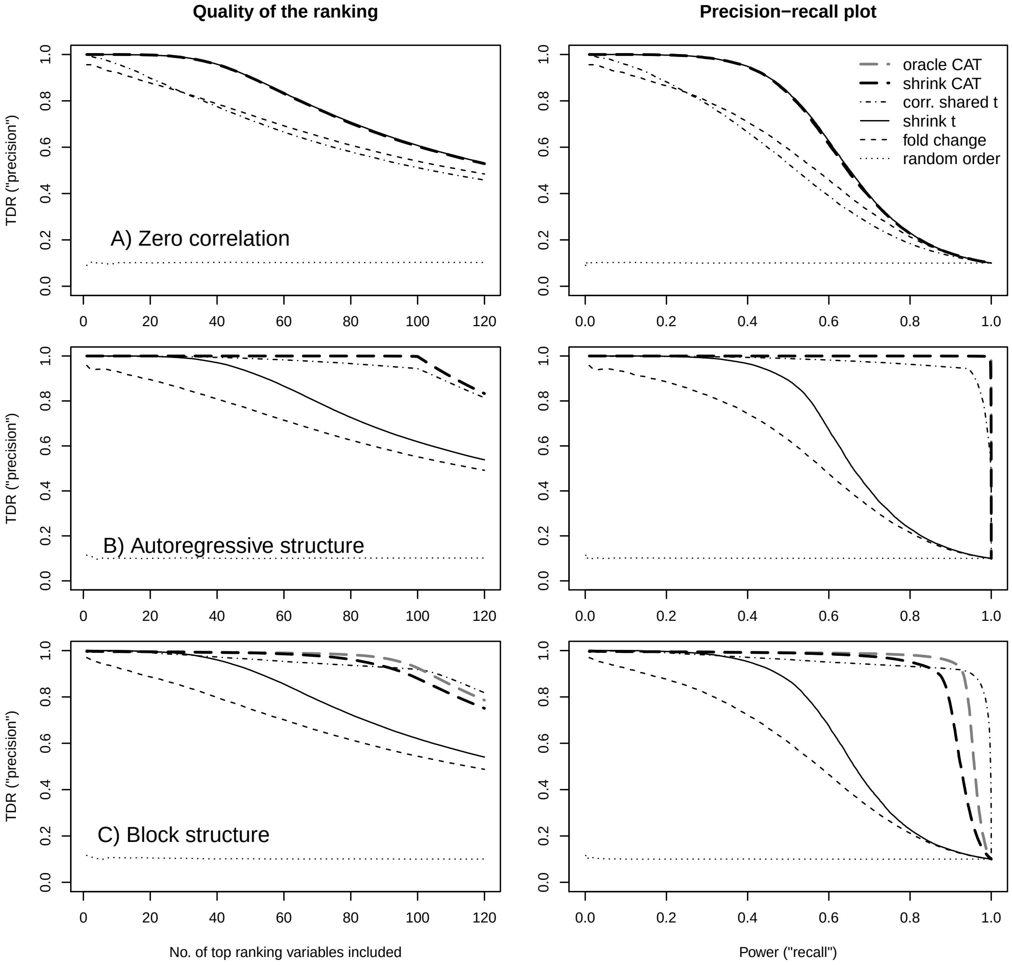


Figure 3.4: True discovery rates (TDR *left column*) and precision-recall curves (*right column*) for the three synthetic correlation structures A–C. Note that for B and C the grouped CAT score was employed, using a correlation neighborhood $|r| \geq 0.85$.

In addition, we included in our study a recently proposed procedure that, like the CAT score, also aims at incorporating information about the correlation among $\boldsymbol{X}$: the correlation-shared t-score introduced by Tibshirani and Wasserman (2006). Correlation-shared t averages over variable-specific Student t-scores in a data-dependent correlation neighborhood. Note that the CAT score and the correlation-shared t-score are based on linear combinations of t-scores, albeit with different weights.

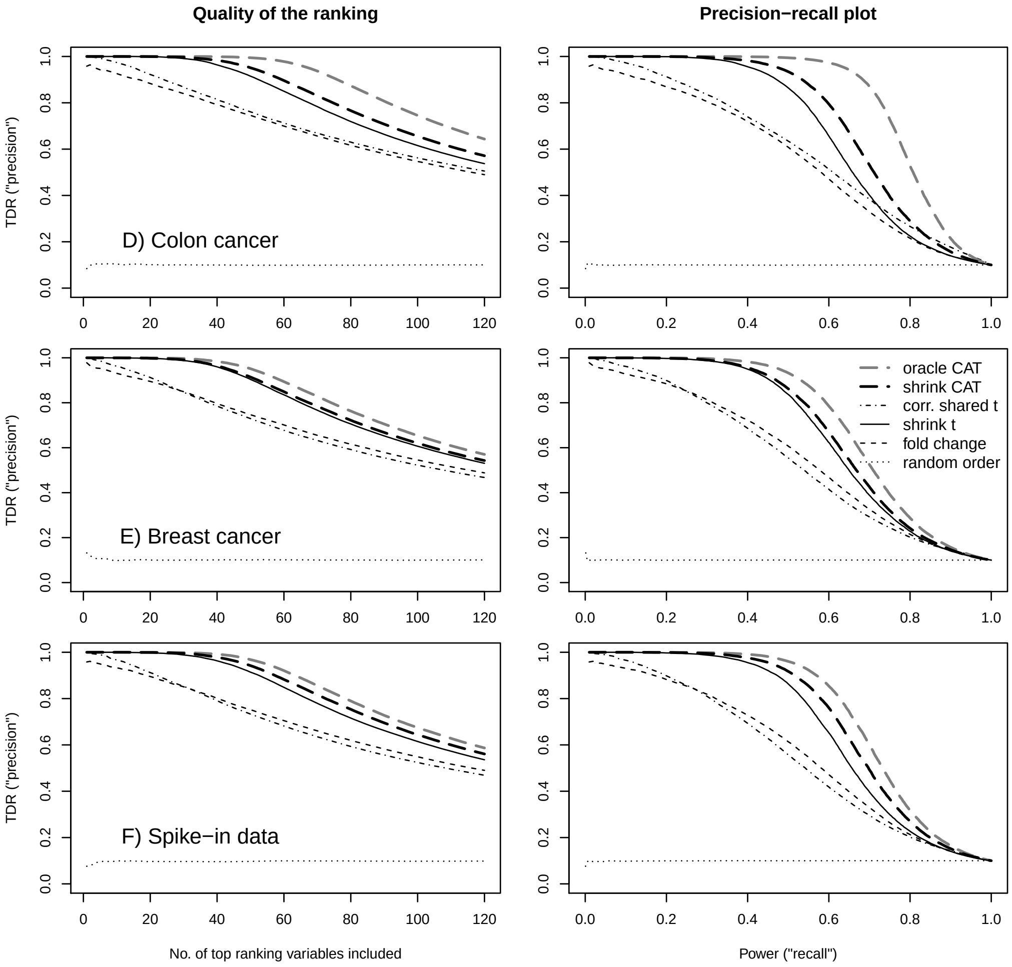


Figure 3.5: True discovery rates (TDR *left column*) and precision-recall curves (*right column*) for the three shrinkage correlation scenarios D–F.

3.4.4 Comparison of variable rankings

For each correlation scenario A–F we generated 500 data sets and computed corresponding rankings using the various *t*-scores and CAT scores discussed above. We then counted false positives, true positives, false negatives, and true negatives for all possible cut-offs in the variable list (1-1000). From this data we estimated the true discovery rates and the power.

A graphical summary of the results are presented in Figure **3.4** and Figure **3.5**. The first column shows the true discovery rates as a function of the number of included top-ranking variables, whereas the second column gives the plots of true discovery rate versus power. The latter graphs, known in the machine learning community as precision-recall plots, highlight methods that simultaneously have large power and large true discovery rates.

The first row in Figure **3.4** shows the control case when there is no correlation present. As expected, the CAT score performs identical to the shrinkage *t* approach. A similar performance is given by the correlation-shared *t* and the fold change statistic, slightly worse than shrinkage *t*- and CAT score.

For the autoregressive and the block structure (scenarios B and C in Figure **3.4**) substantial gains are achieved over the shrinkage *t*-score, both by the CAT score and the correlation-shared *t*-score In particular, in case B these two methods show near-perfect recovery of the gene ranking. The shrinkage *t* approach and fold change remain the second and third best feature ranking approach.

For the shrinkage-estimated correlation structures the picture changes slightly (cf. Figure **3.5**). All these scenarios have in common that there is common background correlation but no very strong individual pairwise correlations exist (cf. Figure **3.3**, bottom row). In this setting the shrinkage CAT score also improves over the shrinkage *t*-score. The oracle CAT score shows that further benefits are possible if the correlation structure was known, or if a better estimator was used. For the scenarios with realistic correlation structure the correlation-shared *t*-score performs similar to the fold change.

In summary, in all the six quite different correlation scenarios the (grouped) CAT score offers in part substantial performance improvements over standard regularized *t*-scores which were represented here by the shrinkage *t*-score. The correlation-shared *t*-score also performs exceptionally well if there are a few highly correlated variables, but otherwise falls back to the efficiency of using the fold-change approach.

3.5 Summary

The correlation-adjusted *t*-score is the result of our attempt to incorporate knowledge on the correlation structure among predictors into the selection of variables. While it is well known that in the absence of correlation the *t*-score provides optimal rankings (Fan and Fan, 2008), the situation is less clear in case of correlated predictor variables. Either this information is disregarded in ranking according to (regularized) versions of the *t*-score or otherwise a full logistic regression model or LDA is fitted that aims at minimizing the prediction error.

Here, we propose a different approach. We utilize the decomposition of the prediction rule in DDA into a weight vector $\boldsymbol{\omega}^{\text{DDA}}$, that is proportional to the *t*-score, and the standardized data $\delta(\boldsymbol{x})^{\text{DDA}}$. Since DDA is the type of discriminant analysis employed in case of uncorrelated variables, we focus on LDA, the generalization of DDA for correlation among the variables. Along the lines, we demonstrate that the prediction rule in LDA can be decomposed into a weight vector $\boldsymbol{\omega}$, that is proportional to the CAT score, and the standardized and decorrelated data $\boldsymbol{\delta}(\boldsymbol{x})^{\text{LDA}}$. Hence, we argue that the CAT score is the generalization of the t-score under correlation and provides a natural weight for variable selection in LDA analysis. Moreover, we elucidate that the CAT score decomposes Hotelling's T^2, a standard multivariate criterion to test for mean differences.

In simulations studies we show that the CAT score improves the ranking of variables in correlated scenarios, since it detects more true positives than competing approaches. Besides, in Section **6.3** and Section **6.4** we analyze the quality of prediction rules on real omics data and conclude that the CAT score also is competitive with modern state-of-the-art techniques for prediction in real omics data.

Chapter 4

Variable selection in regression

Regression summarizes strategies to model a quantitative trait of interest. As in the previous chapter, this thesis focuses on linear approaches to regression. First, the basic notation and definitions of linear regression are established. Facing the small n, large d problem several penalized adaptions of the standard linear regression have been proposed that perform an intrinsic step of variable selection. Here, the most important ones are discussed. Penalized regression techniques aim at optimizing the prediction error. Thus, these methods are listed under the Section **4.2** Prediction. Then, this chapter includes a review on ranking variables according to their importance.

After the introduction to existing techniques we show how variable selection under correlation can be improved by explicitly accounting for the correlation structure. We introduce a novel quantity for variable importance in the linear model, correlation-adjusted correlation, abbreviated as CAR scores. CAR scores are the analogon of CAT scores in case of a quantitative trait of interest. Along the lines we derive the CAR score from the best linear predictor. Furthermore, we show how it facilitates the representation of the decomposition of variances, the centerpiece of the linear model. Especially, we demonstrate that the sum of squared CAR scores adds up to the proportion of variance explained.

In simulations we demonstrate that CAR scores are applicable even in the analysis of high-dimensional data and moreover improve the performance in selecting the true effects and minimizing the prediction error. Finally, we conclude by a comparison of CAT and CAR score.

4.1 Linear regression revisited

4.1.1 The linear regression model and best linear predictor

Linear regression describes the (linear) relationship between one variable of interest and d explaining variables $\boldsymbol{X}$ by the following linear combination

$$Y = \beta_0 + \underbrace{\boldsymbol{\beta}^T}_{1\times d}\underbrace{\boldsymbol{X}}_{d\times 1} + \epsilon \tag{4.1}$$

where

- Y is the dependent variable, outcome, or response with
  - expectation μ_Y, and
  - variance σ^2_Y,
- $\boldsymbol{X}$ are the d predictors or explaining variables ($d \times 1$) with
  - expectations $\boldsymbol{\mu}$, and
  - covariance matrix $\boldsymbol{\Sigma}$ that can be decomposed into a variance matrix $\boldsymbol{V}$ and correlation matrix $\boldsymbol{P}$ according to

    $$\boldsymbol{\Sigma} = \boldsymbol{V}^{1/2}\boldsymbol{P}\boldsymbol{V}^{1/2}.$$
- $\boldsymbol{\beta}$ are the d regression coefficients of size $(d \times 1)$,
- β_0 is the intercept or offset, and
- ϵ is the irreducible error with $E(\epsilon) = 0$.

See e.g. Whittaker (1990) [Chapter 5] for more details. For interpretation the β-coefficients are most important; β_i, with $i \in 1, ..., d$, gives the influence of X_i on Y conditional on all the other $d-1$ variables. In the following this thesis refers to

$$\begin{aligned} X_i \text{ as zero variable, if } & \beta_i = 0 \\ X_i \text{ as nonzero variable, if } & \beta_i \neq 0. \end{aligned}$$

Intercept and β-coefficients are selected to minimize the squared divergence between the established model $\hat{Y} = \beta_0 + \boldsymbol{\beta}^T\boldsymbol{X}$ and the response, the so called prediction error $\mathrm{E}\left((Y - \hat{Y})^2\right)$. As we are going to show in Section **4.1.2** the prediction error is a pivotal quantity in the linear model. The prediction error is minimized by regression coefficients equal to

$$\boldsymbol{\beta} = \Sigma^{-1}\Sigma_{XY} \tag{4.2}$$

where Σ_{XY} is the d-dimensional vector of covariances between $\boldsymbol{X}$ and Y, and an intercept equal to

$$\beta_0 = \mu_Y - \boldsymbol{\beta}^T \boldsymbol{\mu}\,. \tag{4.3}$$

Hence, the best linear predictor equals

$$Y^\star = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{X}\,. \tag{4.4}$$

The coefficients β_0 and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^T$ are *constants*, and not random variables like $\boldsymbol{X}$, Y and $Y^\star$.

Often, it is convenient to center and standardize the response and the predictor variables. With $Y_{\text{std}} = (Y - \mu_Y)/\sigma_Y$ and $\boldsymbol{X}_{\text{std}} = \boldsymbol{V}^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu})$ the predictor equation (Equation **4.4**) can be written as

$$Y^\star_{\text{std}} = (Y^\star - \mu_Y)/\sigma_Y = \boldsymbol{\beta}^T_{\text{std}} \boldsymbol{X}_{\text{std}} \tag{4.5}$$

where $\boldsymbol{\beta}_{\text{std}}$ are the standardized regression coefficients

$$\boldsymbol{\beta}_{\text{std}} = \boldsymbol{V}^{1/2} \boldsymbol{\beta} \sigma_Y^{-1} = \boldsymbol{P}^{-1} \boldsymbol{P}_{XY} \tag{4.6}$$

where $\boldsymbol{P}_{XY}$ is the d-dimensional vector of correlations between $\boldsymbol{X}$ and Y and $\boldsymbol{P}$ is the $d \times d$ matrix of correlations among $\boldsymbol{X}$. The standardized intercept vanishes because of the centering.

In practice, the variable of interest $\boldsymbol{y}$ is represented by a n-dimensional vector of observations and the explaining variables $\boldsymbol{x}$ by a $d \times n$ matrix. Empirical estimates $\boldsymbol{b}$ of the regression coefficients are derived by minimizing the residual sum of squares (RSS)

$$\text{RSS}(\boldsymbol{b}) = \left(\boldsymbol{y} - (b_0 + \boldsymbol{b}^T \boldsymbol{x})\right)^T \left(\boldsymbol{y} - (b_0 + \boldsymbol{b}^T \boldsymbol{x})\right)\,. \tag{4.7}$$

Differentiating with respect to $\boldsymbol{b}$ and setting the derivative to zero leads to the ordinary least squares solution

$$\boldsymbol{b} = \boldsymbol{x}(\boldsymbol{x}^T \boldsymbol{x})^{-1} \boldsymbol{x}^T \boldsymbol{y}\,.$$

According to the Gauss Markov Theorem the least squares solution has the smallest variance of all unbiased estimates (Fahrmeir et al., 2003). Nonetheless, it is possible that there exist biased estimates, like regularized estimates, that have a lower prediction error than the least squares estimate. Additionally, the ordinary least squares estimate requires the matrix $(\boldsymbol{x}^T \boldsymbol{x})$ to be positive definite. Otherwise, $(\boldsymbol{x}^T \boldsymbol{x})$ is not invertible. Matrices are only invertible if they are of full rank. For one, deviations from the full rank are due to either strong correlation among the d explaining variables or even linear dependencies. Moreover, especially in small n, large d situations, estimates of the covariance matrix have a rank at most equal to the size of the samples $n << d$. Then, regularization is needed to derive an estimate of full rank. There exist several strategies for regularization or penalization in regression. Since they aim at minimizing the prediction error, the most important ones are discussed in section Section **4.2**.

4.1.2 The decomposition of variance

The resulting minimal prediction error is

$$\mathrm{E}\left((Y - Y^\star)^2\right) = \sigma_Y^2 - \boldsymbol{\beta}^T \boldsymbol{\Sigma}\, \boldsymbol{\beta}\,.$$

Alternatively, this irreducible error may be written $\mathrm{E}\left((Y - Y^\star)^2\right) = \sigma_Y^2\,(1 - \Omega^2)$ where $\Omega = \mathrm{Cor}(Y, Y^\star)$ and

$$\Omega^2 = \boldsymbol{P}_{YX} \boldsymbol{P}^{-1} \boldsymbol{P}_{XY}$$

is the *squared multiple correlation coefficient*. Furthermore, $\mathrm{Cov}(Y, Y^\star) = \sigma_Y^2\,\Omega^2$ and $\mathrm{E}(Y^\star) = \mu_Y$. The expectation $\mathrm{E}\left((Y - Y^\star)^2\right) = \mathrm{Var}(Y - Y^\star)$ is also called the *unexplained variance* or *noise variance*. Together with the *explained variance* or *signal variance* $\mathrm{Var}(Y^\star) = \sigma_Y^2\,\Omega^2$ it adds up to the *total variance* $\mathrm{Var}(Y) = \sigma_Y^2$. Accordingly, the *proportion of explained variance* is

$$\frac{\mathrm{Var}(Y^\star)}{\mathrm{Var}(Y)} = \Omega^2$$

which indicates that Ω^2 is the central quantity for understanding both nominal prediction error and variance decomposition in the linear model. The *ratio of signal variance to noise variance* is

$$\frac{\mathrm{Var}(Y^\star)}{\mathrm{Var}(Y - Y^\star)} = \frac{\Omega^2}{1 - \Omega^2}\,.$$

A summary of these relations is given in Table **4.1**, along with the empirical error decomposition in terms of observed sum of squares.

If instead of the optimal parameters β_0 and $\boldsymbol{\beta}$ we employ $\beta_0' = \beta_0 + \Delta\beta_0$ and $\boldsymbol{\beta}' = \boldsymbol{\beta} + \Delta\boldsymbol{\beta}$ the minimal prediction error $\mathrm{E}\left((Y - Y^\star)^2\right)$ increases by the *model error*

$$ME(\Delta\beta_0, \Delta\boldsymbol{\beta}) = (\Delta\boldsymbol{\beta})^T\, \boldsymbol{\Sigma}\, \Delta\boldsymbol{\beta} + (\Delta\beta_0)^2\,. \tag{4.8}$$

The *relative model error* is the ratio of the model error and the irreducible error $\mathrm{E}\left((Y - Y^\star)^2\right)$.

Table 4.1: Variance decomposition in terms of squared multiple correlation Ω^2 and corresponding empirical sums of squares.

Level	Total variance	=	unexplained variance	+	explained variance
Population	$\mathrm{Var}(Y)$	=	$\mathrm{Var}(Y - Y^\star)$	+	$\mathrm{Var}(Y^\star)$
	σ_Y^2	=	$\sigma_Y^2\,(1-\Omega^2)$	+	$\sigma_Y^2\,\Omega^2$
Empirical	TSS	=	RSS	+	ESS
	$\sum_{l=1}^n (y_l - \bar{y})^2$	=	$\sum_{l=1}^n (y_l - \hat{y}_l)^2$	+	$\sum_{l=1}^n (\hat{y}_l - \bar{y})^2$
	$\mathrm{df} = n-1$		$\mathrm{df} = n-d-1$		$\mathrm{df} = d$

Abbreviations: $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$; df: degrees of freedom; TSS: total sum of squares; RSS: residual sum of squares; ESS: explained sum of squares.

4.1.3 Classical strategies for variable selection

A rudimental approach to variable selection in the linear model is based on a t-test that examines if the regression coefficients differ from zero. The test utilizes the distribution of the estimated regression coefficients. Under model Equation **4.1** with an error ϵ, that is normally distributed with $N(0, \mathrm{Var}(\epsilon))$, the estimated regression coefficients $\boldsymbol{b}$ follow a multivariate gaussian distribution

$$\boldsymbol{b} \sim N(\boldsymbol{\beta}, \boldsymbol{\Sigma}^{-1}\mathrm{Var}(\epsilon))\,.$$

From the decomposition of variance, as presented in Table **4.1**, it is obvious that $\mathrm{Var}(\epsilon)$ equals the unexplained variance given as $\sigma_Y^2(1-\Omega^2)$. To test the null hypothesis that the coefficient of variable i equals zero, i.e. $b_i = 0$, the following t-score vector is derived

$$\boldsymbol{\tau}_{XY} = \mathrm{diag}\{\boldsymbol{\Sigma}^{-1}\}^{-1/2}\,\boldsymbol{\beta}\,\sigma_Y^{-1}(1-\Omega^2)^{-1/2}\,\sqrt{\mathrm{df}} \tag{4.9}$$

where $\mathrm{df} = n - d - 1$ represents the degrees of freedom. Under the null hypothesis the estimate $\hat{\boldsymbol{\tau}}_{XY}(i)$ follows a t-distribution with $\mathrm{df} = n - d - 1$ degrees of freedom. Using this result it is possible to assign p-values to each variable. In Section **4.3** the connection of $\boldsymbol{\tau}_{XY}$ to partial correlation is discussed.

Stepwise selection comprises heuristic strategies to select variables based on statistics like $\hat{\boldsymbol{\tau}}_{XY}$ or alternatively the F- or Wald-statistic, see e.g. Fahrmeir and Tutz (2001) or Hastie et al. (2009). In backward selection all variables are included and $\hat{\boldsymbol{\tau}}_{XY}$ is computed for all variables, then the variable with the lowest value of $\hat{\boldsymbol{\tau}}_{XY}$ is excluded from the model and $\hat{\boldsymbol{\tau}}_{XY}$ is recomputed for the remaining variables. This step is repeated until there are only variables in the model that have a p-value below a predefined threshold. Such selection strategies suffer from unstable results due to dependencies among the predictor variables since the final model highly depends on the order of the excluded variables. For example, backward and forward selection usually do not agree on the same model (Burnham and Anderson, 2002).

A different approach to variable selection is taken by penalized RSS. Penalized RSS quantify the goodness of fit of a specific model with an additional penalty on the model size $q < d$. Using penalized RSS it is possible to compare models of different size and including different variables. Following George (2000) a general illustration of penalized RSS for a given model of size q is given as

$$\mathrm{RSS}_q^{\mathrm{pen}} = \mathrm{RSS}_q + \lambda \cdot q\widehat{\mathrm{Var}}(\epsilon) \tag{4.10}$$

where RSS_q is the RSS based on the model of dimension q and $\widehat{\mathrm{Var}}(\epsilon) = \frac{RSS}{n-d-1}$ is the estimated residual variance for the full model. The penalization parameter λ is fixed in advance and differs in:

- Akaike's Information Criterion (AIC)

$$\mathrm{RSS}_q^{\mathrm{AIC}} = \mathrm{RSS}_q + 2 \cdot q\widehat{\mathrm{Var}}(\epsilon),$$

- Bayesian Information Criterion (BIC)

$$\mathrm{RSS}_q^{\mathrm{BIC}} = \mathrm{RSS}_q + \log(n) \cdot q\widehat{\mathrm{Var}}(\epsilon),$$

- Risk Inflation Criterion (RIC)

$$\mathrm{RSS}_q^{\mathrm{RIC}} = \mathrm{RSS}_q + 2 \cdot \log(q) \cdot q \cdot \widehat{\mathrm{Var}}(\epsilon),$$

- (minimum) Mallowes' Cp

$$\mathrm{RSS}_q^{\mathrm{Cp}} = \mathrm{RSS}_q + 2 \cdot q\widehat{\mathrm{Var}}(\epsilon).$$

Variable selection using penalized RSS is widespread, still there are two drawbacks. First, the fixed choice of λ has a strong impact on the size of the selected model. Large values of λ favor a small model size and vice versa. In contrast to penalized regression, as discussed in Section **4.2**, the parameter λ is fixed and there is no intrinsic adaption to the data under analysis. Furthermore, penalized RSS is relatively sensitive with respect to small changes in the data (George, 2000).

4.2 Prediction: Penalized regression

Penalized regression strategies aim at minimizing a modified version of the RSS. Additionally to the ordinary RSS, a penalization term is appended that quantifies the size of the regression coefficients. A general representation of penalization is given as

$$\text{argmin}\left\{\text{RSS}(\boldsymbol{b}) + \lambda \sum_{i=1}^{d} \text{pen}(\boldsymbol{b}_i)\right\}$$

where $\text{RSS}(\boldsymbol{b})$ is the residual sum of squares as a function of $\boldsymbol{b}$ (see Equation **4.7**) and $\text{pen}(\boldsymbol{b}_i)$ denotes a penalty with respect to the estimated regression coefficients $\boldsymbol{b}$. Most commonly used are the absolute norm (lasso), the quadratic norm (ridge) or a mixture of both (elastic net) which are introduced in this section. Furthermore, two new approaches, the hyperlasso and minimax concave penalty, are presented. In contrast to the least squares solution regularized estimates are biased, but nonetheless may give a smaller prediction error. In the following the notions penalization, regularization and shrinkage are used interchangeable. The amount of regularization is governed by a shrinkage parameter λ, that is selected to minimize the estimated prediction error. Estimation of the prediction error is mostly performed by cross-validation. For more details see Hastie et al. (2009)[Chapter 7]. Since the parameter estimates are optimized with respect to prediction, regularized regression aims at selecting variables optimal for prediction. Note that the data is to be standardized before analysis. Thus, the intercept is not included in the penalization.

The first regularized regression proposed is ridge regression (Hoerl and Kennard, 1970). Ridge regression is characterized by a quadratic norm (L2) in the penalty

$$\text{argmin}\left\{\text{RSS}(\boldsymbol{b}) + \lambda \sum_{i=1}^{d} \boldsymbol{b}_i^2\right\} .$$

The ridge regression coefficients can be expressed in closed form as

$$\boldsymbol{b}^{ridge} = \boldsymbol{x}(\boldsymbol{x}^T\boldsymbol{x} + \lambda I)^{-1}\boldsymbol{x}^T\boldsymbol{y} .$$

From a Bayesian viewpoint the ridge estimate represents the mean or mode of the posterior distribution characterizing the regression coefficients. The posterior distribution derives from the assumption of a normal likelihood and a conjugate normal prior distribution for $\boldsymbol{\beta}$ with known variance, where $\beta_1, ..., \beta_d$ are assumed independent (Hastie et al., 2009). In contrast to a full Bayesian approach, the posteriori distribution is not explicitly derived, only the mean or mode are needed. Due to the shape of the prior no regression coefficients are forced to zero; thus no intrinsic variable selection takes place.

Lasso regression (Tibshirani, 1996) works with a penalty in absolute norm (L1)

$$\text{argmin}\left\{ \text{RSS}(\boldsymbol{b}) + \lambda \sum_{i=1}^{d} \mid \boldsymbol{b}_i \mid \right\} . \tag{4.11}$$

In contrast to ridge regression there is no closed form of the regression coefficients. As well as the ridge regression the lasso can be derived from a Bayesian representation. Here, the regression coefficients $\beta_1, ..., \beta_d$ (independently) follow a double exponential (laplace) prior distribution (Hoggart et al., 2008) that is a mixture of a normal distribution (N, with expectation β_i and variance σ^2) and a gamma distribution (Ga, with shape σ^2 and scale $\zeta^2/2$)

$$\begin{aligned} p(\beta_i \mid \zeta) &= \int_0^\infty N(\beta_i, \sigma^2) Ga(\sigma^2, \zeta^2/2)\, \partial\sigma^2 \\ &= \frac{\zeta}{2} \exp\{-\zeta \mid \beta_i \mid\} . \end{aligned}$$

Lasso regression automatically performs variable selection by pulling regression coefficients to zero and thus excluding them from the model. The L1 penalty encourages sparse solutions, i.e. only few variables enter a model. In case of correlated sets of variables only one variable is included as a representative for the whole set. Notably, the lasso poses a convex optimization problem. The LARS algorithm allows to compute an entire path of lasso coefficients for a set of regularization parameters (Efron et al., 2004).

The penalty in elastic net regression (Zou and Hastie, 2005) is a mixture of L1 and L2 norm

$$\text{argmin}\left\{ \text{RSS}(\boldsymbol{b}) + \lambda_1 \sum_{i=1}^{d} \mid \boldsymbol{b}_i \mid + \lambda_2 \sum_{i=1}^{d} \boldsymbol{b}_i^2 \right\} . \tag{4.12}$$

In contrast to the sparse solutions of lasso, the elastic net exhibits a grouping property that assigns two variables in strong correlation similar coefficients. Thus, sets of correlated variables are included or excluded jointly from a model. There exists a modification of the LARS algorithm, LARS-EN, to compute an entire path of elastic net coefficients for a set of regularization parameters (Zou and Hastie, 2005).

A generalization of the double exponential prior is the normal exponential gamma (NEG) prior that is used in a strategy called hyperlasso (Hoggart et al., 2008). The NEG prior is given as a mixture of a normal and two gamma distributions

$$\begin{aligned} p(\beta_i \mid \theta_1, \theta_2) &= \int_0^\infty \int_0^\infty N(\beta_i, \sigma^2)\, Ga(\sigma^2, \psi)\, Ga(\psi \mid \theta_1, \theta_2^2)\, \partial\sigma^2 \partial\psi \\ &= c \cdot \exp\{\beta_i^2/(4\theta_2^2)\}\, D_{-2\theta_1 - 1}\{\mid \beta_i \mid /\theta_2\} \end{aligned}$$

where c is an integrating constant and D is the parabolic cylinder function. Characteristic of the NEG prior is an even sharper peak around zero than the lasso, but wider tails. Thus, the hyperlasso favors even sparser solutions compared to the lasso and puts less shrinkage on larger coefficients. The hyperlasso estimates maximize the posterior distribution by applying the CLG algorithm (Hoggart et al., 2008).

There are two more strategies that attempt to minimize the estimation-bias, the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001) and the mimimax concave penalty (MCP) (Zhang, 2010). Both apply an identical penalization as the lasso for small values of the estimated β-coefficients and relax the penalization gradually for larger values to zero. Thus, regressions coefficients are penalized differently depending on their actual size. SCAD and MCP differ in the transition from lasso penalty to no penalization which is governed by two shrinkage parameters $\lambda_1 \geq 0$ and $\lambda_2 > 1$. Since the MCP penalty has given better numerical results than the SCAD (Breheny and Huang, 2011) it is considered here as the representative strategy with varying penalization strength. The MCP penalty is given as

$$\text{pen}^{\text{MCP}}_{\lambda_1,\lambda_2}(b_i) = \begin{cases} \lambda_1 b_i - \frac{b_i^2}{2\lambda_2} & \text{if } b_i \leq \lambda_2\lambda_1 \\ \frac{1}{2}\lambda_2\lambda_1^2 & \text{if } b_i > \lambda_2\lambda_1 \,. \end{cases}$$

In summary, MCP aims at variable selection like lasso without a strong shrinkage of the larger regression coefficients. Since those penalties are nonconvex a special coordinate descent algorithm is developed by Breheny and Huang (2011).

4.3 Ranking variables by importance

Ranking variables according to their importance is an intuitive strategy for variable selection, once the definition of "importance" and moreover the way of its quantification is established. Variable importance may be defined in many different ways, see Firth (1998) for an overview. Here, we consider a variable to be "important" if it is informative about the response and thus if its inclusion in the predictor *increases the explained variance* or, equivalently, *reduces the prediction error*. If there is no correlation among predictors (i.e. if $\boldsymbol{P} = \boldsymbol{I}$) there is general agreement that the marginal correlations $\boldsymbol{P}_{XY}$ provide an optimal way to rank variables (e.g. Fan and Lv, 2008). It is the correlation among predictor variables that facilitates the quantification of importance. So far, a large number of criteria to quantify the importance $\phi(x_i)$ of variable X_i have been suggested. Still, there is no consensus which criterion to use (Grömping, 2007).

Additionally, the discussion of variable importance is hindered by a dissent about desired properties of such a measure. Grömping (2007) lists the following four properties postulated for a measure of variable importance $\phi(x_i)$:

A. Decomposition of Ω^2:

$$\sum_{i=1}^{d} \phi(x_i) = \Omega^2$$

As shown in Section **4.1.2** the proportion of variance explained or the multiple correlation coefficient Ω^2 is the central element to assess the prediction quality of the linear model. A proper decomposition of Ω^2 elucidates which variables are important for the particular linear model, i.e. help to increase the variance explained or equivalently decrease the prediction error. Thus, $\phi(x_i)$ can quantify the "share" (Grömping, 2007) of a variable to the proportion of variance explained.

B. Non-negativity:
A variable should not be negatively important. Negative values in importance or shares are not interpretable.

C. Inclusion:

$$\text{If } \beta_i \neq 0, \text{ then } \phi(x_i) \neq 0$$

Inclusion implies that variables with a nonzero regression coefficient should be allocated a nonzero share.

D. Proper Exclusion:

$$\text{If } \beta_i = 0, \text{ then } \phi(x_i) = 0$$

Exclusion implies that variables with a zero regression coefficient should be of no importance.

Properties A-B are indispensable, but the exclusion property is under discussion. We argue that a zero variable X_i correlated to a nonzero variable is of higher importance than another zero variable X_j that is not related to a nonzero variable. Then, variable X_i contains, in contrast to X_j, at least some information on an interesting variable and thus $\phi(x_i) > \phi(x_j)$. A similar argumentation is found in Grömping (2007). If nonzero variables are uncorrelated with the noise variables this discussion is obsolete. Moreover, we argue that inclusion property leads to adverse effects in the presence of antagonistic variables, i.e when highly correlated variables have opposing effects on the outcome ($\beta_i = -\beta_j$). Including both variables cancels out the effect of both despite β_i and β_j being nonzero. In Section **6.1** we discuss the inclusion property in a case study of a data set that includes antagonistic variables.

In addition, we require:

E. Reduction to marginal correlation:
Since marginal correlation is the optimal criterion in case of no correlation, the measure $\phi(x_i)$ should reduce to the (squared) marginal correlation in case of uncorrelated predictors.

F. Orthogonality:
A special covariance structure is given if the predictor variables are divided into orthogonal subgroups; i.e. the correlation matrix $\boldsymbol{P}$ is characterized by a block structure. Then, orthogonality implies that the sum of the $\phi(x_i)$ of all variables X_i within a block is equal to the squared multiple correlation coefficient of that block with the response.

In the following several of the existing quantities for $\phi(x_i)$ are discussed.

- **Marginal correlation**

  If there is *no correlation among predictors* (i.e. if $\boldsymbol{P} = \boldsymbol{I}$) then there is general agreement that the marginal correlations $\boldsymbol{P}_{XY} = (\rho_1, \ldots, \rho_d)^T$ provide an optimal way to rank variables (e.g. Fan and Lv, 2008). In this special case the standardized predictor equation (Equation **4.5**) simplifies to

  $$Y^{\star}_{\text{std}} = \boldsymbol{P}^T_{XY} \boldsymbol{X}_{\text{std}} \,.$$

  For $\boldsymbol{P} = \boldsymbol{I}$ the marginal correlations represent the influence of each standardized covariate in predicting the standardized response. Moreover, in this case the sum of the squared marginal correlations $\Omega^2 = \sum_{i=1}^{d} \rho_i^2$ equals the squared multiple correlation coefficient. Thus, the contribution of each variable X_i to reducing relative prediction error is ρ_i^2 — recall from Table **4.1** that $\text{Var}(Y - Y^{\star})/\sigma_Y^2 = 1 - \Omega^2$. For this reason in the uncorrelated setting

  $$\phi^{\text{uncorr}}(x_i) = \rho_i^2$$

  is justifiably the canonical measure of variable importance for X_i.

  However, for general $\boldsymbol{P}$, i.e. in the presence of correlation among predictors, the squared marginal correlations do not provide a decomposition of Ω^2 as $\boldsymbol{P}^T_{XY}\boldsymbol{P}_{XY} \neq \Omega^2$. Thus, they are not suited as a general variable importance criterion.

- **Standardized regression coefficients**

  From Equation **4.5** one may consider standardized regression coefficients $\boldsymbol{\beta}_{\text{std}}$ (Equation **4.6**) as generalization of marginal correlations to the case of correlation among predictors. However, while the $\boldsymbol{\beta}_{\text{std}}$ properly reduce to marginal correlations for $\boldsymbol{P} = \boldsymbol{I}$ the standardized regression coefficients do not lead to a decomposition of Ω^2 as

$\boldsymbol{\beta}_{\text{std}}^T \boldsymbol{\beta}_{\text{std}} = \boldsymbol{P}_{YX} \boldsymbol{P}^{-2} \boldsymbol{P}_{XY} \neq \Omega^2$. Further objections to using $\boldsymbol{\beta}_{\text{std}}$ as a measure of variable importance are discussed in Bring (1994).

- **Partial correlation**

Another common way to rank predictor variables and to assign p-values is by means of t-scores $\boldsymbol{\tau}_{XY} = (\tau_1, \ldots, \tau_d)^T$ (which in some texts are also called standardized regression coefficients even though they are not to be confused with $\boldsymbol{\beta}_{\text{std}}$). As presented in Equation **4.9** the t-scores are directly computed from regression coefficients via

$$\begin{aligned} \boldsymbol{\tau}_{XY} &= \text{diag}\{\boldsymbol{P}^{-1}\}^{-1/2} \boldsymbol{\beta}_{\text{std}} (1 - \Omega^2)^{-1/2} \sqrt{\text{df}} \\ &= \text{diag}\{\boldsymbol{\Sigma}^{-1}\}^{-1/2} \boldsymbol{\beta} \sigma_Y^{-1} (1 - \Omega^2)^{-1/2} \sqrt{\text{df}}. \end{aligned} \tag{4.13}$$

The constant df is the degree of freedom and diag$\{\boldsymbol{M}\}$ the matrix $\boldsymbol{M}$ with its off-diagonal entries set to zero.

Completely equivalent to t-scores in terms of variable ranking are the *partial correlations* $\tilde{\boldsymbol{P}}_{XY} = (\tilde{\rho}_1, \ldots, \tilde{\rho}_d)^T$ between the response Y and predictor X_j conditioned on all the remaining predictors $X_{\neq j}$. The t-scores can be converted to partial correlations using the relationship

$$\tilde{\rho}_i = \tau_i / \sqrt{\tau_i^2 + \text{df}}\ .$$

Interestingly, the value of df specified in the t-scores cancels out when computing $\tilde{\rho}_j$. An alternative but equivalent route to obtain the partial correlations is by inversion and subsequent standardization of the joined correlation matrix of Y and $\boldsymbol{X}$ (e.g. Opgen-Rhein and Strimmer, 2007b).

The p-values computed in many statistical software packages for each variable in a linear model are based on empirical estimates of $\boldsymbol{\tau}_{XY}$ with df $= n - d - 1$. Assuming normal $\boldsymbol{X}$ and Y the null distribution of the estimated t-score follows the Student t-distribution with $n - d - 1$ degrees of freedom. Exactly the same p-values are obtained from the empirical partial correlations $\tilde{r}_i$ which have null-density $f(\tilde{r}_i) = |\tilde{r}_i| \,\text{Beta}\left(\tilde{r}_i^2; \frac{1}{2}, \frac{\kappa-1}{2}\right)$ with $\kappa = \text{df} + 1 = n - d$ and $\text{Var}(\tilde{r}_j) = \frac{1}{\kappa}$.

Despite being widely used, a key problem of partial correlations $\tilde{\boldsymbol{P}}_{XY}$ (and hence also of the corresponding t-scores) for use in variable ranking and assigning variable importance is that in the case of vanishing correlation $\boldsymbol{P} = \boldsymbol{I}$ they do *not* properly reduce to the marginal correlations $\boldsymbol{P}_{XY}$. This can be seen already from the simple case with three variables Y, X_1, and X_2 with partial correlation

$$\rho_{Y,X_1|X_2} = \frac{\rho_{Y,X_1} - \rho_{Y,X_2}\rho_{X_1,X_2}}{\sqrt{1-\rho^2_{Y,X_2}}\sqrt{1-\rho^2_{X_1,X_2}}}$$

which for $\rho_{X_1,X_2} = 0$ is not identical to ρ_{Y,X_1} unless ρ_{Y,X_2} also vanishes.

- **Hoffman-Pratt product measure**

  First suggested by Hoffman (1960) and later defended by Pratt (1987) is the following alternative measure of variable importance

  $$\phi^{\mathrm{HP}}(x_i) = (\boldsymbol{\beta}_{\mathrm{std}})_i \, \rho_i = (\boldsymbol{P}^{-1}\boldsymbol{P}_{\boldsymbol{X}Y})_i \, \rho_i \, .$$

  By construction, $\sum_{i=1}^{d} \phi^{\mathrm{HP}}(x_i) = \Omega^2$, and if correlation among predictors is zero then $\phi^{\mathrm{HP}}(x_i) = \rho_i^2$. Moreover, the Hoffman-Pratt measure satisfies the orthogonal compatibility criterion (Genizi, 1993).

  However, in addition to these desirable properties the Hoffman-Pratt variable importance measure also exhibits two severe defects. First, $\phi^{\mathrm{HP}}(x_i)$ may become negative, and second the relationship of the Hoffman-Pratt measure with the original predictor equation is unclear. Therefore, the use of $\phi^{\mathrm{HP}}(x_i)$ is discouraged by most authors (cf. Grömping, 2007).

- **Genizi's measure**

  More recently, Genizi (1993) proposed the variable importance measure

  $$\phi^{\mathrm{G}}(x_i) = \sum_{j=1}^{d} \left((\boldsymbol{P}^{1/2})_{ij} \, (\boldsymbol{P}^{-1/2}\boldsymbol{P}_{\boldsymbol{X}Y})_j \right)^2 \, .$$

  Here and in the following $\boldsymbol{P}^{1/2}$ is the uniquely defined matrix square root with $\boldsymbol{P}^{1/2}$ symmetric and positive definite. See Section **5.1** for more information.

  Genizi's measure provides a decomposition $\sum_{i=1}^{d} \phi^{\mathrm{G}}(x_i) = \Omega^2$, reduces to the squared marginal correlations in case of no correlation, and obeys the orthogonality criterion. In contrast to $\phi^{\mathrm{HP}}(x_i)$ the Genizi measure is by construction also non-negative. However, as with the Hoffman-Pratt measure the connection of $\phi^{\mathrm{G}}(x_i)$ with the original predictor equation is unclear.

- **Quantities based on averaging over all possible orderings**

  Some authors (e.g. Kruskal, 1987) propose to quantify the importance of a variable by computing Ω^2 for all possible orderings of predictor variables $\boldsymbol{X}$ and then, averaging over all orderings. Since such approaches are computational feasible only for very small sets of variables, they are omitted here.

- **Variable importance in regression trees**

  Random forests (Breiman, 2001) are not only used for prediction, they also offer a measure of variable importance as a side-product. A random forest is an ensemble of regression or classification trees that are trained on different bootstrap samples or subsamples of the training data. Prediction is based on an average over all trees. Permutation of a variable X_i breaks a possible relationship with the trait of interest and thus is used to fake the absence of the variable from the model. Variable importance is then defined as the decrease in prediction accuracy before and after permutation. For more details, especially on the algorithms see Breiman (2001). Random forests tend to exhibit a biased behavior of selecting variables (Strobl et al., 2007). Especially, when there are variables with differing scales or in case of categorical variables different number of classes, random forests tend to prefer some variables only due to their data structure. A conditional criterion for variable importance is presented by Strobl et al. (2008) to overcome this bias.

To summarize, so far there is no measure for variable importance that

- is defined in the linear model, and
- is applicable to high-dimensional data, and
- gives a clear connection with the original predictor equation (Equation **4.5**), and
- is non-negative, and
- reduces to marginal correlation in case of uncorrelated predictors, and
- decomposes Ω^2.

4.4 Decorrelation: The CAR score

In the following section we introduce a novel quantity for variable importance that facilitates the selection of variables under correlation in the linear model. First, we define the CAR score and sketch how it relates to several quantities in the linear model. Notably, it is an intermediate between marginal correlation and standardized regression coefficients. The CAR score is derived from a predictive point of view, it is the weight of a decorrelated (and standardized) variable on the best linear predictor. Furthermore, we show that it meets most properties requested for criteria of variable importance. Especially, the CAR score decomposes the proportion of variance explained.

In an extensive simulation study we show that CAR scores are applicable to high-dimensional data and most of all outperform established strategies in terms of prediction error and number of true positives selected. Finally, we demonstrate that the CAR score in regression is the analogon to the CAT score in classification.

4.4.1 Definition of the CAR score

The CAR scores $\boldsymbol{\omega}$ are defined as

$$\boldsymbol{\omega} = \boldsymbol{P}^{-1/2} \boldsymbol{P}_{XY} \tag{4.14}$$

i.e. as the marginal correlations $\boldsymbol{P}_{XY}$ adjusted by the factor $\boldsymbol{P}^{-1/2}$. Accordingly, the acronym "CAR" is an abbreviation for correlation-adjusted (marginal) correlation since it is the correlation $\boldsymbol{P}_{XY}$ between $\boldsymbol{X}$ and Y adjusted for the correlation $\boldsymbol{P}$ among $\boldsymbol{X}$. Decorrelation is performed by the Mahalanobis transform of the correlation matrix of $\boldsymbol{X}$, analogically to the CAT score. See Section **5.1** for more details of the Mahalanobis transform. The CAR scores $\boldsymbol{\omega}$ are constant population quantities and not random variables.

Table **4.2** summarizes some connections of CAR scores with various other quantities from the linear model. For instance, CAR scores may be viewed as intermediates between marginal correlations and standardized regression coefficients, as demonstrated in Figure **4.4.1**. If correlation among predictors vanishes the CAR scores become identical to the marginal correlations.

Further insights into the interpretation of CAR scores can be gained by a comparison with partial correlation. The partial correlation between Y and a predictor X_i is obtained by first removing the linear effect of the remaining $d-1$ predictors $X_{\neq i}$ from both Y and X_i and subsequently computing the correlation between the respective remaining residuals. In contrast, with CAR scores the response Y is left unchanged whereas all d predictors are simultaneously orthogonalized, i.e. the linear effect of the other variables

$$\begin{tikzcd}
\boxed{\text{correlation}} \arrow[r, "\boldsymbol{P}^{-1/2}"] & \boxed{\quad\boldsymbol{\omega}\quad} \arrow[r, "\boldsymbol{P}^{-1/2}"] & \boxed{\quad\boldsymbol{\beta}_{std}\quad} \\
\boldsymbol{P}_{XY} & \boldsymbol{P}^{-1/2}\boldsymbol{P}_{XY} & \boldsymbol{P}^{-1}\boldsymbol{P}_{XY}
\end{tikzcd}$$

Figure 4.1: The CAR score as an intermediate between marginal correlation and standardized regression coefficients.

Table 4.2: Relationship between CAR scores $\boldsymbol{\omega}$ and common quantities from the linear model.

Criterion	Relationship with CAR scores $\boldsymbol{\omega}$
Regression coefficient	$\boldsymbol{b} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\omega}\,\sigma_Y \quad \leftrightarrow \quad \boldsymbol{\omega} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{b}\,\sigma_Y^{-1}$
Std regression coeff.	$\boldsymbol{b}_{\text{std}} = \boldsymbol{P}^{-1/2}\boldsymbol{\omega} \quad \leftrightarrow \quad \boldsymbol{\omega} = \boldsymbol{P}^{1/2}\boldsymbol{b}_{\text{std}}$
Marginal correlation	$\boldsymbol{P}_{XY} = \boldsymbol{P}^{1/2}\boldsymbol{\omega} \quad \leftrightarrow \quad \boldsymbol{\omega} = \boldsymbol{P}^{-1/2}\boldsymbol{P}_{XY}$
Regression t-score	$\boldsymbol{\tau}_{XY} = (\boldsymbol{P}\,\text{diag}\{\boldsymbol{P}^{-1}\})^{-1/2}\,\boldsymbol{\omega}\,(1-\boldsymbol{\omega}^T\boldsymbol{\omega})^{-1/2}\sqrt{\text{df}}$

$X_{\neq i}$ on X_i is removed simultaneously from all predictors (Hyvärinen et al., 2001, Section 6.5). Subsequently, the CAR score is found as the correlation between the "residuals", i.e. the unchanged response and the decorrelated predictors. Thus, CAR scores may be viewed as a multivariate variant of the so-called part correlations.

4.4.2 Derivation from the best linear predictor

Using CAR scores the (standardized) best linear predictor (Equation **4.5**) can be written in the simple form

$$Y^\star_{\text{std}} = \boldsymbol{\omega}^T\delta(\boldsymbol{X}) = \sum_{i=1}^{d}\omega_i\delta_i(\boldsymbol{X})\,, \tag{4.15}$$

where

$$\delta(\boldsymbol{X}) = \boldsymbol{P}^{-1/2}\boldsymbol{V}^{-1/2}(\boldsymbol{X}-\boldsymbol{\mu}) = \boldsymbol{P}^{-1/2}\boldsymbol{X}_{\text{std}}$$

are the Mahalanobis-decorrelated and standardized predictors with $\text{Var}(\delta(\boldsymbol{X})) = \boldsymbol{I}$, and

$$\boldsymbol{\omega} = \boldsymbol{P}^{-1/2}\,\boldsymbol{P}_{XY}$$

is the CAR score vector as defined in Equation **4.14**. Thus, the CAR scores $\boldsymbol{\omega}$ are the weights that describe the influence of each decorrelated and standardized variable in predicting the standardized response. Furthermore, with $\mathrm{Cor}(\boldsymbol{X}_{\mathrm{std}}, Y) = \boldsymbol{P}_{XY}$ we have

$$\boldsymbol{\omega} = \mathrm{Cor}(\delta(\boldsymbol{X}), Y)$$

that is CAR scores are the correlations between the response and the decorrelated covariates.

4.4.3 The decomposition of variance in terms of CAR scores

As demonstrated in Section **4.1.2** the decomposition of total variance into explained and unexplained variance is the pivotal element of linear regression. In particular, the decomposition can be expressed in terms of CAR scores. Using Equation **4.15** the explained variance, as the variance of the best linear predictor, is rewritten to

$$\begin{aligned}\mathrm{Var}(Y^\star) &= \sigma_Y^2 \mathrm{Var}\left(\boldsymbol{\omega}^T \delta(\boldsymbol{X})\right) = \sigma_Y^2 \boldsymbol{\omega}^T \underbrace{\mathrm{Var}(\delta(\boldsymbol{X}))}_{\boldsymbol{I}_d} \boldsymbol{\omega} \\ &= \sigma_Y^2 \boldsymbol{\omega}^T \boldsymbol{\omega}\,.\end{aligned}$$

Consequently, the nominal mean squared prediction error in terms of CAR scores can be expressed as

$$\mathrm{Var}(Y - Y^\star) = \mathrm{E}\left((Y - Y^\star)^2\right) = \sigma_Y^2 \left(1 - \boldsymbol{\omega}^T \boldsymbol{\omega}\right).$$

Thus, (decorrelated) variables with small CAR scores contribute little to improve the prediction error or to reduce the unexplained variance. Altogether, the decomposition of total variance $\mathrm{Var}(Y) = \sigma_Y^2$ into explained and unexplained variance can be rewritten in terms of CAR scores

$$\begin{aligned}\overbrace{\mathrm{Var}(Y)}^{\text{Total variance}} &= \overbrace{\mathrm{Var}(Y^\star)}^{\text{Explained variance}} + \overbrace{\mathrm{Var}(Y - Y^\star)}^{\text{Unexplained variance}} \\ \sigma_Y^2 &= \sigma_Y^2(\boldsymbol{\omega}^T\boldsymbol{\omega}) + \sigma_Y^2(1 - \boldsymbol{\omega}^T\boldsymbol{\omega})\,.\end{aligned} \tag{4.16}$$

We argue that the CAR score is the central quantity to assess which variables contribute to the explained variance or equivalently reduce the unexplained variance. Moreover, it is apparent by Equation **4.16** that the proportion of variance explained simplifies to

$$\begin{aligned}\frac{\text{Explained Variance}}{\text{Total Variance}} &= \frac{\sigma_Y^2(\boldsymbol{\omega}^T\boldsymbol{\omega})}{\sigma_Y^2} \\ &= \boldsymbol{\omega}^T\boldsymbol{\omega} = \sum_{i=1}^d \omega_i^2\,.\end{aligned} \tag{4.17}$$

Evidently, the sum of squared CAR-scores adds up to the proportion of variance explained or squared multiple correlation coefficient. Due to this decomposition, we argue that the CAR score is the natural criterion to assess the contribution of a variable to the proportion of variance explained. Note that in the set-up of discriminant analysis the sum of squared CAT-scores adds up to Hotelling's T^2. This suggests that Hotelling's T^2 in classification is the analogon to the proportion of variance explained in the linear model.

4.4.4 The CAR score as quantity for variable importance

The decomposition of Ω^2 in Equation **4.17** suggests to define

$$\phi^{\text{car}}(x_i) = \omega_i^2$$

as a measure of variable importance. $\phi^{\text{car}}(x_i)$ is always non-negative [B.], reduces to ρ_i^2 for uncorrelated explanatory variables [E.], and leads to the canonical decomposition of the multiple correlation coefficient [A.]

$$\Omega^2 = \sum_{i=1}^{d} \phi^{\text{car}}(x_i)\,.$$

Furthermore, it is easy to see that $\phi^{\text{car}}(x_i)$ satisfies the orthogonal compatibility criterion [F.] demanded in Genizi (1993). Interestingly, Genezi's own importance measure $\phi^{\text{G}}(x_i)$ can be understood as a weighted average $\phi^{\text{G}}(x_i) = \sum_{k=1}^{d} (\boldsymbol{P}^{1/2})_{ik}^2 \phi^{\text{car}}(x_k)$ of squared CAR scores. In short, what we propose here is to first Mahalanobis-decorrelate the predictors to establish a canonical basis, and subsequently we define the importance of a variable X_i as the natural weight ω_i^2 in this reference frame.

As we will further discuss in Section **4.5**, the CAR score does not meet the exclusion property [D.] since it assigns zero variables correlated to a nonzero variable a share unequal zero. Due to the antagonistic property, covered in Section **4.4.5**, the CAR score also does not meet the inclusion property [C.], since the CAR scores of two positively correlated nonzero variables with opposing β-coefficients tend to zero. This is motivated by the fact that including both of the variables in the model results in canceling out the effect of both. In Section **6.1** we illustrate the antagonistic property on a reference data set on the progression of diabetes and discuss how different approaches on variable selection handle antagonistic variables.

To sum up, the CAR score meets the most important properties for quantities of variable importance, i.e. it decomposes the proportion of variance explained, satisfies the orthogonal compatibility, is non-negative, and reduces to marginal correlation in case of uncorrelated predictors. In contrast, the CAR score does *not* meet the inclusion and exclusion property. But we argue that there exist settings when both properties may be disadvantageous in the multivariate framework of the linear model. First, the exclusion property is not met if a zero variable correlated to a nonzero variable is allocated a share unequal zero. Still, it might be more helpful to include the nonzero variable because it provides at least partial information on the variable of interest. Second, the inclusion property is adverse in case of two antagonistic variables since including both cancels out the effect of both.

4.4.5 More properties of the CAR score

Beyond the concept of variable importance the CAR exhibits an additive property beneficial for the analysis of sets of variables. Then we present exploratory and theoretical tools that give guidance on how to select the model size. Moreover, we report grouping, antagonistic, and oracle properties that illustrate the behavior or CAR scores in variable selection.

- **Grouped CAR score**

  Due to the additivity of squared CAR scores it is straightforward to define a *grouped* CAR score for a set of variables as the sum of the individual squared CAR scores

  $$\omega_{\text{grouped}} = \sqrt{\sum_{g \in \text{set}} \omega_g^2}\,.$$

  As with the grouped CAT score, presented in Section **3.3.3**, we also may add a sign in this definition. An estimate of the squared grouped CAR score is an example of a simple global test statistic that may be useful if prespecified set of variables exist as e.g. in studying gene set enrichment (e.g. Ackermann and Strimmer, 2009).

- **Accumulated squared CAR score**

  Another related summary is the *accumulated squared CAR score* Ω_q^2 for the largest $q \leq d$ predictors. Arranging the CAR scores in decreasing order of absolute magnitude $\omega_{(1)}, \ldots, \omega_{(d)}$ with $\omega_{(1)}^2 > \ldots > \omega_{(d)}^2$ this can be written as

  $$\Omega_q^2 = \sum_{i=1}^{q} \omega_{(i)}^2\,. \tag{4.18}$$

  A plot of the accumulated CAR scores is an explanatory tool to visualize how much variance can be explained by the most important predictors. Typically, the most important variables have the largest shares of the proportion of variance explained.

Table 4.3: Threshold parameter λ for some classical model selection procedures.

Criterion	Reference	Penalty parameter
AIC	Akaike (1974)	$\lambda = 2$
C_p	Mallows (1973)	$\lambda = 2$
BIC	Schwarz (1978)	$\lambda = \log(n)$
RIC	Foster and George (1994)	$\lambda = 2\log(d)$

- **CAR scores and information criteria for model selection**

  CAR scores define a canonical ordering of the explanatory variables. Thus, variable selection using CAR scores is a simple matter of thresholding (squared) CAR scores. Intriguingly, this provides a direct link to model selection procedures using information criteria, as presented in Section **4.1.3**. Using Table **4.1** we rewrite the penalized RSS (Equation **4.10**) as

$$\begin{aligned}\mathrm{RSS}_q^{\mathrm{pen}} &= \mathrm{RSS}_q + \lambda \cdot q\widehat{\mathrm{Var}}(\epsilon)\\ &= n\hat{\sigma}_Y^2(1-\hat{\Omega}_q^2) + \lambda \cdot q\hat{\sigma}_Y^2(1-R^2)\end{aligned}$$

where $\hat{\Omega}_q^2$ is the sum of the q accumulated squared (estimated) CAR scores in the smaller model as described in Equation **4.18** and R^2 is the estimated proportion of variance explained in the full model. This connection further simplifies to

$$\begin{aligned}\frac{\mathrm{RSS}_q^{\mathrm{pen}}}{n\hat{\sigma}_Y^2} &= 1-\widehat{\Omega}_q^2 + \frac{\lambda q(1-R^2)}{n}\\ &= 1-\sum_{j=1}^{q}\left(\hat{\omega}_{(j)}^2 - \frac{\lambda(1-R^2)}{n}\right).\end{aligned} \tag{4.19}$$

This quantity decreases with q as long as $\hat{\omega}_{(q)}^2 > \hat{\omega}_c^2 = \frac{\lambda(1-R^2)}{n}$. Therefore, in terms of CAR scores classical model selection is equivalent to thresholding $\hat{\omega}_j^2$ at critical level $\hat{\omega}_c^2$, where predictors with $\hat{\omega}_j^2 \leq \hat{\omega}_c^2$ are removed. If n is large or for a perfect fit ($R^2 = 1$) all predictors are retained. More information on how to determine a cut-off is given in Section **5.3**.

- **Grouping property**

  A favorable feature of the elastic net procedure for variable selection is the grouping property which enforces the simultaneous selection of highly correlated predictors (Zou and Hastie, 2005). Model selection using CAR scores also exhibits the grouping property because predictors that are highly correlated have nearly identical CAR scores. This can directly be seen from the definition $\boldsymbol{\omega} = \boldsymbol{P}^{1/2}\boldsymbol{\beta}_{\text{std}}$ of the CAR score. For two predictors X_1 and X_2 and correlation $\text{Cor}(X_1, X_2) = \rho$ a simple algebraic calculation shows that the difference between the two squared CAR scores equals

  $$\omega_1^2 - \omega_2^2 = \left((\beta_{\text{std}})_1^2 - (\beta_{\text{std}})_2^2 \right) \sqrt{1-\rho^2}\,.$$

  Therefore, the two squared CAR scores become identical with growing absolute value of the correlation between the variables. This grouping property is intrinsic to the CAR score itself and not a property of an estimator.

- **Canceling out antagonistic variables**

  In addition to the grouping property the CAR score also exhibits an important behavior with regard to antagonistic variables. If the regression coefficients of two variables have opposing signs and these variables are in addition positively correlated then the corresponding CAR scores decrease to zero. For example, with $(\boldsymbol{\beta}_{\text{std}})_2 = -(\boldsymbol{\beta}_{\text{std}})_1$ we get

  $$\omega_1 = -\omega_2 = (\boldsymbol{\beta}_{\text{std}})_1 \sqrt{1-\rho}\,.$$

  This implies that antagonistic positively correlated variables will be bottom ranked. A similar effect occurs for protagonistic variables that are negatively correlated, as with $(\boldsymbol{\beta}_{\text{std}})_1 = (\boldsymbol{\beta}_{\text{std}})_2$ we have

  $$\omega_1 = \omega_2 = (\boldsymbol{\beta}_{\text{std}})_1 \sqrt{1+\rho}$$

  which decreases to zero for large negative correlation (i.e. for $\rho \to -1$).

- **An oracle version**

  Further insight into the CAR score is obtained by considering an "oracle version" where it is known in advance which predictors are truly non-null. Specifically, we assume that the regression coefficients can be written as

  $$\boldsymbol{\beta}_{\text{std}} = \begin{pmatrix} \boldsymbol{\beta}_{\text{std, non-null}} \\ 0 \end{pmatrix}$$

and that there is no correlation between null and non-null variables so that the correlation matrix $\boldsymbol{P}$ has block-diagonal structure

$$\boldsymbol{P} = \begin{pmatrix} \boldsymbol{P}_{\text{non-null}} & 0 \\ 0 & \boldsymbol{P}_{\text{null}} \end{pmatrix} .$$

The resulting oracle CAR score

$$\boldsymbol{\omega} = \boldsymbol{P}^{1/2}\boldsymbol{\beta}_{\text{std}} = \begin{pmatrix} \boldsymbol{\omega}_{\text{non-null}} \\ 0 \end{pmatrix}$$

is exactly zero for the null variables. Therefore, asymptotically the null predictors will be identified by the CAR score with probability one as long as the employed estimator is consistent.

4.4.6 Estimation

To repeat, the CAR score, as presented in Equation **4.14**, is a *population quantity*. In practice the CAR score needs to be estimated, more precisely, suitable estimates $\boldsymbol{R}$ and $\boldsymbol{R}_{XY}$ of the two correlation matrices $\boldsymbol{P}$ and $\boldsymbol{P}_{XY}$ need to be devised.

If there are more observations than variables, $n \gg d$, empirical estimates for correlation can be used, like the sample correlation estimate. Otherwise regularized estimates need to be employed. Here, we concentrate on the shrinkage estimate $\boldsymbol{R}^{\text{shrink}}$ for the correlation matrix $\boldsymbol{P}$ as presented in (Schäfer and Strimmer, 2005) and already discussed in Section **3.3.4**. Using also the shrinkage estimate $\boldsymbol{R}_{XY}^{\text{shrink}}$ for the correlation $\boldsymbol{P}_{XY}$ the shrinkage CAR score estimate is given by

$$\hat{\boldsymbol{\omega}}_{\text{shrink}} = (\boldsymbol{R}^{\text{shrink}})^{-1/2}\, \boldsymbol{R}_{XY}^{\text{shrink}} . \tag{4.20}$$

Note that the shrinkage estimate has the special property of providing an intermediate between the (marginal) shrinkage correlation $\boldsymbol{R}_{XY}^{\text{shrink}}$ and the original CAR score $\boldsymbol{R}^{-1/2}\boldsymbol{R}_{XY}^{\text{shrink}}$. The shrinkage parameter λ is set in a data-driven fashion, i.e. if the estimates of the correlations are highly variable, and thus insecure, λ tends to one and no correlation among $\boldsymbol{X}$ is incorporated.

More information on the Mahalanobis transform is presented in Section **5.1**. An efficient algorithm for calculating the inverse matrix square-root $\boldsymbol{R}^{-1/2}$ for the shrinkage correlation estimator is described in Section **5.2**.

4.5 Simulation studies

4.5.1 Design of the simulation study

In our simulations we broadly followed the setup employed in Zou and Hastie (2005), Witten and Tibshirani (2009) and Wang et al. (2011).

Specifically, we considered the following scenarios:

- *Example 1:* 8 variables with $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$. The predictors exhibit autoregressive correlation with $\text{Cor}(X_i, X_j) = 0.5^{|i-j|}$.

- *Example 2:* As Example 1 but with $\text{Cor}(X_i, X_j) = 0.85^{|i-j|}$.

- *Example 3:* 40 variables with $\boldsymbol{\beta} = (3, 3, 3, 3, 3, -2, -2, -2, -2, -2, 0, \dots, 0)^T$. The correlation between all pairs of the first 10 variables is set to 0.9, and otherwise set to 0.

- *Example 4:* 40 variables with $\boldsymbol{\beta} = (3, 3, -2, 3, 3, -2, 0, \dots, 0)^T$. The pairwise correlations among the first three variables and among the second three variables equals 0.9 and is otherwise set to 0.

The intercept was set to $\beta_0 = 0$ in all scenarios. We generated n samples $\boldsymbol{x}_l$ by drawing from a multivariate normal distribution with unit variances, zero expectations, and correlation structure $\boldsymbol{P}$ as indicated for each simulation scenario. To compute $y_l = \boldsymbol{\beta}^T \boldsymbol{x}_l + \varepsilon_l$, where $l \in 1, ..., n$, we sampled the error ε_l from a normal distribution with zero mean and standard deviation σ (so that $\text{Var}(\varepsilon) = \text{Var}(Y - Y^\star) = \sigma^2$). In Examples 1 and 2 the dimension is $d = 8$ and the sample sizes considered were $n = 50$ and $n = 100$ to represent a large sample setting. In contrast, for Examples 3 and 4 the dimension is $d = 40$ and sample sizes were small (from $n = 10$ to $n = 100$). In order to vary the ratio of signal and noise variances we used different degrees of unexplained variance ($\sigma = 1$ to $\sigma = 6$). For fitting the regression models we employed a training data set of size n. The tuning parameter of each approach was optimized using an additional independent validation data set of the same size n. In the CAR, partial correlation (PCOR) and Genizi approach the tuning parameter corresponds directly to the number of included variables, whereas for elastic net, lasso, and boosting the tuning parameter(s) correspond(s) to a regularization parameter.

For each estimated set of regression coefficients $\boldsymbol{b}$ we computed the model error (Equation **4.8**) and the model size. All simulations were repeated 200 times, and the average relative model error as well as the mean true and mean false positives were reported. For estimating CAR scores and associated regression coefficients we used in the large sample cases (Examples 1 and 2) the empirical estimator and and otherwise (Examples 3 and 4) shrinkage estimates.

For comparison we fitted in our study lasso (LASSO) and elastic net (E(LASTIC) NET) regression models using the algorithms available in the R package `scout` (Witten and Tibshirani, 2009). In addition, we employed the boosting algorithm (BOOST) for linear models as implemented in the R package `mboost` (Hothorn and Bühlmann, 2006), ordinary least squares with no variable selection (OLS), with partial correlation ranking (PCOR) and with variable ranking by the Genizi method (GENIZI).

4.5.2 Results from the simulation study

The results are summarized in Table **4.4** and Table **4.5**. In all investigated scenarios model selection by CAR scores is competitive with elastic net regression, and typically outperforms the lasso and OLS with no variable selection and OLS with variable selection by partial correlation. It is also in most cases distinctively better than boosting. Genizi's variable selection criterion also performs very well, with a similar performance to CAR scores in many cases, except for Example 2. Table **4.4** and Table **4.5** also show the true and false positives for each method. The regression models selected by the CAR score approach often exhibit the largest number of true positives and the smallest number of false positives, which explains its effectiveness.

Figure **4.2** shows the distribution of the estimated regression coefficients for the investigated methods over the 200 repetitions for Example 3 with $n = 50$ and $\sigma = 3$. This figure demonstrates that using CAR scores — unlike lasso, elastic net, and boosting — recovers the regression coefficients of variables X_6 to X_{10} that have negative signs. Moreover, in this setting the CAR score regression coefficients have a much smaller variability than those obtained using the OLS-Genizi or PCOR method.

The simulations for Examples 1 and 2 represent cases where the null variables X_3, X_4, X_6, X_7, and X_8 are correlated with the non-null variables X_1, X_2 and X_5. In such a setting the variable importance $\phi^{\text{CAR}}(x_i)$ assigned by squared CAR scores to the null-variables is nonzero. For illustration, we list in Table **4.6** the population quantities for Example 1 with $\sigma = 3$. The squared multiple correlation coefficients is $\Omega^2 = 0.70$ and the ratio of signal variance to noise variance equals $\Omega^2/(1 - \Omega^2) = 2.36$. Standardized regression coefficients $\boldsymbol{\beta}_{\text{std}}$, as well as partial correlations $\tilde{\boldsymbol{P}}_{XY}$ are zero whenever the corresponding regression coefficient $\boldsymbol{\beta}$ vanishes. In contrast, marginal correlations $\boldsymbol{P}_{XY}$, CAR scores $\boldsymbol{\omega}$ and the variable importance $\phi^{\text{CAR}}(x_i)$ are all nonzero even for $\beta_i = 0$. This implies that for large sample size in the setting of Example 1 all variables (but in particular, also X_3, X_4, and X_6) carry information about the response, albeit only weakly and indirectly for variables with $\beta_i = 0$.

As described in Section **4.3**, in the literature on variable importance the axiom of "proper exclusion" is frequently encountered, i.e. it is demanded that the share of Ω^2 allocated to a variable X_i with $\beta_i = 0$ is zero. The

Table 4.4: Average relative model error (x 1000) and its standard deviation as well as the mean true and false positives (TP+FP) in alternating rows for Examples 1 and 2. These simulations represent large sample settings ($d = 8$ with $n = 50$ to $n = 100$).

	CAR *	E NET	LASSO	BOOST	OLS	PCOR	GENIZI
Example 1 (true model size = 3)							
$n = 50$							
$\sigma = 1$	**107 (5)**	135 (7)	132 (6)	390 (24)	217 (8)	**107 (5)**	109 (6)
	3.0+1.2	3.0+1.9	3.0+1.8	3.0+2.6	3.0+5.0	3.0+0.7	3.0+1.3
$\sigma = 3$	**119 (7)**	130 (6)	148 (6)	151 (6)	230 (9)	153 (8)	129 (7)
	3.0+1.3	3.0+2.6	3.0+1.9	3.0+3.5	3.0+5.0	2.9+0.9	3.0+1.3
$\sigma = 6$	143 (6)	**127 (5)**	152 (6)	149 (8)	227 (8)	163 (6)	139 (6)
	2.5+1.2	2.8+2.4	2.6+2.0	2.8+3.7	3.0+5.0	2.3+1.4	2.5+1.1
$n = 100$							
$\sigma = 1$	**53 (3)**	64 (3)	59 (3)	219 (18)	97 (4)	54 (3)	55 (3)
	3.0+1.0	3.0+1.9	3.0+1.5	3.0+2.4	3.0+5.0	3.0+0.8	3.0+1.2
$\sigma = 3$	**55 (3)**	58 (2)	59 (3)	78 (3)	99 (3)	59 (3)	56 (4)
	3.0+1.2	3.0+2.1	3.0+1.9	3.0+3.6	3.0+5.0	3.0+0.8	3.0+1.0
$\sigma = 6$	65 (3)	**64 (3)**	69 (3)	66 (3)	97 (3)	76 (3)	65 (3)
	2.8+1.2	2.9+2.4	2.9+2.1	3.0+3.7	3.0+5.0	2.6+1.3	2.8+1.5
Example 2 (true model size = 3)							
$n = 50$							
$\sigma = 1$	**110 (5)**	147 (7)	134 (6)	716 (55)	230 (9)	120 (8)	130 (6)
	3.0+1.4	3.0+2.4	3.0+2.0	3.0+3.1	3.0+5.0	3.0+0.9	3.0+2.3
$\sigma = 3$	127 (5)	**124 (5)**	139 (6)	165 (7)	220 (8)	178 (9)	158 (8)
	2.8+1.6	3.0+3.0	2.8+2.2	2.8+3.5	3.0+5.0	2.4+1.6	2.8+2.1
$\sigma = 6$	121 (5)	**95 (4)**	121 (6)	110 (5)	232 (9)	165 (7)	135 (5)
	2.2+1.5	2.7+3.2	2.2+1.9	2.5+3.4	3.0+5.0	1.8+1.5	2.2+1.6
$n = 100$							
$\sigma = 1$	**49 (3)**	67 (3)	61 (3)	325 (28)	95 (3)	52 (3)	60 (3)
	3.0+1.1	3.0+2.2	3.0+1.9	3.0+3.0	3.0+5.0	3.0+1.0	3.0+2.0
$\sigma = 3$	**62 (3)**	63 (3)	64 (3)	83 (4)	101 (4)	78 (4)	62 (4)
	3.0+1.5	3.0+2.7	3.0+2.2	3.0+3.3	3.0+5.0	2.8+1.2	3.0+1.9
$\sigma = 6$	64 (3)	**53 (2)**	59 (2)	54 (2)	100 (4)	77 (3)	66 (3)
	2.6+1.7	2.9+3.1	2.6+2.1	2.7+3.3	3.0+5.0	2.0+1.4	2.7+1.8

* using empirical CAR estimator.

Table 4.5: Average relative model error (x 1000) and its standard deviation as well as the mean true and false positives (TP+FP) in alternating rows for Examples 3 and 4. These simulations represent small sample settings ($d = 40$ with $n = 10$ to $n = 100$).

	CAR *	E NET	LASSO	BOOST	OLS	PCOR	GENIZI
Example 3 (true model size = 10)							
$n = 10$							
$\sigma = 3$	**1482 (44)**	1501 (45)	1905 (75)	2203 (66)	—		
	6.1+7.0	6.3+11.5	2.1+4.7	2.4+13.7	—		
$n = 20$							
$\sigma = 3$	**838 (30)**	950 (26)	1041 (29)	1421 (44)	—		
	6.4+2.7	5.6+6.2	2.5+4.2	2.8+12.0	—		
$n = 50$							
$\sigma = 3$	**358 (11)**	571 (10)	608 (8)	805 (12)	5032 (214)	888 (27)	364 (12)
	8.5+0.6	5.2+2.9	3.3+3.3	4.2+13.0	10.0+30.0	2.5+2.2	8.4+1.1
$n = 100$							
$\sigma = 3$	172 (6)	488 (4)	525 (6)	569 (8)	693 (14)	406 (10)	**155 (5)**
	9.5+0.7	6.0+6.8	5.9+10.8	7.1+17.3	10.0+30.0	6.9+3.1	9.6+0.6
Example 4 (true model size = 6)							
$n = 10$							
$\sigma = 6$	**835 (24)**	1061 (34)	1684 (60)	1113 (39)	—		
	3.5+9.3	4.5+20.2	1.6+6.4	1.5+9.8	—		
$n = 20$							
$\sigma = 6$	**527 (18)**	767 (25)	925 (40)	791 (22)	—		
	4.2+7.0	4.4+13.2	2.4+7.5	2.0+9.4	—		
$n = 50$							
$\sigma = 6$	**200 (11)**	226 (9)	293 (14)	359 (11)	4991 (176)	1075 (67)	204 (7)
	4.9+3.0	4.3+4.7	3.0+4.0	3.3+12.9	6.0+36.0	2.8+5.0	5.5+0.8
$n = 100$							
$\sigma = 6$	**87 (4)**	107 (4)	112 (3)	168 (4)	699 (16)	232 (8)	94 (4)
	5.4+1.2	4.5+2.9	3.5+2.8	3.8+12.2	6.0+36.0	4.6+1.7	5.8+0.9

* using shrinkage CAR estimator.

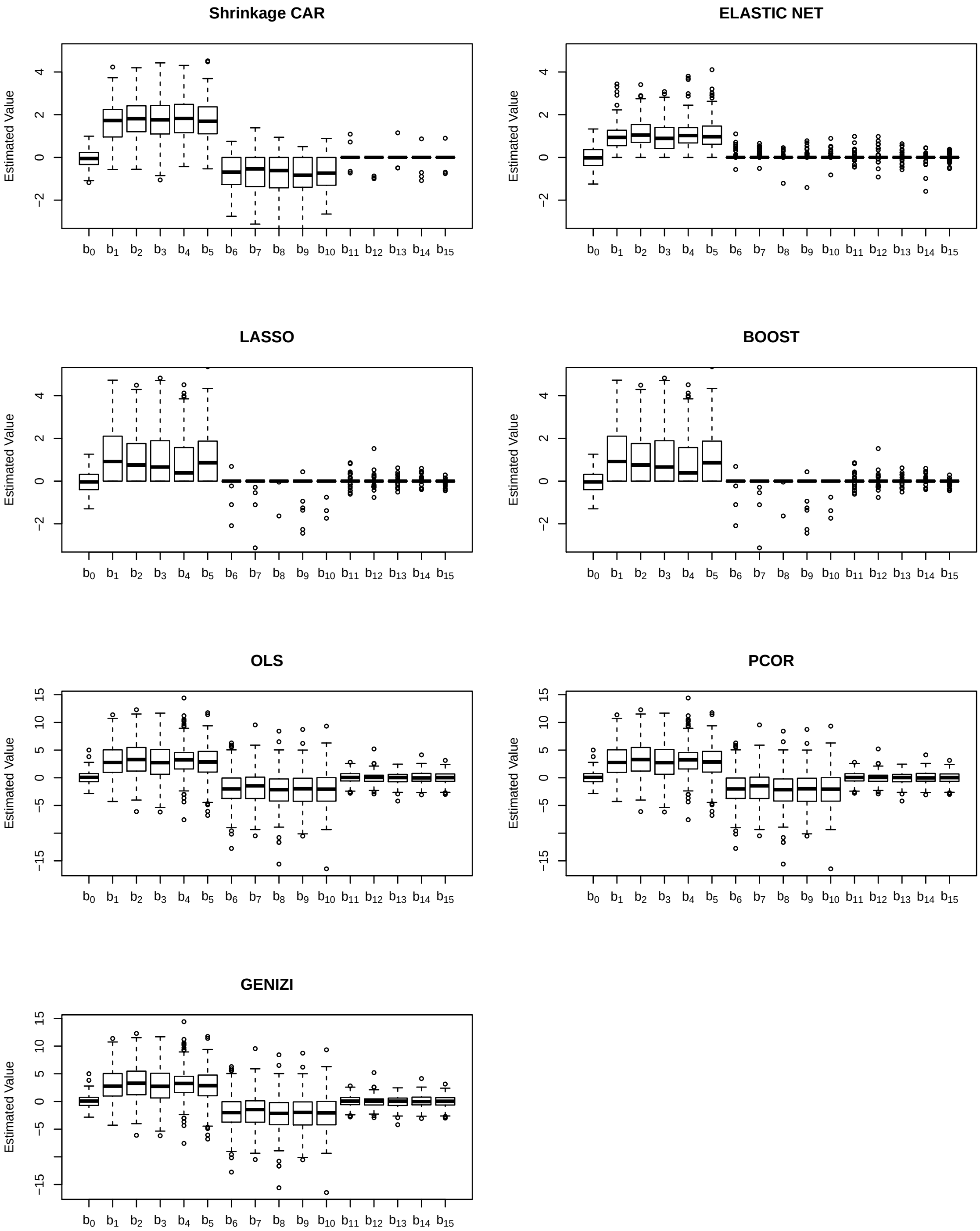


Figure 4.2: Distribution of estimated regression coefficients for Example 3 with $n = 50$ and $\sigma = 3$. Coefficients for variables X_{16} to X_{40} are not shown but are similar to those of X_{11} to X_{15}. The scale of the plots for OLS, PCOR and GENIZI is different from that of the other four methods.

Table 4.6: Population quantities for Example 1 with $\sigma = 3$.

Quantity	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
$\boldsymbol{\beta}$	3	1.5	0	0	2	0	0	0
$\boldsymbol{\beta}_{\text{std}}$	0.55	0.27	0	0	0.36	0	0	0
$\tilde{\boldsymbol{P}}_{XY}$	0.65	0.36	0	0	0.46	0	0	0
$\boldsymbol{P}_{XY}$	0.70	0.59	0.36	0.32	0.43	0.22	0.11	0.05
$\boldsymbol{\omega}$	0.60	0.40	0.15	0.13	0.36	0.10	0.04	0.02
ϕ^{CAR}	0.36	0.16	0.02	0.02	0.13	0.01	0.00	0.00

Numbers are rounded to two digits after the point.

squared CAR scores violate this principle if null and non-null variables are correlated. However, in our view this violation makes perfect sense, as in this case the null variables are informative about Y and thus may be useful for prediction. Moreover, because of the existence of equivalence classes in graphical models one can construct an alternative regression model with the same fit to the data that shows no correlation between null and non-null variables but which then necessarily includes additional variables. A related argument against proper exclusion is found in Grömping (2007).

4.6 Comparison of CAT and CAR score

Decorrelation offers an intuitive recipe for selecting variables under correlation. If there is no correlation among predictor variables there is consensus that the t-score in case of binary traits (Fan and Fan, 2008), respectively the marginal correlation in case of quantitative traits (Fan and Lv, 2008) are the optimal criteria to select variables. In the previous sections we have presented generalizations of the t-score and correlation to accommodate correlation among predictors, the CAT and the CAR score. To repeat, the CAT score in classification is the analog to the CAR score in regression. Hence, both scores share important characteristics and exhibit related behavior, as summarized in Table **4.7**.

In particular, CAT and CAR score are defined as the Mahalanobis-decorrelated marginal quantities optimal for variable selection in case of no correlation, either the t-score $\boldsymbol{\tau}$ or the marginal correlations $\boldsymbol{P}_{XY}$. Moreover, while the CAT score decomposes Hotelling's T^2, the CAR score decomposes the squared multiple correlation coefficient or proportion of variance explained. This suggests that Hotelling's T^2 in classification is the corresponding quantity to the squared multiple correlation coefficient. Already Hotelling (1931) mentioned the "affinity" of Hotellings T^2 with the multiple correlation coefficient due to similar geometrical interpretations. Here, the connection of CAT and CAR scores provides more evidence that Hotellings T^2 and the multiple correlation coefficient are related quantities with respect to different scales of the outcome.

Table 4.7: Comparison of CAT and CAR scores.

	CAT	CAR
Response Y	Binary	Metric
Definition	$\boldsymbol{\tau}^{\text{adj}} = \boldsymbol{P}^{-1/2}\boldsymbol{\tau}$	$\boldsymbol{\omega} = \boldsymbol{P}^{-1/2}\boldsymbol{P}_{XY}$
Marginal quantity	$\boldsymbol{\tau} = (\frac{1}{n_1} + \frac{1}{n_2})^{-1/2}\boldsymbol{V}^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$	$\boldsymbol{P}_{XY}$
Decomposition	Hotelling's T^2 $T^2 = \sum_{i=1}^{d}(\tau_i^{\text{adj}})^2$	Squared multiple correlation $\Omega^2 = \sum_{j=1}^{d}\omega_j^2$
Global test statistic for a set of size s	$T_s^2 = \sum_{i=1}^{s}(t_i^{\text{adj}})^2$	$R_s^2 = \sum_{j=1}^{s}\hat{\omega}_j^2$
Null distribution for empirical statistic under normality	$\frac{n-s-1}{(n-2)s}T_s^2 \sim F(s, n-s-1)$ with $n = n_1 + n_2$	$R_s^2 \sim \text{Beta}(\frac{s}{2}, \frac{n-s-1}{2})$

4.7 Summary

Correlation-adjusted marginal correlations $\boldsymbol{\omega}$, or CAR scores, are our contribution to the discussion on quantifying variable importance. This approach is based on simultaneous orthogonalization of the covariables by Mahalanobis-decorrelation and subsequently estimating the remaining correlation between the response and the sphered predictors. The CAR score meets most important properties postulated for measures of variable importance, especially it decomposes the proportion of variance explained. Furthermore, in contrast to other quantities, it is applicable to high-dimensional data.

Beyond the notion of variable importance we argue that the CAR score is the central quantity to understand nominal prediction error and the variance decomposition. In particular, the CAR score offers an elegant reformulation of the decomposition of variances

$$\begin{aligned}
\overbrace{\text{Var}(Y)}^{\text{Total variance}} &= \overbrace{\text{Var}(Y^\star)}^{\text{Explained variance}} + \overbrace{\text{Var}(Y - Y^\star)}^{\text{Unexplained variance}} \\
\sigma_Y^2 &= \sigma_Y^2(\boldsymbol{\omega}^T\boldsymbol{\omega}) + \sigma_Y^2(1 - \boldsymbol{\omega}^T\boldsymbol{\omega})\,.
\end{aligned}$$

Thus, we argue that the CAR score is the central quantity to assess which variables contribute to the explained variance or equivalently reduce the unexplained variance.

In an extensive simulation study, we demonstrate that the CAR score exhibits superior performance not only in ranking but also in prediction. It outperforms competing approaches in terms of true positives in ranking and in terms of model error in prediction. Interestingly, elastic net, lasso, and boosting fail to recover negative regression coefficients in contrast to the linear model, partial correlation, and the CAR score.

Chapter 5

Computational issues

The estimation and handling of correlation matrices in high-dimensional data is a complicated task. Here, we first discuss the Mahalanobis transform that performs decorrelation in an unique way. Furthermore, we present an efficient algorithm to compute the matrix power of high-dimensional matrices and thus allows the computation of CAT and CAR scores even in high-dimensions. Next, we illustrate a subtle trick using simple analysis that allows an enormous reduction in storage and computation time. Additionally, we provide information on how to determine the model size for CAT and CAR scores.

5.1 Special properties of the Mahalanobis transform

The computation of CAT and CAR scores relies on decorrelation by the Mahalanobis transform which derives from the Mahalanobis distance. The Mahalanobis distance is a metric alternative to the Euclidean that is used for non-spherical distributions. It quantifies the distance of a point $\boldsymbol{x}$ to the expectation vector $\boldsymbol{\mu}$ of a multivariate distribution as

$$D^2(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$$

where $\boldsymbol{\Sigma}^{-1}$ is the covariance of the distribution. Then, the Mahalanobis transform $\delta(\boldsymbol{x})$ by the correlation matrix $\boldsymbol{P}$ is defined as

$$\begin{aligned} \delta(\boldsymbol{x}) &= \boldsymbol{P}^{-1/2}\boldsymbol{x} \\ &= \boldsymbol{U}\boldsymbol{M}^{-1/2}\boldsymbol{U}^T\boldsymbol{x} \end{aligned} \tag{5.1}$$

where $\boldsymbol{M}$ is a diagonal matrix containing the eigenvalues and $\boldsymbol{U}$ is the orthogonal eigenvector system, derived from an eigendecomposition of the correlation matrix $\boldsymbol{P}$.

Importantly, the Mahalanobis transform has a number of properties not shared by other decorrelation transforms with $\text{Var}(\boldsymbol{\delta}(\boldsymbol{x})) = \text{diag}(\sigma^2)$. First, it is the unique linear transformation that minimizes $\text{E}\left((\boldsymbol{\delta}(\boldsymbol{x}) - \boldsymbol{x})^T(\boldsymbol{\delta}(\boldsymbol{x}) - \boldsymbol{x})\right)$, see Genizi (1993) and Hyvärinen et al. (2001, Section 6.5). Therefore, the Mahalanobis-decorrelated data $\boldsymbol{\delta}(\boldsymbol{x})$ are nearest to the original data $\boldsymbol{x}$. Second, as $\boldsymbol{P}^{-1/2}$ is positive definite $\boldsymbol{\delta}(\boldsymbol{x})^T\boldsymbol{x} > 0$ for any $\boldsymbol{x}$ which implies that $\boldsymbol{\delta}(\boldsymbol{x})^T$ and $\boldsymbol{x}$ are informative about each other also on a componentwise level (for example they must have the same sign). The correlation of the corresponding elements in $\boldsymbol{x}$ and $\boldsymbol{\delta}(\boldsymbol{x})$ is given by $\text{Cor}((\boldsymbol{x})_i, \boldsymbol{\delta}(\boldsymbol{x})_i) = (\boldsymbol{P}^{1/2})_{ii}$.

5.2 Computationally efficient calculation of shrinkage estimators of CAT and CAR scores

As mentioned in the proceeding chapters CAT and CAR score are population quantities and any suitable estimate for the correlation matrix $\boldsymbol{P}$ and the t-score, respectively correlation $\boldsymbol{P}_{XY}$, can be used. If there are more observations than variables we advise to use empirical estimates. In contrast, in large d, small n settings, as typically given in the analysis of omics data, regularization is needed. Here, we focus on shrinkage estimates. Still, it is possible, that the dimension d becomes prohibitively large to apply CAT and CAR scores in the standard fashion since the Mahalanobis transform of the $d \times d$ correlation matrix $\boldsymbol{R}$ is too computer-intensive. Using computational economies akin to those discussed in Hastie and Tibshirani (2004) we now show that computation of $-\frac{1}{2}$ matrix power of the estimated correlation matrix and subsequent calculation of estimates of CAT and CAR scores can be done in a computationally highly effective way, even when direct computation of CAT and CAR scores via Equation **3.18** and Equation **4.20** is infeasible.

As established in Equation **3.19** we concentrate on the shrinkage correlation estimator of Schäfer and Strimmer (2005), given by

$$\boldsymbol{R}_{\text{shrink}} = \lambda \boldsymbol{I}_d + (1 - \lambda)\boldsymbol{R}_{\text{empirical}}$$

where $\boldsymbol{R}_{\text{empirical}}$ is the empirical correlation matrix and λ the shrinkage intensity. As target we use the d-dimensional identity matrix $\boldsymbol{I}_d$. Additional information about the structure of the correlation matrix $\boldsymbol{P}$ (or its inverse) may also be taken into account by setting a different target. If the correlation matrix exhibits a known pattern, e.g., a block-diagonal structure, then it is advantageous to employ a correspondingly structured estimator (e.g., Tai and Pan, 2007; Li and Li, 2008; Guillemot et al., 2008).

Using singular value decomposition as in Equation **5.1** the empirical correlation matrix can be written as

$$\boldsymbol{R}_{\text{empirical}} = \lambda/(1-\lambda)\boldsymbol{U}\boldsymbol{M}\boldsymbol{U}^T$$

where $\boldsymbol{M}$ is positive definite matrix of size $m \times m$, $\boldsymbol{U}$ an orthonormal matrix of size $d \times m$, and $m = \text{rank}(\boldsymbol{R}_{\text{empirical}}) << d$. This simplifies the shrinkage estimator to

$$\boldsymbol{R}_{\text{shrink}} = \lambda(\boldsymbol{I}_d + \boldsymbol{U}\boldsymbol{M}\boldsymbol{U}^T)\,.$$

Then, we define

$$\boldsymbol{Z} = \boldsymbol{R}_{\text{shrink}}/\lambda = \boldsymbol{I}_d + \frac{1-\lambda}{\lambda}\boldsymbol{R}_{\text{empirical}} = \boldsymbol{I}_d + \boldsymbol{U}\boldsymbol{M}\boldsymbol{U}^T\,.$$

Subsequently, to calculate the α-th power of $\boldsymbol{Z}$ we use the identity[1]

$$\boldsymbol{Z}^\alpha = \boldsymbol{I}_d - \boldsymbol{U}(\boldsymbol{I}_m - (\boldsymbol{I}_m + \boldsymbol{M})^\alpha)\boldsymbol{U}^T \tag{5.2}$$

that requires only the computation of the α-th power of the $m \times m$ matrix $\boldsymbol{I}_m + \boldsymbol{M}$. This trick enables substantial computational savings when the number of samples (and hence the rank m of the correlation matrix) is much smaller than d.

We note that identity Equation **5.2** is related but not identical to the well-known Woodbury matrix identity for the inversion of a matrix. For $\alpha = -1$ our identity reduces to

$$\boldsymbol{Z}^{-1} = \boldsymbol{I}_d - \boldsymbol{U}(\boldsymbol{I}_m - (\boldsymbol{I}_m + \boldsymbol{M})^{-1})\boldsymbol{U}^T,$$

whereas the Woodbury matrix identity equals

$$\boldsymbol{Z}^{-1} = \boldsymbol{I}_d - \boldsymbol{U}(\boldsymbol{I}_m + \boldsymbol{M}^{-1})^{-1}\boldsymbol{U}^T.$$

We use Equation **5.2** to compute the α-th matrix power of $\boldsymbol{R}_{\text{shrink}}$ using

$$\boldsymbol{R}^\alpha_{\text{shrink}} = \lambda^\alpha(\boldsymbol{I}_d - \underbrace{\boldsymbol{U}}_{d\times m}(\boldsymbol{I}_m - \underbrace{(\boldsymbol{I}_m + \boldsymbol{M})}_{m\times m}{}^\alpha)\underbrace{\boldsymbol{U}^T}_{m\times d})\,.$$

This implies we only have to compute the matrix power of the $m \times m$ matrix $\boldsymbol{I}_m + \boldsymbol{M}$ to obtain $\boldsymbol{R}^\alpha$. Moreover, for efficiently calculating CAT and CAR scores it is crucial to note that it is not at all necessary neither to store or to compute the full $d \times d$ sized matrix $\boldsymbol{R}^{-1/2}_{\text{shrink}}$. For example, the shrinkage CAR

[1] The validity of the identity can be verified by noting that the eigenvalues of $(\boldsymbol{I}_d + \boldsymbol{U}\boldsymbol{M}\boldsymbol{U}^T)^\alpha$ and of the righthand side of Equation **5.2** are identical (which implies similarity between the two matrices) and that no further rotation is needed for identity.

score of Equation **4.20** is given as

$$\begin{aligned}\boldsymbol{R}_{XY}^{\text{adj}} &= \boldsymbol{R}_{\text{shrink}}^{-1/2}\boldsymbol{R}_{XY}^{\text{shrink}} \\ &= \lambda^{-1/2}(\boldsymbol{I}_d - \boldsymbol{U}(\boldsymbol{I}_m - (\boldsymbol{I}_m + \boldsymbol{M})^{-1/2})\boldsymbol{U}^T)\boldsymbol{R}_{XY}^{\text{shrink}} \\ &= \lambda^{-1/2}(\underbrace{\boldsymbol{R}_{XY}^{\text{shrink}}}_{d\times 1} - \underbrace{\boldsymbol{U}(\boldsymbol{I}_m - (\boldsymbol{I}_m + \boldsymbol{M})^{-1/2})}_{d\times m}(\underbrace{\boldsymbol{U}^T\boldsymbol{R}_{XY}^{\text{shrink}}}_{m\times 1}))\,.\end{aligned} \tag{5.3}$$

Consequently, Equation **5.3** allows to obtain shrinkage estimates of CAT and CAR scores effectively even in high-dimensions as none of the matrices employed in Equation **5.3** is larger than $d \times m$, and most are even smaller ($d \times 1$ or $m \times 1$), all without actually computing the shrinkage correlation matrix $\boldsymbol{R}_{\text{shrink}}$ in Equation **3.18** and Equation **4.20**.

5.3 On determining the model size

So far this thesis has focused on the two criteria CAT and CAR score to quantify the importance of variables. To perform variable selection it is essential to correctly assess variable importance, but also to determine the optimal model size. This is an intricate area of intense research and a comprehensive review is beyond the scope of this work. Still this chapter gives an introduction to the most important concepts.

Again there are different aspects if the variables are selected with respect to prediction or to ranking. Moreover, it is vital to select the approach to set a cut-off with respect to the data at hand. For example, p-values based on the null distribution are advised for data sets with only few variables; whereas FDR is applicable on high-dimensional data only. Here, we provide an introductory sketch of different techniques to select the model size in variable selection using CAT or CAR scores.

5.3.1 Distribution under the null hypotheses of CAT and CAR score

For small data sets variable selection by constructing tests based on the distribution under the null hypotheses is widespread. Here, we provide instructions which distributions to use for squared CAT and CAR scores as well as for the global test statistics of a set of variables. Variable selection by p-values is advised only for small data sets with few variables due to multiple testing. This means that the significance level of each test must be adjusted to meet the overall error rate. Corrections like Bonferroni's are too conservative for larger numbers of variables. The concept of testing according to Neyman and Pearson explicitly focuses on discerning the TP effects correctly, while controlling the FP effects at a certain significance level

(Rüger, 1998). Thus, classical variable selection based on p-values is the most adequate technique in ranking to include a prespecified number of FPs. In Section **6.1** we illustrate the use of the null distribution on the benchmark diabetes data including ten variables.

1. Under the null hypothesis $\mathrm{H}_0 : t_i^{\mathrm{adj}} = 0$ the CAT score follows a *t-distribution.*

   To begin with the rescaled Hotellings T^2 (Equation **3.17**) follows a F-distribution with $\mathrm{df}_1 = d$ and $\mathrm{df}_2 = (n-d-1)$ under the null hypothesis (Hotelling, 1931)

   $$\frac{(n-d-1)}{d(n-2)}T^2 \sim F(d, n-d-1)\,. \tag{5.4}$$

   For $d = 1$ the F- distribution is equivalent to the t-distribution with $n-2$ degrees of freedom

   $$\frac{(n-2)}{1(n-2)}t^2 \sim F(1, n-2) \Leftrightarrow t \sim t(n-2)\,.$$

   This suggests that under the null hypothesis the squared CAT score $(t_i^{\mathrm{adj}})^2$ for variable $i \in 1, ..., d$ follows the F-distribution with $(1, n-2)$ degrees of freedom, respectively the CAT score (t_i^{adj}) follows a t-distribution with $(n-2)$ degrees of freedom

   $$(t_i^{\mathrm{adj}})^2 \sim F(1, n-2) \Leftrightarrow t_i^{\mathrm{adj}} \sim t(n-2)\,.$$

   For about $n = 30$ observations the t-distribution can be approximated by a normal distribution (Fahrmeir et al., 2003). Furthermore, we derive from Equation **5.4** a null distribution for the global test statistic $T_s^2 = \sum_{i=1}^{s}(t_i^{\mathrm{adj}})^2$ for a set of variables of size s as

   $$\frac{(n-s-1)}{s(n-2)}T_s^2 \sim F(s, n-s-1)\,.$$

2. Under the null hypothesis $\mathrm{H}_0 : \hat{\omega}_i^2 = 0$ the squared CAR score follows a *Beta distribution*.

The coefficient of determination R^2 follows the Beta distribution with $\mathrm{Beta}(\frac{d}{2}, \frac{n-d-1}{2})$. It is defined as the proportion of explained to total variance

$$R^2 = \frac{\widehat{\mathrm{Var}}(Y^\star)}{\widehat{\mathrm{Var}}(Y)} = \frac{\widehat{\mathrm{Var}}(Y^\star)}{\widehat{\mathrm{Var}}(Y - Y^\star) + \widehat{\mathrm{Var}}(Y^\star)} . \tag{5.5}$$

To derive the distribution of the estimated variance of the best linear predictor $Y^\star$ we use the connection of the χ^2-distribution to the Beta distribution. Suppose that $X \sim \chi^2(\mathrm{df}_1)$ and $Y \sim \chi^2(\mathrm{df}_2)$, then

$$\frac{X}{X+Y} \sim \mathrm{Beta}(\mathrm{df}_1/2, \mathrm{df}_2/2) . \tag{5.6}$$

Due to the decomposition of $R^2 \sim \mathrm{Beta}(\frac{d}{2}, \frac{n-d-1}{2})$ in Equation **5.5** and the additivity property of the χ^2 distribution we arrive at

- $\widehat{\mathrm{Var}}(Y^\star) \sim \chi^2(d)$,
- $\widehat{\mathrm{Var}}(Y - Y^\star) \sim \chi^2(n - d - 1)$.

Following the connection of CAR scores to the best linear predictor in Equation **4.16** we find that $\widehat{\mathrm{Var}}(Y^\star) = \sigma_Y^2 \hat{\boldsymbol{\omega}}^T \hat{\boldsymbol{\omega}} \sim \chi^2(d)$.

Now we define the observed best linear predictor based on variable $i \in 1, ..., d$ as

$$(Y_i^\star - \hat{\mu}_i)/\hat{\sigma}_Y^2 := \hat{\omega}_i \delta(x_i)$$

with $\widehat{\mathrm{Var}}(Y_i^\star) = \hat{\sigma}_Y^2 \widehat{\mathrm{Var}}(\hat{\omega}_i \delta(x_i)) = \hat{\sigma}_Y^2 \hat{\omega}_i^T \hat{\omega}_i$.

This suggests that the variance of $Y^\star$ is equal to the sum over all d variances of $Y_i^\star$ and thus the sum $\sum_{i=1}^d \widehat{\mathrm{Var}}(Y_i^\star)$ also follows a χ^2-distribution with d degrees of freedom

$$\widehat{\mathrm{Var}}(Y^\star) = \hat{\sigma}_Y^2 \hat{\omega}^T \hat{\omega} = \hat{\sigma}_Y^2 \sum_{i=1}^d \hat{\omega}_i^2 = \sum_{i=1}^d \widehat{\mathrm{Var}}(Y_i^\star) \sim \chi^2(d).$$

Since the χ^2-distribution is additive and under the null hypothesis all d variances $\widehat{\mathrm{Var}}(Y_i^\star)$ have the same weight, the degrees of freedom are split equally. Hence, the variance of one particular best linear predictor $\widehat{\mathrm{Var}}(Y_i^\star)$ follows the χ^2-distribution under the null hypotheses with one degree of freedom. With the distribution of $Y_i^\star$ in mind, we define analog to Equation **5.5**

$$R_i^2 := \frac{\widehat{\text{Var}}(Y_i^\star)}{\widehat{\text{Var}}(Y)} = \frac{\widehat{\text{Var}}(Y_i^\star)}{\widehat{\text{Var}}(Y - Y_i^\star) + \widehat{\text{Var}}(Y_i^\star)} = \frac{\hat{\sigma}_Y^2 \hat{\omega}_i^2}{\hat{\sigma}_Y^2} = \hat{\omega}_i^2 \,,$$

where

- $\widehat{\text{Var}}(Y_i^\star) \sim \chi^2(1)$,
- $\widehat{\text{Var}}(Y - Y_i^\star) \sim \chi^2(n-2)$.

Using the relation of Beta to χ^2-distribution in Equation **5.6** it is evident that R_i^2 and thus the squared CAR score follows a Beta distribution with 1 and $(n-2)$ degrees of freedom

$$\hat{\omega}_i^T \hat{\omega}_i \sim \text{Beta}(1/2, (n-2)/2) \,.$$

Thus, the null distribution of the empirical CAR scores under the null hypothesis is identical to that of the empirical marginal correlations, regardless of the correlation among $\boldsymbol{X}$ (Hotelling, 1953).

5.3.2 Penalized residual sum of squares

For CAR scores there is another option to select a cut-off due to an intrinsic link with information criteria based on penalized RSS, as generally presented in Equation **4.10**. Section **4.4.5** provides the theoretical aspects, especially Equation **4.19** suggests simple rules to set a cut-off. First, the empirical squared CAR scores are ordered, so that $\hat{\omega}_{(1)}^2 > ... > \hat{\omega}_{(d)}^2$. Depending on the penalty this leads to the following concrete rules for determining the model size:

- AIC, respectively Cp include all $\hat{\omega}_{(q)}^2$ that

$$\hat{\omega}_{(q)}^2 > 2 \cdot (1 - R^2)/n \,.$$

- RIC includes all $\hat{\omega}_{(q)}^2$ that

$$\hat{\omega}_{(q)}^2 > 2 \log(d) \cdot (1 - R^2)/n \,.$$

- BIC includes all $\hat{\omega}_{(q)}^2$ that

$$\hat{\omega}_{(q)}^2 > \log(n) \cdot (1 - R^2)/n \,.$$

The penalized RSS criteria above are estimates for the in-sample prediction error (Hastie et al., 2009) that is defined as

$$\widehat{\mathrm{Err}}_{in} = \overline{\mathrm{err}} + \hat{\mathrm{err}}_{\mathrm{over}}$$

where $\overline{\mathrm{err}}$ is the average training error, as presented in Equation **2.1**, and $\hat{\mathrm{err}}_{\mathrm{over}}$ is an estimate for the average over-optimism due to the estimation of the prediction error using the training data. The in-sample prediction error is of no direct use for future predictions, but still can help to compare different models (Hastie et al., 2009). To sum up, penalized RSS are formulated with respect to optimizing prediction and thus select a model size with the best predictive performance.

In Section **6.1** we illustrate the use of penalized RSS cut-offs on the benchmark diabetes data including ten variables and compare the model sizes with the ones obtained from the empirical null distribution.

5.3.3 Estimation of the prediction error by cross-validation

A different approach to estimating the prediction error are computer-intensive techniques as cross-validation and bootstrapping. In contrast to the penalized RSS considered in Section **5.3.2**, they estimate the expected prediction or generalization error (Equation **2.2**) which is the quantity to evaluate the future predictive performance most accurately. In the following we concentrate on the cross-validation approach to estimate the prediction error. If there are enough observations a validation set can be used to assess the predictive performance. But most often data is scarce and a resampling scheme called K-fold cross-validation is applied. K-fold cross-validation divides all n observations available in K ideally equal parts. Then, for the kth part of the observations, with $k \in 1, ..., K$, the prediction rule is fitted for all observations but the ones in the kth part. Subsequently, the prediction error for the kth part is the difference between the actual observations in the kth part and the prediction based on all observations but the ones in the kth part. This is repeated for all $k \in 1, ..., K$ parts and then the average error over all K-folds is used as an estimate for the expected prediction error. Model selection based on cross-validation selects the model with the lowest estimated prediction error.

Section **6.3.5** discusses the regression analysis of a data set on expression of $d = 403$ genes. Here, the number of variables is too small to fit a reliable FDR estimate, still it is too large to apply p-value testing. Hence, we focus on the prediction error estimated by cross-validation to select the model-size for the CAR score.

5.3.4 False (non) discovery rate

The false discovery rate (FDR) is the most powerful and intuitive approach to variable selection in high-dimensional set-ups. It allows to control the expected proportion of incorrectly rejected null hypotheses in multiple testing. First proposed by Schweder and Spjøtvoll (1982) and Benjamini and Hochberg (1995), FDR is now ubiquitous in the analysis of high-dimensional data. Precisely, the FDR is defined as "the proportion of the rejected null hypothesis which are erroneously declared significant" (Benjamini and Hochberg, 1995). There are different approaches of deriving the FDR, namely the local false discovery rate and the tail area-based false discovery. For an overview see Strimmer (2008b) or Efron (2008).

Notably, in FDR analysis the variables are first ranked according to their p-value or the summary statistic z employed in the specific set-up, that is a one-to-one transformation of the corresponding p-value. This suggests that FDR is tightly linked to the concept of variable ranking. Then, the distribution of the ranking statistic is fitted using a mixture of a null density f_0 and an alternative density f_A. For instance the summary statistic z is modeled by

$$f(z) = \eta_0 f_0(z) + (1 - \eta_0) f_A(z)$$

Importantly, it is assumed that the mixing constant $\eta_0 \in [0, 1]$ is near 1, so that there is only a small proportion of TP effects. Moreover, it is essential that the number of hypothesis d is large, so that the fit of the distributions is stable.

The tail area-based false discovery can be seen as corrected p-value; for the specific summary statistic z it is interpreted as

$$\text{Fdr}(z_i) := \text{Prob}(\text{“not interesting”} \mid Z \geq z_i)\,.$$

In contrast the local false discovery rate defines an empirical Bayesian posterior probability (Efron et al., 2001)

$$\text{fdr}(z_i) := \text{Prob}(\text{“not interesting”} \mid Z = z_i)\,.$$

While the FDR is especially constructed to include a specified number of FP in a ranking, it is less useful in determining a set of variables in prediction. For prediction it is often beneficial to include more variables since the aim is *not* to discover the TP effects, but to build a classifier that predicts the outcome as precise as possible. Thus, Ahdesmäki and Strimmer (2010) advocate the control of the false nondiscovery rate (FNDR) that is the expected proportion of alternative hypotheses incorrectly specified as null variable. Using the local FDR, there is a exact relation between FNDR and FDR

$$\text{fndr}(z_i) := \text{Prob}(\text{“interesting”} \mid Z = z_i) = 1 - \text{fdr}(z_i)\,.$$

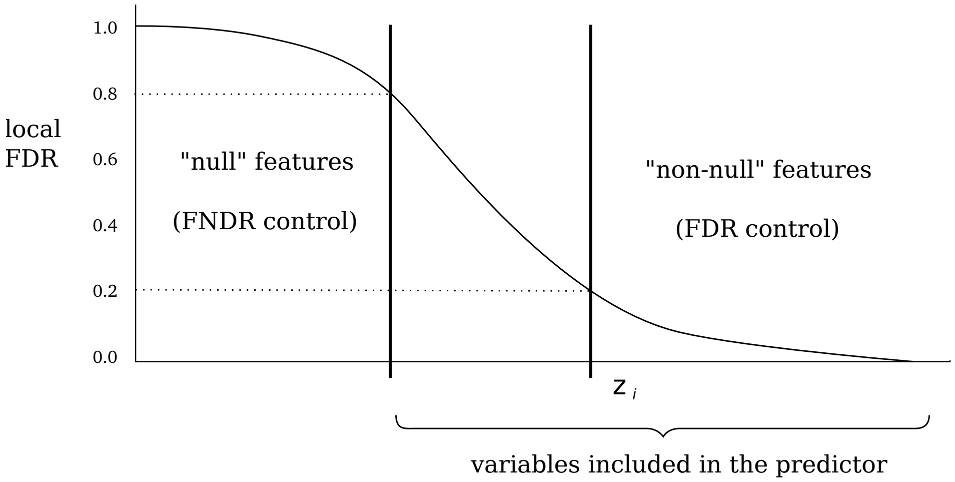


Figure 5.1: Local false discovery rate as function of the summary score z_i, e.g. squared CAT or CAR score. Depending on z_i the variables fall into three distinct areas: The rejection zone on the left, an acceptance zone on the right, and a "buffer" zone in between. Variables in both, the buffer and acceptance zone, are included by the FNDR approach. The figure is adapted from Ahdesmäki and Strimmer (2010).

In particular, the FNDR cut-off aims at correctly discerning variables that are null variables, but to relax the cut-off and to include more variables that are likely to be no true effects. Consequently, the use of FNDR leads to the selection of a larger set of variables that is a superset of the variables selected by FDR. Ahdesmäki and Strimmer (2010) refer to the variables additionally included as "buffer zone". See Figure **5.1** for an illustration of the difference between variables included by FNDR or by FDR control.

In Section **6.3** we present four classification tasks on high-dimensional transcriptomics data where we set the cut-offs using FNDR and FDR.

Chapter 6

Application to experimental data

To illustrate application and performance of CAT and CAR scores in the analysis of high-dimensional omics data, this chapter comprises analysis of genomics data in Section **6.2**, transcriptomics in Section **6.3**, and metabolomics data in Section **6.4**. But first, we start with a case study on variable selection in the reference clinical data set on the progression of diabetes in Section **6.1**.

6.1 Clinical data: Analysis of the diabetes data

In this section we reanalyze a low-dimensional benchmark data set on the disease progression of diabetes discussed in Efron et al. (2004) and Hastie et al. (2009) as case study on variable selection. There are $d = 10$ covariates, age (`age`), sex (`sex`), body mass index (`bmi`), blood pressure (`bp`) and six blood serum measurements (`s1`, `s1`, `s2` `s3` , `s4`, `s5`, `s6`), on which data were collected from $n = 442$ patients. As $d < n$ we use empirical estimates of CAR scores and ordinary least squares regression coefficients in our analysis. The data were centered and standardized beforehand.

A particular challenge of the diabetes data set is that it contains two variables (`s1` and `s2`) that are highly positively correlated ($r = 0.897$) but behave in an antagonistic fashion. Specifically, their regression coefficients have the opposite signs so that in prediction the two variables cancel each other out. Figure **6.1** shows all regression models that arise when covariates are added to the model in the order of decreasing variable importance given by $\phi^{\text{CAR}}(x_i)$. As can be seen from this plot, the variables `s1` and `s2` are ranked least important and included only in the two last steps.

For the empirical estimates the exact null distributions are available, therefore we also computed p-values for the estimated CAR scores, marginal correlations $\boldsymbol{P}_{XY}$ and partial correlations $\tilde{\boldsymbol{P}}_{XY}$, and selected those variables for inclusion with a p-value smaller than 0.05. In addition, we computed lasso, elastic net and boosting regression models. Note that the use of partial correlations for selecting variables is equivalent to the t-test on the

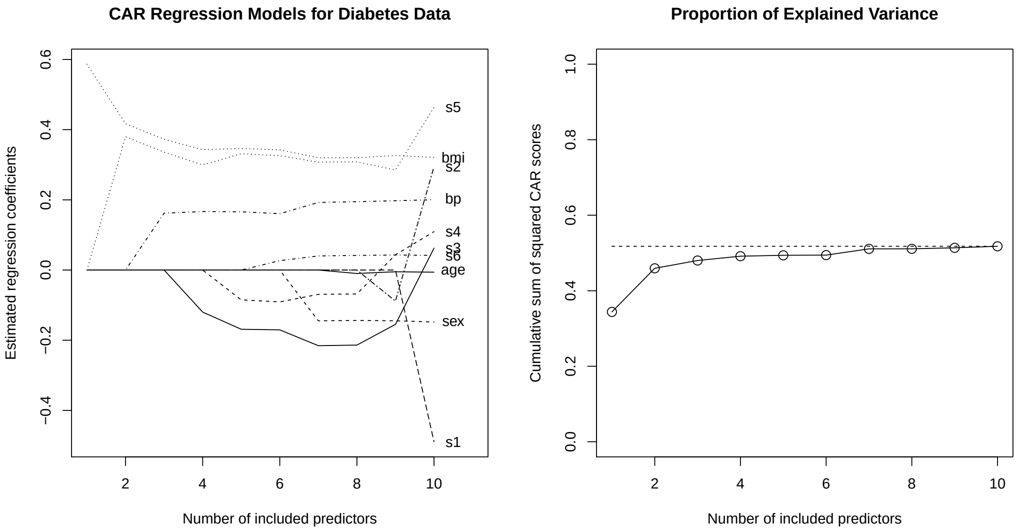


Figure 6.1: Estimates of regression coefficients for the diabetes study on the left. Variables are included in the order of empirical squared CAR scores, and the corresponding regression coefficients are estimated by ordinary least squares. The antagonistic correlated variables `s1` and `s2` are included only in the last two steps. Accumulated squared CAR score on the right as indicator of the proportion of variance explained by increasing model size. The largest gains are achieved by the first three variables.

regression-coefficients. See Equation **4.13** for the exact connection. The results are summarized in Table **6.1**. All models include `bmi`, `bp` and `s5` and thus agree that those three explanatory variables are most important for prediction of diabetes progression. Using marginal correlations and the elastic net both lead to large models of size 9 and 10, respectively, whereas the CAR feature selection results in a smaller model. The CAR model and the model determined by partial correlations are the only ones not including either of the variables `s1` or `s2`.

In addition, we also compared CAR models selected by the various penalized RSS approaches. Using the C_p / AIC rule on the empirical CAR scores results in 8 included variables, RIC leads to 7 variables, and BIC to the same 6 variables as in Table **6.1**. The accumulated squared CAR score is illustrated in Figure **6.1**. Here, it is evident that the largest shares of the proportion of variance explained are achieved by the first six variables. Then, the gain in proportion of variance explained diminishes to stagnates at the level of the overall coefficient of determination.

Table 6.1: Ranking of variables and selected models (in bold type) using various variable selection approaches on the diabetes data.

Rank	$\tilde{P}_{XY}$	P_{XY}	CAR	Elastic Net	Lasso	Boost
age	10	**8**	8	**10**	—	—
sex	**4**	10	7	**4**	**5**	**5**
bmi	**1**	**1**	**1**	**1**	**1**	**1**
bp	**2**	**3**	**3**	**3**	**3**	**3**
s1	5	**7**	9	**9**	**6**	**6**
s2	6	**9**	10	**7**	—	—
s3	9	**5**	**4**	**5**	**4**	**4**
s4	7	**4**	**5**	**6**	—	—
s5	**3**	**2**	**2**	**2**	**2**	**2**
s6	8	**6**	**6**	**8**	**7**	**7**
Model size	4	9	6	10	7	7

6.2 Genomics data: Analysis of SNP data

GWAS are now routinely conducted to search for genetic factors indicative of or even causally linked to disease. Typically, the aim of such a study is to identify a small subset of SNPs associated with a phenotype of interest. From an analysis point of view the screening for relevant genetic biomarkers is best cast as a problem of statistical variable selection. More precisely, SNPs are selected by ranking them accordingly to their association to the phenotype. In GWAS variable selection is very challenging as the full set of SNPs is often very large while both the effect of each potentially causal SNP as well as their number is very small (e.g. Guan and Stephens, 2011).

To date, most GWAS are based on single-SNP analyzes where each SNP is considered independently of all others and association with the phenotype is computed using a univariate test statistic such as variants of the t-score, quantities from contingency tables, as the ATT statistic (Armitage, 1955) or the χ^2 test, or marginal correlation (Foulkes, 2009). The advantage of this approach is that it is computationally inexpensive. However, it implicitly assumes complete independence of markers and thus ignores the dependency structure among SNPs, e.g., due to linkage or interaction among SNPs. In order to increase statistical efficiency and to exploit the correlation among predictive SNPs several authors have recently started to investigate simultaneous SNP selection using fully multivariate approaches. This was pioneered for GWAS in the seminal paper of Hoggart et al. (2008) that introduced the NEG regression model, a shrinkage-based approach to select relevant SNPs. More recently, LASSO regression was employed to

GWAS by Wu et al. (2009) and MCP regression in an overview article by Ayers and Cordell (2010). For more details of these penalized regression strategies see Section **4.2**. Furthermore Guan and Stephens (2011) developed Bayesian variable selection regression explicitly for application on GWAS. Boosting (Hothorn and Bühlmann, 2006) is another promising multivariate approach advocated for high-dimensional variable selection that has not yet been investigated for GWAS.

Here, we conduct a systematic comparison of these state-of-the-art simultaneous SNP selection procedures using data from the GAW17 consortium (Almasy et al., 2011). These data are particularly suited for investigating relative performance as the true causal SNPs are known. In a recent study Ayers and Cordell (2010) investigated methods for simultaneous SNP selection with focus on binary traits. We extend this comparison study by considering quantitative phenotypes and by including additional variable selection approaches, in particular the CAR score and boosting. Since the GAW17 data comprises a list of the true causal effects it is possible to compare the rankings of the approaches listed, especially we report the true positives included, model size, and the quality of the top 100 SNPs.

6.2.1 GAW 17 unrelated data and preprocessing

The Genetic Analysis Workshop (GAW) is an initiative that provides reference data for statistical genetic analysis aiming at a common ground for comparing and evaluating existing approaches and novel strategies. To test the performance of CAT and CAR scores we use the mini-exome data set compiled for the GAW17 workshop held 13-16 October 2010 in Boston (`http://www.gaworkshop.org/gaw17/`). This data set is a combination of real sequence data and simulated synthetic phenotypes, where the true causal SNPs are known. Compilation and simulation of the phenotypes is described in detail in Almasy et al. (2011). The data are available by request from Jean MacCluer, see `http://www.gaworkshop.org/gaw17/data.html` for details.

We focus here on the GAW 17 unrelated data with metric phenotypes Q1 and Q2. The corresponding sequence data matrix contains information on 24,487 SNPs for $n = 697$ individuals. For each phenotype there are $B = 200$ simulations. By construction, phenotype Q1 has a residual heritability of 0.44 and is influenced by 39 SNPs in 9 genes, whereas Q2 has a lower residual heritability of 0.29 and is influenced by 72 SNPs in 13 genes. This suggests that discovery of true causal SNPs should be less challenging for Q1 than for Q2.

In the preprocessing of the sequences we first recoded the alleles in the raw data into 0, 1, 2 assuming an additive effects model. Second, we standardized the data matrix to column mean zero and column variance 1. Subsequently, we removed duplicate predictors so that 15,076 unique

SNPs remained. The set of true causal SNPs for both Q1 and Q2 also contains each a duplicate, reducing the number of true unique SNPs to 38 and 71. Finally, we further filtered out synonymous SNPs, as we are interested only in non-synonymous mutations. The resulting predictor matrix $\boldsymbol{X}$ is of size $697 \times 8,020$, i.e. $d = 8,020$ unique non-synonymous SNPs are simultaneously considered for selection.

For preprocessing the response variables Q1 and Q2 we removed the influence of the three non-genetic covariates sex, age, and smoking by linear regression. The resulting residuals were standardized to mean zero and variance 1 which yielded $B = 200$ response vectors $y_1^{(b)}$ and $y_2^{(b)}$, where $b \in 1, ..., B$, each of size 697×1.

6.2.2 Relative performance of the rankings generated by the investigated methods

Table 6.2: Software used in the comparison study. The R packages are available from the R software archive CRAN at `http://cran.r-project.org/`.

Method	Software	Reference
CAR	R package `care`	see Appendix Section **B**
COR	R package `care`	see Appendix Section **B**
NEG	`HLasso` program	Hoggart et al. (2008)
MCP	R package `ncvreg`	Breheny and Huang (2011)
BOOST	R package `mboost`	Hothorn and Bühlmann (2006)
LASSO	R package `glmnet`	Friedman et al. (2010)

For each of the $B = 200$ response vectors for Q1 and Q2 we computed a regression model including all $d = 8,020$ SNPs as potential predictors. Following Ayers and Cordell (2010) we focused on regularized regression approaches. Specifically, we used the following five methods, all of which have been shown to be powerful tools for variable selection in large-scale regression settings:

- CAR: Variable ranking by shrinkage CAR score,
- NEG: Regression with normal exponential gamma (NEG) prior (Hoggart et al., 2008),
- MCP: Regression with MCP penalty (Zhang, 2010),
- BOOST: Boosting (Schapire, 1990), and
- LASSO: Lasso regression (Tibshirani, 1996).

The corresponding software implementations are listed in Table **6.2**. As a reference for comparison we additionally included two baseline methods:

- COR: Univariate SNP ranking by marginal correlation, and
- RND: Random ordering of all SNPs.

All methods except CAR and COR combine regularization with variable selection. Thus, for determining model sizes for CAR scores and COR we used a local FNDR thresholding with a cutoff of 0.2 as suggested in Ahdesmäki and Strimmer (2010) using the R package `fdrtool` (Strimmer, 2008a,b).

Generally, all software were run with default settings. The regularization parameters required by the NEG, MCP, BOOST and CAR approaches were set to fixed values optimizing the overall performance of each method. Specifically, for CAR and MCP we employed $\lambda = 0.1$, for BOOST $\nu = 0.1$ and for NEG $\lambda = 85$. For LASSO we used the built-in cross-validation routines.

Table 6.3: Median model sizes and average true positives for phenotypes Q1 and Q2 for all investigated methods summarized across the 200 repetitions (column 2 and 3). For comparison, we also show the average true positives at the specified model size for CAR, COR and RND (columns 4-6). The best performing method is shown in bold, the second best in italic.

	Method	Model	TP	TP	TP	TP
		Size	Method	CAR	COR	RND
Q1						
	CAR	218	**9.635**	**9.635**	*8.490*	1.110
	COR	319	*9.280*	**10.990**	*9.280*	1.625
	NEG	1390	*15.310*	**17.565**	14.375	6.595
	MCP	20	*4.110*	**4.190**	3.945	0.115
	BOOST	53	*5.835*	**5.905**	5.495	0.250
	LASSO	37	*5.185*	**5.205**	4.890	0.175
Q2						
	CAR	135	**6.885**	**6.885**	*6.195*	1.250
	COR	19	**2.225**	*2.165*	**2.225**	0.190
	NEG	1632	20.21	**28.08**	*25.90*	14.50
	MCP	29	2.745	**2.820**	*2.760*	0.275
	BOOST	59	*3.920*	**4.335**	3.820	0.585
	LASSO	15	1.500	*1.875*	**1.970**	0.135

The aim of this study is to compare simultaneous SNP selection methods with regard to their ability to discover the true known SNPs. For this purpose we investigated the respective SNP rankings and the corresponding

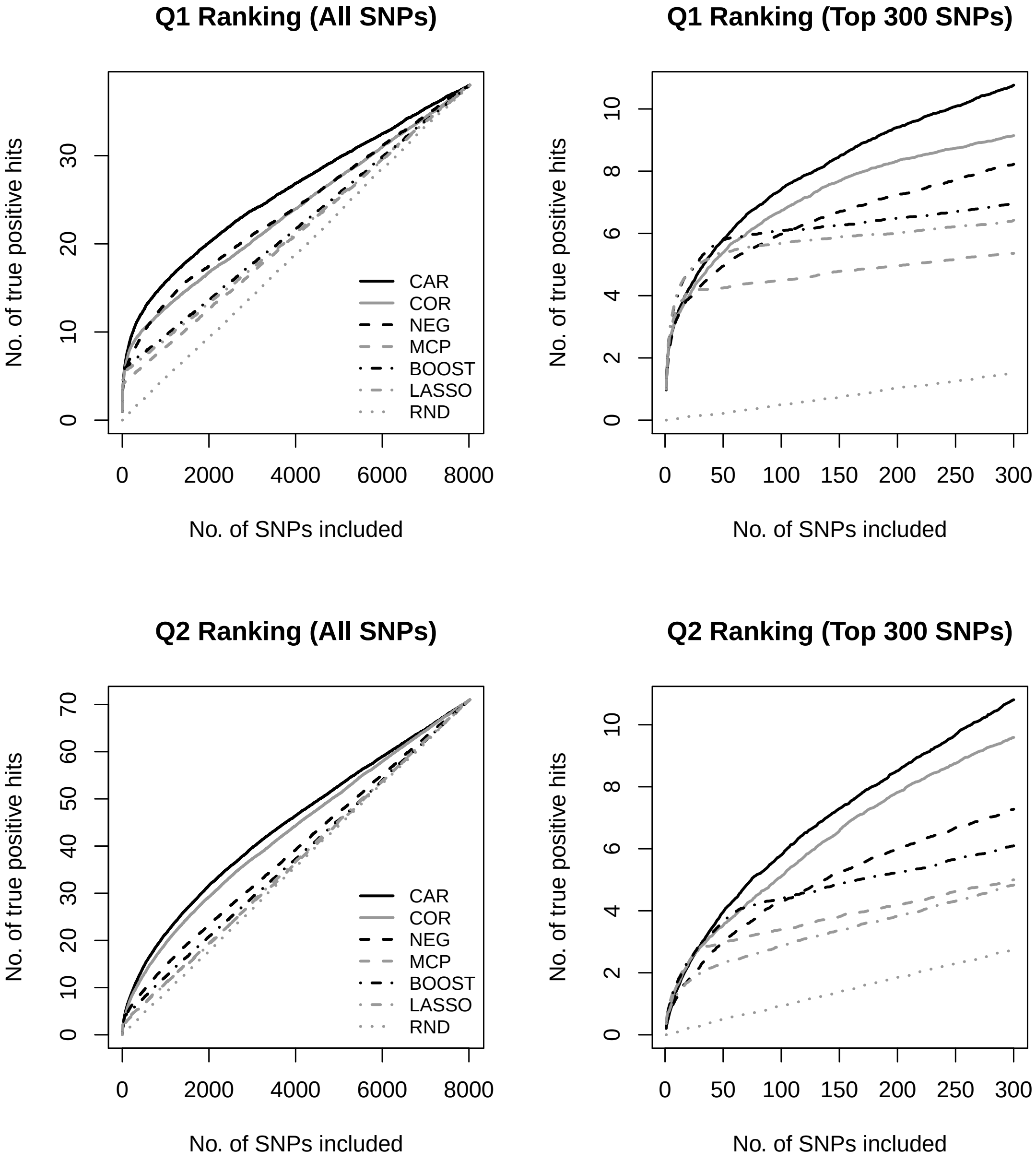


Figure 6.2: Average true positives resulting from SNP rankings of the investigated approaches for phenotype Q1 (top row) and Q2 (bottom row). For Q1 there are 38 true SNPs and for Q2 71 true SNPs.

true positives, the size of the selected models, and the variability across the 200 repetitions. In Figure **6.2** and the associated Table **6.3** we compare the effectiveness of SNP rankings for phenotypes Q1 and Q2. For Q1 all methods uniformly outperform marginal correlation, i.e. at the model size determined by each procedure the number of true positives is larger than that for marginal correlations at the same cutoff. Thus, for Q1 all multivariate SNP selection approaches improve over univariate selection. Moreover, as can be seen from Figure **6.2** (top row) and Table **6.3** SNP ranking by CAR scores regardless the chosen cutoff is better in terms of true positive than all other competing approaches. For the more challenging phenotype Q2 the situation is similar. CAR scores almost always provide the most effective ranking (see lower part of Table **6.3**) but intriguingly for this phenotype it is also the only multivariate method that improves consistently over marginal correlation. Boosting provides a competitive ranking up to the first 60 SNPs included. In Table **6.3** we also list the median model sizes for each regression approach. BOOST, LASSO, and MCP generally lead to small numbers of selected SNPs (less than 60), CAR and COR variable sets are medium sized and NEG chooses very large number of SNPs.

Table 6.4: True SNPs found among the top 100 SNPs ordered by CAR scores in at least 50 of the 200 repetitions for Q1 and Q2.

	SNP	Frequency	MAF	BETA
Q1				
	FLT1 | C13S523	200	0.066714	0.64997
	FLT1 | C13S522	200	0.027977	0.61830
	FLT1 | C13S524	164	0.004304	0.62223
	KDR | C4S1877	145	0.000717	1.07706
	FLT1 | C13S431	110	0.017217	0.74136
	KDR | C4S1878	101	0.164993	0.13573
	KDR | C4S1884	95	0.020803	0.29558
	VEGFC | C4S4935	91	0.000717	1.35726
	ARNT | C1S6533	88	0.011478	0.56190
	VEGFA | C6S2981	69	0.002152	1.20645
Q2				
	VNN1 | C6S5380	138	0.170732	0.24437
	SIRT1 | C10S3050	72	0.002152	0.97060
	LPL | C8S442	69	0.015782	0.49459
	VNN3 | C6S5441	59	0.098278	0.27053
	BCHE | C3S4875	59	0.000717	1.09484
	VNN3 | C6S5449	57	0.010043	0.66909
	SIRT1 | C10S3048	54	0.002152	0.83224
	BCHE | C3S4869	54	0.000717	1.01569

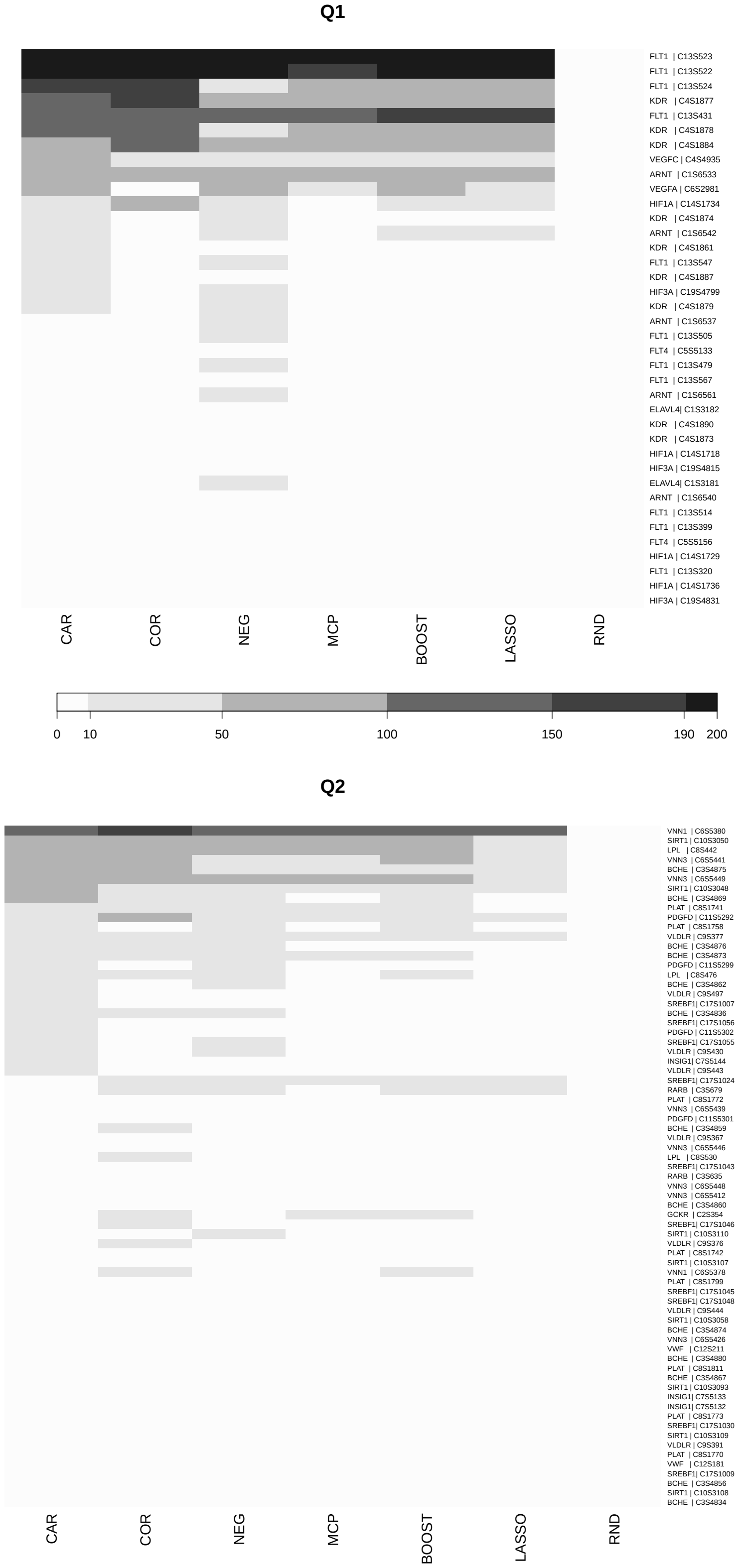


Figure 6.3: Frequency of occurrence of each true SNP among the top 100 SNPs selected by each approach for phenotype Q1 (top row) and for Q2 (lower row) for the 200 repetitions. Note that the SNPs are ordered according to the first column.

In further investigation of these results we identified the actual true SNPs recovered by each SNP selection approach. Specifically, we counted which of the 38, respectively 71 true causal SNPs for Q1 and Q2 were found among the first 100 top ranking SNPs using the 200 repetitions available for each phenotype. The result is shown as a heatmap in Figure **6.3** and visualizes the relative difficulty of recovering the individual causal SNPs. In Q1, there are two SNPs on top of the heatmap that are consistently detected by all methods. Then, there is a large block primarily recovered by CAR score and correlation, but not by the other approaches. Finally, there are some moderate detections only in CAR scores and NEG regression. Half of the true positives are hardly discovered by any method. The comparison with randomly ordered SNPs (column RND) shows that those SNPs only appear by chance. For Q2, there is only a single SNP that is consistently included in all models. As in Q1, it is followed by a small group of detections most prominent in CAR score and correlation. Finally, there are some moderate findings for both, the CAR score and NEG, and some only for correlation. In addition, hierarchical clustering of the columns (methods) in this heatmap (tree not shown in figure) reveals a basic similarity pattern among the methods: CAR and COR cluster together, NEG and MCP regression, and LASSO and BOOST.

In Table **6.4** we list the SNPs identified by CAR score among the top 100 SNPs in at least 50 of 200 repetitions along with their minor allele frequency (MAF) and BETA values. The BETA value measures the effect size in the actual simulation of the phenotype (Almasy et al., 2011). Interestingly, most of the SNPs recovered by CAR scores are rare SNPs with comparatively strong effects, i.e. large BETA values. Common SNPs are found as well, then also with small effect values. Thus, CAR scores are successful in achieving a high true positive rate because they not only allow to identify common SNPs but also SNPs with small MAF if a strong signal is present (large BETA).

In order to facilitate replication of our results we provide R code (R Development Core Team, 2012) covering all analysis steps from preprocessing the raw data to plotting of figures at `http://strimmerlab.org/software/care/`. The data are available by request from Jean MacCluer, see `http://www.gaworkshop.org/gaw17/data.html` for details.

6.3 Transcriptomics data: Analysis of gene expression data

The analysis of transcriptomics data is well established since several years. Especially, transcriptomics data is often used to identify biomarkers and to subsequently construct prediction rules for clinical diagnosis. Hence, this section reports results of classification in four benchmark data sets. Here, we are especially interested in the quality of the prediction by the three most established approaches, DDA, LDA, and PAM where the selection of relevant variables is based on t-scores, CAT scores, and shrunken centroids, respectively. Variable selection in the first four classification tasks is based on the FDR and FNDR as described in Section **5.3.4**. Finally, we investigate relations of gene-expression in the human brain with age to compare the performance of the CAR score to lasso and elastic net.

6.3.1 Classification of prostate cancer

The first classification task is based on a publication by Singh et al. (2002) who discuss stratification of patients suffering from prostate cancer. The authors investigate gene-expression of 102 patients and probands to derive a prediction rule based on gene-expression to identify patients at risk for recurrence of cancer. The variable of interest is binary, healthy ($n_1 = 50$) versus cancer ($n_2 = 52$), where the proportion of case to control is balanced. Thus, the presented task is classification of two groups. Additionally Singh et al. (2002) search genes correlated to clinical parameters, like for example the Gleason score, and provide gene rankings.

Here, we focus on prediction since we can compare different approaches by their ability to correctly classify patients and probands. The data provided comprises only training data that is gene-expression of $d = 6,033$ genes from $n = 102$ observations and a factor variable indicating the presence of cancer. Following Ahdesmäki and Strimmer (2010) we compute 200 estimates of the prediction error that is based on 200 splits from a balanced 10-fold cross-validation with 20 repetitions. Mean prediction error and corresponding standard errors are reported in Table **6.5** and illustrated in Figure **6.4**.

First, all approaches improve over the standard DDA model including all variables. Comparing variable selection based on t-score and CAT score we find that the LDA model using CAT scores consistently improves over the DDA model assuming independence. Furthermore, it is beneficial to include more variables by controlling FNDR, both in LDA and in DDA. The PAM algorithm fails to provide a stable model size, it ranges from 172-482. Still its mean prediction error is smaller than in the FDR approach of DDA and LDA that include only 53 to 62 variables. If DDA and LDA use the 166, respectively 131 variables designated by the FNDR approach, both approaches have a lower prediction error than PAM.

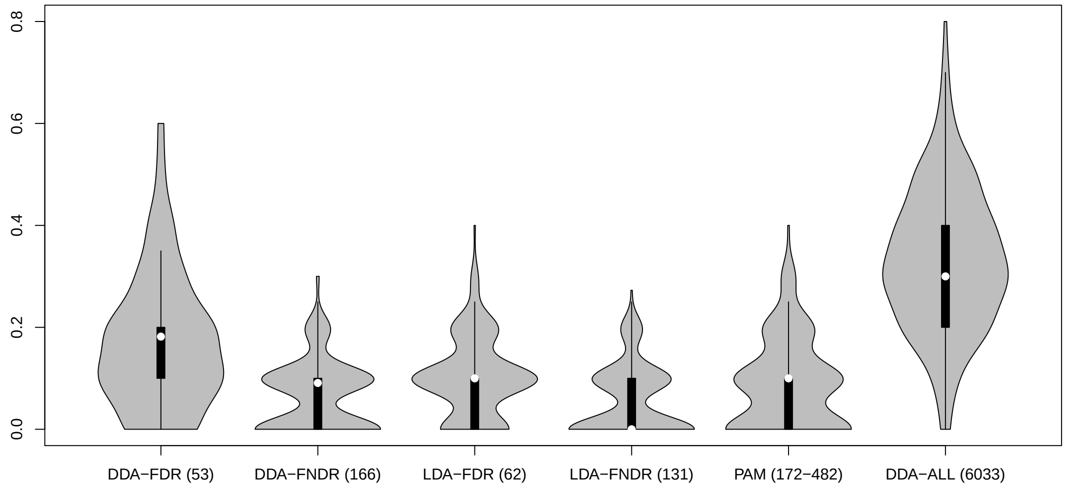


Figure 6.4: Classification of prostate cancer: Violinplots of the cross-validated prediction error based on 200 estimates with model size in brackets.

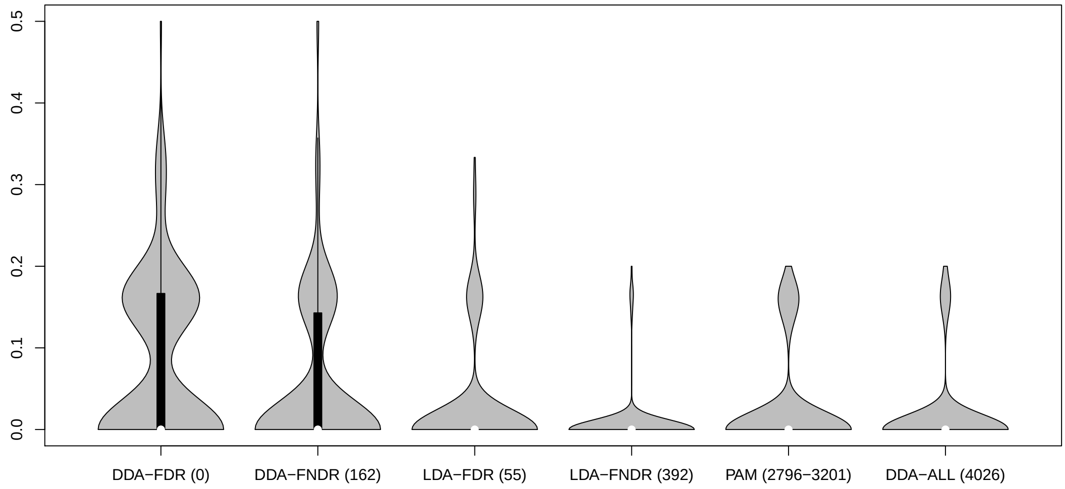


Figure 6.5: Classification of lymphoma: Violinplots of the cross-validated prediction error based on 200 estimates with model size in brackets.

Table 6.5: Cross-validated prediction errors with standard error (SE) and model size in classification of prostate cancer.

Model	Size	Prediction error	SE
DDA (FDR)	53	0.1682	0.0093
DDA (FNDR)	166	0.0640	0.0049
LDA (FDR)	62	0.0990	0.0055
LDA (FNDR)	131	0.0550	0.0048
PAM	72-482	0.0859	0.0063
DDA	6033	0.3327	0.0099

6.3.2 Classification of lymphoma

Next, we focus on multiclass prediction. The first data set describes the expression profile of $d = 4,026$ genes in $K = 3$ subtypes of lymphoma which was compiled by Alizadeh et al. (2000). Aim of this study is to find a genetic signature to discriminate between the three most prevalent subtypes of lymphoma. Altogether there are $n = 62$ observations, comprising $n_1 = 42$ in group 1, $n_2 = 9$ in group 2, and $n_3 = 11$ in group 3. Information on the mean prediction error based on 200 estimates from a balanced 10-fold cross-validation with 20 repetitions is given in Table **6.6** and illustrated in Figure **6.5**.

Table 6.6: Cross-validated prediction errors with standard error (SE) and model size in classification of lymphoma.

Model	Size	Prediction error	SE
DDA (FDR)	0	0.0805	0.0072
DDA (FNDR)	162	0.0536	0.0068
LDA (FDR)	55	0.0261	0.0047
LDA (FNDR)	392	0.0066	0.0023
PAM	2796-3201	0.0261	0.0045
DDA	4026	0.0165	0.0035

Above all, we find that only one approach performing variable selection improves over the DDA model containing all genes; it is LDA employing CAT scores for ranking and FNDR cut-off. With respect to the cut-off, models build on the FNDR rule perform distinctly better than models build on the FDR rule. Thus, it is beneficial to include more variables that are possibly false positives than to choose a too strict cut-off. Notably, in DDA not a single gene passes the FDR cut-off. Yet again considering correlation among genes in the LDA model provides a lower prediction error than DDA based on the independence assumption. This suggests variable selection using CAT scores yields a more informative set of variables than the *t*-score. The PAM algorithm fails to build a stable model, but gives a mean prediction error that is on par with LDA combined with the FDR cut-off.

6.3.3 Classification of small round blue cell tumour

A rather simple classification task is to discriminate between four subtypes of small, round blue cell tumors (SRBCTs) of childhood by the expression profile. While it is difficult to distinguish the four subtypes by standard histology, we show that it is possible to construct classifiers on gene-expression data that classify errorless. Khan et al. (2001) provide gene-expression data of $d = 2,308$ genes measured on $n = 82$ patients. Frequencies of the four subtypes of SRBCT are presented in Table **6.7**.

Table 6.7: Frequencies of the four subtypes of SRBCT.

Neuroblastoma	18
Rhabdomyosarcoma	25
Burkitt lymphomas	11
Ewing family of tumors	29

The mean prediction error over 200 repetitions is given in Table **6.8** and illustrated in Figure **6.6**. In particular, DDA as well as LDA using FNDR for variable selection derive perfect classification rules with no misclassification. Moreover, also DDA and LDA using FDR substantially improve over PAM and DDA using all variables.

Table 6.8: Cross-validated prediction errors with standard error (SE) and model size in classification of SRBCT.

Model	Size	Prediction error	SE
DDA (FDR)	62	0.0015	0.0011
DDA (FNDR)	90	0	0
LDA (FDR)	76	0.0014	0.0010
LDA (FNDR)	89	0	0
PAM	39-87	0.0105	0.0028
DDA	2308	0.0436	0.0057

6.3.4 Classification of brain cancer

A more difficult task is the discrimination of $K = 5$ subtypes of brain cancer. The data set including expression levels of $d = 5,597$ genes from $n = 42$ patients is provided in the R package `rda` (Guo et al., 2007). Pomeroy et al. (2002) compiled the data to derive a prediction rule to improve classification, since diagnosis on the basis of morphologic appearance is difficult. Correct diagnosis is vital for providing appropriate therapy as the outcome in terms of overall survival highly depends on the specific subtype of cancer. The frequency of the five subgroups is well balanced as can be seen from Table **6.9**.

Table 6.9: Frequencies of the five subtypes of brain cancer.

group	1	2	3	4	5
frequency	10	10	10	4	8

Surprisingly, the only technique improving over DDA on all genes is the LDA approach performing variable selection by CAT scores and setting the cut-off by FNDR. Again PAM failed to provide a model of stable size ranging from 197 to $5,597$ included genes. Again as depicted in Table **6.10** the FDR cut-off is rather strict and especially for DDA leads to a model including only eight genes, that performs rather poorly. Thus, we advice to set the cut-off by controlling the FNDR that includes a larger set of variables, i.e. 25 more genes in DDA and 89 in LDA. This leads to a decrease in prediction error of 0.1760 in DDA and 0.0594 in LDA.

Table 6.10: Cross-validated prediction errors with standard error (SE) and model size in classification of brain cancer.

Model	Size	Prediction error	SE
DDA (FDR)	8	0.3543	0.0175
DDA (FNDR)	33	0.1783	0.0137
LDA (FDR)	23	0.2110	0.0124
LDA (FNDR)	102	0.1516	0.0113
PAM	197-5597	0.1927	0.0121
DDA	5597	0.1618	0.0121

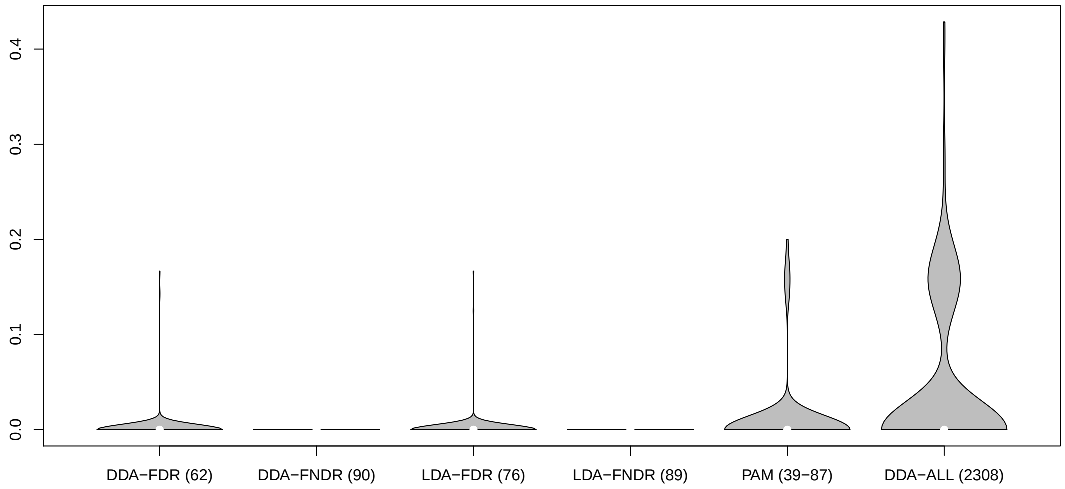


Figure 6.6: Classification of SRBCT: Violinplots of the cross-validated prediction error based on 200 estimates with model size in brackets.

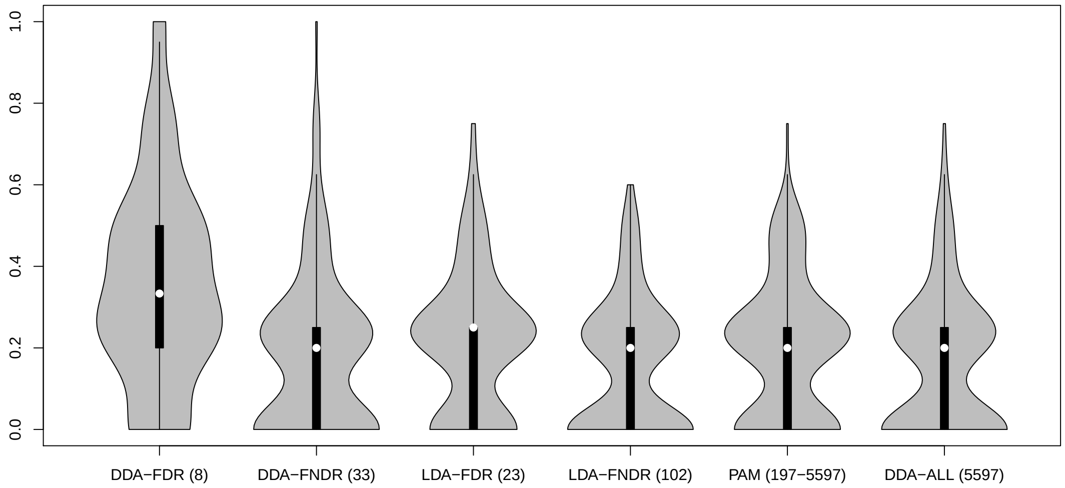


Figure 6.7: Classification of brain cancer: Violinplots of the cross-validated prediction error based on 200 estimates with model size in brackets.

6.3.5 Correlating gene-expression with age

To conclude the analysis of transcriptomics data we address a regression task. We analyze data from a gene-expression study investigating the relation of aging and gene-expression in the human frontal cortex (Lu et al., 2004). Specifically, the age of $n = 30$ patients was recorded, ranging from 26 to 106 years, and the expression of $d = 12,625$ genes was measured by microarray technology. In our analysis we used the age as metric response Y and the genes as explanatory variables $\boldsymbol{X}$. Our aim in the study of this data set is to find genes related to age or more precisely, a ranking of genes most affected by aging. Although it is quite absurd to derive a prediction rule for age based on gene-expression, it is quite essential to understand the functional changes in gene-expression depending on aging. Since we have no knowledge about the true underlying effects of aging, we decided to use the prediction error to assess the performance of variable selection.

Table 6.11: Cross-validated prediction errors with standard error (SE) in brackets resulting from lasso and elastic net in comparison with regression models of the CAR score for respective model size. Additionally, the prediction error for the CAR model, including 60 variables, with the lowest prediction error is given.

Model (Size)	Prediction error (SE)
LASSO (36)	0.4006 (0.0011)
ELASTIC NET (85)	0.3417 (0.0068)
CAR (36)	0.3357 (0.0070)
CAR (60)	0.3049 (0.0064)
CAR (85)	0.2960 (0.0059)

In preprocessing we removed genes with negative values and log-transformed the expression values of the remaining $d = 11,940$ genes. We centered and standardized the data and computed empirical marginal correlations. Subsequently, based on marginal correlations we filtered out all genes with local false non-discovery rates smaller than 0.2, following Ahdesmäki and Strimmer (2010). Thus, in this prescreening step we retained the $d = 403$ variables with local false-discovery rates smaller than 0.8.

On this 30×403 data matrix we fitted regression models using shrinkage CAR, lasso, and elastic net. The optimal tuning parameters were selected by minimizing the prediction error estimated by 5-fold cross-validation with 100 repeats. Cross-validation included model selection as integrative step, e.g., CAR scores were recomputed in each repetition in order to avoid downward bias. A summary of the results is found in Table **6.11**. The

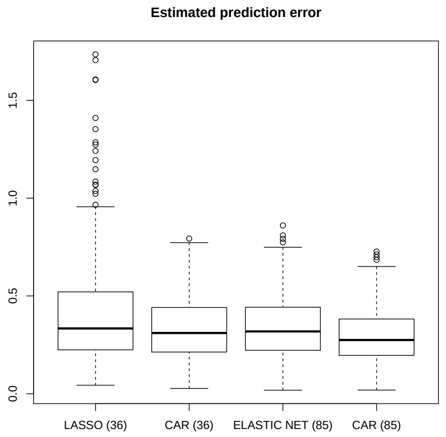


Figure 6.8: Comparison of cross-validated prediction errors of lasso and CAR regression model for corresponding model size of 36 included genes and comparison of elastic net and CAR regression model for corresponding model size of 85 included genes. The model size is given in brackets.

prediction error of the elastic net regression model is substantially smaller than that of the lasso model, at the cost of 49 additionally included covariates. The regression model suggested by the CAR approach for the same model sizes improves over both models Figure **6.8**. As can be seen from Figure **6.9** the optimal CAR regression model has a size of about 60 predictors. The inclusion of additional explanatory variables does not substantially improve prediction accuracy.

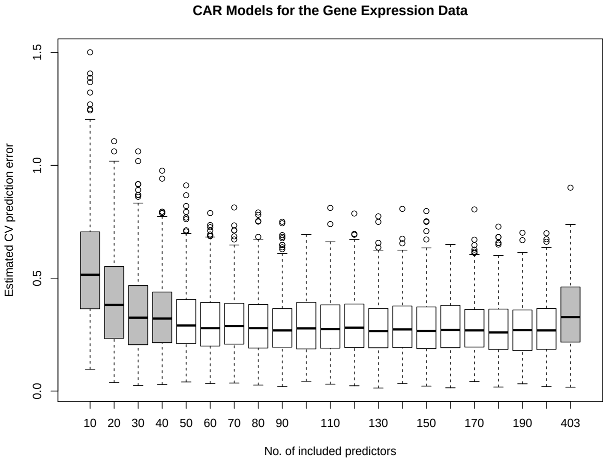


Figure 6.9: Comparison of cross-validated prediction errors of CAR regression models of various sizes.

6.4 Metabolomics data: Ranking of markers for prostate cancer

The analysis of metabolomics data has emerged only recently and in contrast to transcriptomics data there are no clear reference data, yet. To illustrate the application of the CAT score on metabolomics data we investigate a subset of preprocessed data from a recent metabolomic study concerning prostate cancer (Sreekumar et al., 2009). Here, we use the preprocessed data as kindly provided by Dr. Sreekumar and Dr. Chinnaiyan.

In particular, we focus on the effect of correlation between metabolites on variable ranking. The original study investigated three groups of tissues, benign, localized cancer and metastatic prostate cancer. Here, we focused on the two types of cancer tissue. Specifically, we compared 12 samples of clinically localized prostate cancers versus 14 samples of metastatic prostate cancers. For each sample the concentrations of 518 metabolites were measured.

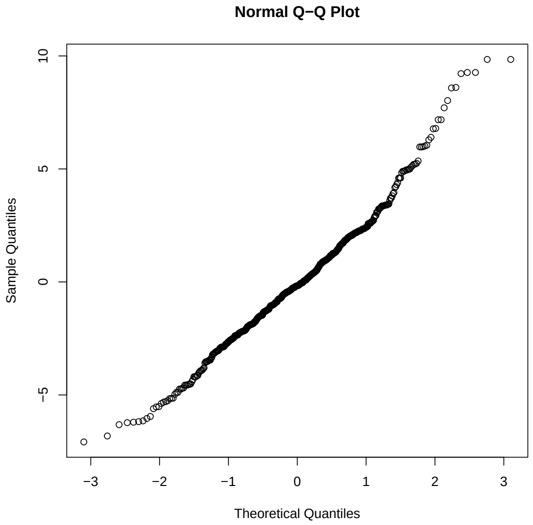


Figure 6.10: Plot of normal versus empirical quantiles for the grouped CAT scores computed from the metabolomic prostate data. The linearity in the central part indicates a normal null model as approximate of the t-distribution.

We computed a shrinkage t-score and a shrinkage CAT score for each of the 518 metabolites. For the latter we applied grouping of features with a correlation threshold of $|r| \geq 0.85$. Since $n = 26$ we approximate the t-distribution by the normal distribution (Fahrmeir et al., 2003) and illustrate this by a Q-Q-plot of CAT scores versus a normal distribution in Figure **6.10**. By inspection of this diagnostic plot we see that the null model of the grouped CAT scores, represented by the linear middle part, is approximately a normal distribution. The deviations from normality at the tails correspond to the alternative distribution containing the high-ranked metabolites of interest.

The ten top ranking metabolic features that differentiate between localized and metastatic cancer according to t-scores and CAT scores, respectively, are listed in Table **6.12**. Overall, the two rankings differ quite notably, as expected in the presence of correlation. In particular, at the top of the list there are differences due to very strong correlation between the substrate X-5207 and Nicotinamide ($r = 0.9444$) and likewise between Guanosine and X-3390 ($r = 0.9389$). Unlike with t-scores, in a grouped CAT score analysis the features in these two pairs are treated as a unit. Jointly, the correlated markers outperform other individual markers with respect to distinguishing between the two phenotypic groups.

Table 6.12: The top ten ranking metabolites according to the shrinkage t and the grouped CAT scores, respectively. Note that nicotinamide and X-5207, as well as guanosine and X-3390, are strongly correlated.

Rank	shrinkage t	grouped CAT score
1	Ciliatine	Nicotinamide
2	Inosine	X-5207
3	Putrescine	Guanosine
4	X-3390	X-3390
5	Palmitate	Ciliatine
6	Glycerol	Putrescine
7	Ribose	Inosine
8	X-3102	Citrate
9	Myristate	Uridine
10	X-4620	X-2867

To compare the quality of the two top ten lists we explore the predictive quality of the ten top ranked metabolites by reporting the prediction error that was estimated by a balanced 5-fold cross-validation with 100 repetitions. Corresponding prediction errors of a DDA model based on the top ten variables listed by the shrinkage t and of the LDA model based on the top ten variables listed by the shrinkage CAT score are reported in Table **6.13**. Cross-validation included model selection based on the respective scores and including the top ten variables in each step. Evidently, the CAR score provides a top ten of metabolites that discriminates more precisely between the cancer subtypes than the metabolite set selected by shrinkage t.

Table 6.13: Frequency table and mean of the prediction error in a balanced 5-fold cross-validation with 100 repetitions (500 splits) for shrinkage *t* (DDA) and CAT score (LDA)

error in CV	0	1/6	1/5	1/4	1/3	2/5	1/2	3/5	mean
shrinkage *t*	442	10	26	5	6	6	3	3	0.0316
grouped CAT score	459	11	20	4	1	2	2	2	0.0203

Chapter 7

Conclusion

In this thesis we have introduced a multivariate framework for variable selection and biomarker identification that explicitly takes account of the correlation structure among markers. We were concerned with the two most important applications of variable selection: Classification and regression. We presented two novel scores that quantify the importance of a variable in a multivariate setting. Specifically, we proposed the correlation-adjusted t (CAT) score in classification, and the correlation-adjusted (marginal) correlation (CAR) score in regression.

In Chapter **3** we addressed variable selection in classification. First, we reviewed existing approaches to select variables for prediction and ranking. Then, in Section **3.3** we introduced the CAT score that we defined as the Mahalanobis-decorrelated t-score vector. We derived the CAT score from LDA where it quantifies the influence of a decorrelated and standardized variable on the prediction rule.

Chapter **4** discussed variable selection in regression. We outlined penalized regression for prediction and presented the concept of variable importance for ranking. Subsequently, we introduced the CAR score in Section **4.4** as the Mahalanobis-decorrelated marginal correlation vector. Like the CAT score, the CAR score is derived from a predictive point of view; it quantifies the weight of a standardized and decorrelated variable on the prediction of the standardized response. Moreover, the CAR score represents the correlation between the response and the decorrelated covariables. We argued that the CAR score is the central quantity to assess which variables contribute to the explained variance, or equivalently reduce the unexplained variance, since the decomposition of variances in the linear model can be rewritten in terms of CAR scores $\boldsymbol{\omega}$ as

$$\begin{aligned}
\overbrace{\text{Var}(Y)}^{\text{Total variance}} &= \overbrace{\text{Var}(Y^\star)}^{\text{Explained variance}} + \overbrace{\text{Var}(Y - Y^\star)}^{\text{Unexplained variance}} \\
\sigma_Y^2 &= \sigma_Y^2(\boldsymbol{\omega}^T\boldsymbol{\omega}) + \sigma_Y^2(1 - \boldsymbol{\omega}^T\boldsymbol{\omega})\,.
\end{aligned}$$

Both CAT score and CAR score are embedded in the same multivariate framework of *decorrelation* and hence share important properties. CAT and CAR score adjust the quantities deemed optimal in case of no correlation by the Mahalanobis transform and thus incorporate information on the correlation structure among covariables into the selection of variables. Moreover, they represent intermediates between these marginal quantities and the standardized β-coefficients from logistic, respectively linear regression. Notably, CAT and CAR score decompose the multivariate statistics established to assess the effect of sets of variables. The CAT score decomposes Hotelling's T^2, and the CAR score the proportion of variance explained. This suggests that Hotelling's T^2 in LDA is the analogon to the proportion of variance explained in linear regression.

Generally, we emphasized that strategies for variable selection differ, whether they aim at *prediction*, or *ranking*:

- We derived both, the CAT and the CAR score, from a *predictive point of view* in LDA and linear regression. LDA is the generalization of DDA to correlation among the predictor variables; similarly, linear regression generalizes marginal correlation to correlation among the predictor variables. We showed that both scores quantify the influence of a decorrelated and standardized variable on the respective prediction rule. Thus, we argued that CAT and CAR score are effective criteria to assess the importance of a variable in the multivariate, predictive framework of LDA and linear regression, respectively. To support this argument we conducted extensive simulation studies, and analyzed transcriptomics and metabolomics data where we showed that our approaches are competitive, or outperform other techniques in terms of a lower prediction error.

- CAT and CAR score quantify the importance of a variable in the multivariate framework. Hence, our scores can be employed to *rank variables according to their importance*. In classification, a variable is considered important if it discriminates between the given groups. In linear regression, we considered a variable important if it provides a contribution or share to the variance explained.

  Still, there are two properties that need to be taken into account when using our approaches in ranking. First, in the presence of highly correlated variables we advised to treat these variables as a group and quantify the importance of the whole group by the grouped versions of our scores. Otherwise the share of variance explained by the whole set of variables is split into equal, but decisively smaller parts for each variable. To recover the joint effect of such a group of variables the grouped versions of our scores evaluate the share of all variables in the group combined. Additionally, highly correlated variables have approximately equal scores due to a grouping property and are thus found on adjacent positions in a ranking. Investigating the quality of rankings from simulations and genomics studies we ascertained that our approaches overall detect more true positives than competing scores.

Both quantities are defined on the population level and hence are independent of any inference paradigm. In Chapter **5** we derived shrinkage estimates for CAT and CAR scores, and presented an efficient algorithm that allows the application of our approach to high-dimensional omics data. Thus, the CAT and the CAR score are multivariate quantities for variable importance that are defined in a parametric framework, and applicable on high-dimensional data.

Finally, Chapter **6** illustrated the identification of biomarkers by CAT and CAR scores in high-dimensional omics data. In particular, we analyzed data from genomics (Section **6.2**), transcriptomics (Section **6.3**), and metabolomics experiments (Section **6.4**). We were able to affirm that variable selection by CAT and CAR scores improves the identification of biomarkers in terms of more true positives detected and lower prediction error than competing approaches.

For investigating molecular processes in the cell, new high-throughput technologies are continuously emerging, such as time-resolved mass-spectrometry for proteomics, or next generation sequencing for transcriptomics (RNA-seq). These “next generation” data are again high-dimensional and also exhibit an intricate correlation structure due to complex patterns of cellular regulation. Therefore, statistical tools for high-dimensional variable selection under correlation, such as the CAT and CAR score developed in this thesis, will be advantageous to improve the identification of biomarkers in future molecular experimental studies.

Bibliography

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

Ahdesmäki, M. and Strimmer, K. (2010). Feature selection in omics prediction problems using cat scores and false non-discovery rate control. *Ann. Appl. Statist.*, 4:503–519.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control*, 19:716–723.

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., and Staudt, L. M. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511.

Allison, M. (2008). Is personalized medicine finally arriving? *Nature Biotechnology*, 26:509–519.

Almasy, L., Dyer, T. D., Peralta, J. M., Kent Jr., J. W., Charlesworth, J. C., Curran, J. E., and Blangero, J. (2011). Genetic Analysis Workshop 17 mini-exome simulation. *BMC Proceedings*, 5 (Suppl. 9):S2.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA*, 96:6745–6750.

Alter, O., Brown, P. O., and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA*, 97:10101–10106.

Ardlie, K. G., Kruglyak, L., and Seielstad, M. (2002). Patterns of linkage disequilibrium in the human genome. *Nature Reviews Genetics*, 3:299–309.

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics*, 11:375–386.

Ayers, K. L. and Cordell, H. (2010). SNP selection in genome-wide and candidate gene studies via penalized logistic regression. *Genet. Epidemiol.*, 34:879–891.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, 57:289–300.

Bickel, P. J. and Levina, E. (2004). Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10:989–1010.

Box, J. F. (1987). Guinness, Gosset, Fisher, and small samples. *Statist. Sci*, 2:45–52.

Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Ann. Applied Statistics*, 5:232–253.

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Bring, J. (1994). How to standardize regression coefficients. *The American Statistician*, 48:209–213.

Burnham, K. P. and Anderson, D. (2002). *Model selection and multi-model inference*. Springer, 2 edition.

Choe, S. E., Boutros, M., Michelson, A. M., Church, G. M., and Halfon, M. S. (2005). Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control data set. *Genome Biology*, 6:R16.

Crick, F. (1958). On protein synthesis. *Symposia of the Society for Experimental Biology*, 12:138–163.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227:561– 563.

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *J. Amer. Statist. Assoc.*, 70:892–896.

Efron, B. (2008). Microarrays, empirical Bayes, and the two-groups model. *Statist. Sci.*, 23:1–22.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression (with discussion). *Ann. Statist.*, 32:407–499.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.*, 96:1151–1160.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95:14863–14868.

Fahrmeir, L., Künstler, R., Pigeot, I., and Tutz, G. (2003). *Statistik*. Springer.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate statistical modelling based on generalized linear models*. Springer, 2nd edition.

Fan, J. and Fan, Y. (2008). High-dimensional classification using features annealed independence rules. *Ann. Statist.*, 36:2605–2637.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96:1348–1361.

Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space (with discussion). *J. R. Statist. Soc. B*, 70:849–911.

Firth, D. (1998). Relative importance of explanatory variables. In *Conference on Statistical Issues in Social Sciences, Stockholm, October 1998*.

Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.*, 22:1947–1975.

Foulkes, A. S. (2009). *Applied Statistical Genetics with R*. Springer, New York.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Statist. Soft.*, 39:1–13.

Genizi, A. (1993). Decomposition of R^2 in multiple regression with correlated regressors. *Statistica Sinica*, 3:407–420.

George, E. I. (2000). The variable selection problem. *J. Amer. Statist. Assoc.*, 95:1304–1308.

Golub, T., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P. ., Coller, H., Loh, M. L., Downing, J., andC. D. Bloomfield, M. A. C., and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537.

Grömping, U. (2006). Relative importance in linear regression in R: the package relaimpo. *J. Statist. Soft.*, 17:1.

Grömping, U. (2007). Estimators of relative importance in linear regression based on variance decomposition. *The American Statistician*, 61:139–147.

Guan, Y. and Stephens, M. (2011). Bayesian variable selection regression for genome-wide association studies, and other large-scale problems. *Ann. Appl. Statist.*, 5:1780–1815.

Guillemot, V., Le Brusquet, L., Tenenhaus, A., and Frouin, V. (2008). Graph-constrained discriminant analysis of functional genomics data. In *IEEE International Conference on Bioinformatics and Biomedicine*, Philadelphia, PA, USA.

Guo, Y., Hastie, T., and Tibshirani, T. (2007). Regularized discriminant analysis and its application in microarrays. *Biostatistics*, 8:86–100.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *JMLR*, 3:1157–1182.

Hamburg, M. A. and Collins, F. S. (2010). The path to personalized medicine. *NEW ENGL J MED*, 363:301–304.

Hand, D. J. (2006). Classifier technology and the illusion of progress. *Statist. Sci.*, 21:1–14.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition.

Hastie, T. and Tibshirani, T. (2004). Efficient quadratic regularization for expression arrays. *Biostatistics*, 5:329–340.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O. P., Wilfond, B., Borg, A., and Trent, J. (2001). Gene-expression profiles in hereditary breast cancer. *N. Engl. J. Med.*, 344:539–548.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67.

Hoffman, P. J. (1960). The paramorphic representation of clinical judgment. *Psychol. Bull.*, 57:1116–131.

Hoggart, C. J., Whittaker, J. C., M. De Iorio, and Balding, D. J. (2008). Analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genetics*, 4:e1000130.

Holmes, E., Wilson, I., and Nicholson, J. (2008). Metabolic phenotyping in health and disease. *Cell*, 134:714–717.

Hotelling, H. (1931). The generalization of Student's ratio. *Ann. Math. Statist.*, 2:360–378.

Hotelling, H. (1953). New light on the correlation coefficient and its transforms. *J. R. Statist. Soc. B*, 15:193–232.

Hothorn, T. and Bühlmann, P. (2006). Model-based boosting in high dimensions. *Bioinformatics*, 22:2828–2829.

Huber, W., von Heydebreck, A., Sueltmann, H., Poustka, A., and Vingron, M. (2003). Parameter estimation for the calibration and variance stabilization of microarray data. *Statist. Appl. Genet. Mol. Biol.*, 2:3.

Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A., and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18:S96–S104.

Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent Component Analysis*. John Wiley & Sons.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, 409:860–921.

Kerr, M. K., Martin, M., and Churchill, G. A. (2001). Analysis of variance for gene expression microarray data. *J. Comp. Biol.*, 7:819–837.

Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., and Meltzer, P. S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Med.*, 7:673–679.

Kruskal, W. (1987). Relative importance by averaging over orderings. *The American Statistician*, 41:6–10.

Läuter, J., Horn, F., Rosolowski, M., and Glimm, E. (2009). High-dimensional data analysis: selection of variables, data compression and graphics — applications to gene expression. *Biometr. J.*, 51:235–251.

Lewis, C. (2002). Genetic association studies: Design, analysis and interpretation. *Briefings in Bioinformatics*, 3:146–153.

Li, C. and Li, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24:1175–1182.

Liebermeister, W. (2002). Linear modes of gene expression determined by independent component analysis. *Bioinformatics*, 18:51–60.

Lu, T., Pan, Y., Kao, S.-Y., Li, C., Kohane, I., Chan, J., and Yankner, B. A. (2004). Gene regulation and DNA damage in the ageing human brain. *Nature*, 429:883–891.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, 15:661–675.

Nilsson, R., Pena, J. M., Björkegren, J., and Tegnér, J. (2007). Consistent feature selection for pattern recognition in polynomial time. *JMLR*, 8:589–612.

Opgen-Rhein, R. and Strimmer, K. (2007a). Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. *Statist. Appl. Genet. Mol. Biol.*, 6:9.

Opgen-Rhein, R. and Strimmer, K. (2007b). From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC Systems Biology*, 1:37.

Opgen-Rhein, R. and Strimmer, K. (2007c). Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process. *BMC Bioinformatics*, 8 (Suppl. 2):S3.

Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., Baehner, F. L., Walker, M. G., Watson, D., Park, T., Hiller, W., Fisher, E. R., Wickerham, D. L., Bryant, J., and Wolmark, N. (2004). A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *NEW ENGL J MED*, 351:2817–26.

Pollard, T. D. and Earnshaw, W. C. (2007). *Cell Biology*. Elsevier.

Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., Sturla, L. M., Angelo, M., McLaughlin, M. E., Kim, J. Y. H., Goumnerova, L. C., Black, P. M., Lau, C., Allen, J. C., Zagzag, D., Olson, J. M., Curran, T., Wetmore, C., Biegel, J. A., Poggio, T., Mukherjee, S., Rifkin, R., Califano, A., Stolovitzky, G., Louis, D. N., Mesirov, J. P., Lander, E. S., and Golub, T. R. (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415:436–442.

Pratt, J. W. (1987). Dividing the indivisible: using simple symmetry to partion variance explained. In Pukkila, T. and Puntanen, S., editors, *Proceeding of Second Tampere Conference in Statistics*, pages 245–260. University of Tampere, Finland.

R Development Core Team (2012). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.-H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J. P., Poggio, T., Gerald, W., Loda, M., Lander, E. S., and Golub, T. R. (2001). Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci*, 98:15149–15154.

Rüger, B. (1998). *Test- und Schätztheorie: Band 1, Grundlagen*. Oldenbourg Wissenschaftsverlag.

Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statist. Appl. Genet. Mol. Biol.*, 4:32.

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, 5:197–227.

Schilsky, R. L. (2010). Personalized medicine in oncology: The future is now. *Nature Reviews Drug Discovery*, 9:363–366.

Schuster, S. (2008). Next-generation sequencing transforms today's biology. *Nature Methods*, 5:16–19.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, 6:461–464.

Schweder, T. and Spjøtvoll, E. (1982). Plots of p-values to evaluate many tests simultaneously. *Biometrika*, 69:493–502.

Shendure, J. and Hanlee, J. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, 26:1135 – 1145.

Singh, D., Febbo, P. G., Ross, K., 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., and Sellers, W. R. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1:203–209.

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statist. Appl. Genet. Mol. Biol.*, 3:3.

Sreekumar, A., Poisson, L. M., Rajendiran, T. M., Khan, A. P., Cao, Q., Yu, J., Laxman, B., Mehra, R., Lonigro, R. J., Li, Y., Nyati, M. K., Ahsan, A., Kalyana-Sundaram, S., Han, B., Cao, X., Byun, J., Omenn, G. S., Ghosh, D., Pennathur, S., Alexander, D. C., Berger, A., Shuster, J. R., Wei, J. T., Varambally, S., Beecher, C., and Chinnaiyan, A. M. (2009). Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature*, 457:910–914.

Stella, C., Beckwith-Hall, B., Cloarec, O., Holmes, E., Lindon, J., Powell, J., van der Ouderaa, F., Bingham, S., Cross, A., and Nicholson, J. (2006). Susceptibility of human metabolic phenotypes to dietary modulation. *J Proteome Res*, 5:2780–2788.

Strimmer, K. (2008a). fdrtool: a versatile R package for estimating local and tail area-based false discovery rates. *Bioinformatics*, 24:1461–1462.

Strimmer, K. (2008b). A unified approach to false discovery rate estimation. *BMC Bioinformatics*, 9:303.

Strobl, C., Boulesteix, A. L., Kneib, T., and Augustin, T. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, 9:1471–2105.

Strobl, C., Boulesteix, A. L., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8:1471–2105.

Student (1908). The probable error of a mean. *Biometrika*, 6:1–25.

Tai, F. and Pan, W. (2007). Incorporating prior knowledge of gene functional groups into regularized discriminant analysis of microarray data. *Bioinformatics*, 23:3170–3177.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B*, 58:267–288.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of multiple cancer type by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. USA*, 99:6567–6572.

Tibshirani, R. and Wasserman, L. (2006). Correlation-sharing for detection of differential gene expression. *arXiv*, math.ST:math/0608061.

Tusher, V., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA*, 98:5116–5121.

Veer, L. J., v. Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., and Witteveen, A. T. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–535.

Wang, S., Nan, B., Rosset, S., and Zhu, J. (2011). Random lasso. *Ann. Applied Statistics*, 5:468–485.

Wang, Z., Gerstein, M., and Snyder, M. (2009). Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10:57–63.

Watson, J. and Crick, F. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171:737–738.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley, New York.

Wishart, D., Tzur, D., Knox, C., Eisner, R., Guo, A. C., Young, N., Cheng, D., K. Jewell, D. A., Sawhney, S., Fung, C., Nikolai, L., Lewis, M., Coutouly, M., Forsythe, I., Tang, P., Shrivastava, S., Jeroncic, K., Stothard, P., Amegbey, G., Block, D., Hau, D., Wagner, J., Miniaci, J., Clements, M., Gebremedhin, M., Guo, N., Zhang, Y., Duggan, G. E., MacInnis, G. D., Weljie, A. M., Dowlatabadi, R., Bamforth, F., D. Clive, R. G., Li, L., Marrie, T., Sykes, B. D., Vogel, H. J., and Querengesser, L. (2007). Hmdb: The human metabolome database. *Nucleic Acids Research*, 35:521–526.

Witten, D. M. and Tibshirani, R. (2009). Covariance-regularized regression and classification for high-dimensional problems. *J. R. Statist. Soc. B*, 71:615–636.

Witten, D. M. and Tibshirani, R. (2011). Penalized classification using fisher's linear discriminant. *J. R. Statist. Soc. B*, 73:753–772.

Woodcock, J. (2007). The prospects for "personalized medicine' ' in drug development and drug therapy. *Clinical Pharmacology & Therapeutics*, 81:164–169.

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., and Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25:714–721.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, 38:894–942.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B*, 67:301–320.

Appendix A

Notation

A.1 Symbols

$\boldsymbol{X} = (X_1, ..., X_d)^T$	explanatory variables (column vector of length $d \times 1$)
$i \in 1, ..., d$	index for variables
$j \in 1, ..., d$	index for variables
$\boldsymbol{x}$	observed explanatory variables
$l \in 1, ..., n$	index for observations
Y	variable of interest
$\boldsymbol{y}$	observed variable of interest

$\boldsymbol{\mu}$	vector of expectations of $\boldsymbol{X}$ (column vector of length $d \times 1$)
$\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$	vector of expectations in group 1, respectively 2
$\bar{\boldsymbol{x}}$	mean vector
$\boldsymbol{\Sigma} = \boldsymbol{V}^{1/2}\boldsymbol{P}\boldsymbol{V}^{1/2}$	covariance matrix of $\boldsymbol{X}$
$\boldsymbol{V}$	variance matrix of $\boldsymbol{X}$
	with σ_i^2, with $i \in 1, ..., d$, on its diagonal
$\boldsymbol{S}$	sample variance matrix of $\boldsymbol{x}$
	with s_i^2, with $i \in 1, ..., d$, on its diagonal
$\boldsymbol{P}$	correlation matrix of $\boldsymbol{X}$
ρ	correlation between two explanatory variables
$\boldsymbol{R}$	sample correlation matrix of $\boldsymbol{x}$
$n_1 + n_2 = n$	sample size in group 1, respectively 2

$k \in 1, ..., K$	class index for categorical variable of interest
$\boldsymbol{\tau}$	t-score vector with scale factor c_n on population level
$\boldsymbol{t}$	empirical t-score vector
$\boldsymbol{\beta} = (\beta_1, ..., \beta_d)^T$	β-coefficients ($d \times 1$) from logistic regression
β_0	intercept from logistic regression
$\boldsymbol{\omega}$	weight vector from decomposition of the prediction rule
$\delta(\boldsymbol{x})$	standardized and decorrelated data
$\boldsymbol{\tau}^{adj}$	CAT score vector with scale factor c_n on population level
$\boldsymbol{t}^{adj}$	empirical CAT score vector
λ	shrinkage parameter
$\boldsymbol{I}_d$	identity matrix of size $d \times d$

Y	variable of interest (quantitative)
μ_Y	expectation of Y
σ_Y^2	variance of Y
$\bar{y}$	mean of Y
$\boldsymbol{\beta} = (\beta_1, ..., \beta_d)^T$	regression coefficients, column vector of length d
β_0	intercept
$\boldsymbol{b}$	estimated regression coefficients
b_0	estimated intercept
$Y^\star = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{X}$	best linear predictor
$\boldsymbol{P}_{XY} = (\rho_1, \dots, \rho_d)^T$	(marginal) correlations between $\boldsymbol{X}$ and Y
$\boldsymbol{R}_{XY}$	estimated (marginal) correlations
$\boldsymbol{\Sigma}_{XY}$	covariance between $\boldsymbol{X}$ and Y
Ω^2	multiple correlation coefficient
R^2	empirical multiple correlation coefficient or coefficient of determination
$\boldsymbol{\omega}$	the CAR score on population level (column vector, $d \times 1$)
$\hat{\boldsymbol{\omega}}$	empirical CAR score
λ	shrinkage parameter
ζ, θ	parameter for specification of prior distributions
$\phi(x_i)$	a measure of importance of variable X_i
r	estimated correlation coefficient
$\tilde{\boldsymbol{P}}_{XY} = (\tilde{\rho}_1, \dots, \tilde{\rho}_d)^T$	partial correlation coefficients
$\tilde{r}$	partial estimated correlation coefficient
$\boldsymbol{\tau}_{XY} = (\tau_1, \dots, \tau_d)^T$	t-score vector to test β-coefficients in regression

$\boldsymbol{R} = \boldsymbol{U}\boldsymbol{M}\boldsymbol{U}^T$	eigenvalue decomposition of the correlation matrix $\boldsymbol{R}$
$\boldsymbol{U}$	orthogonal eigenvector system
$\boldsymbol{M}$	diagonal matrix containing the eigenvalues
f_0	null distribution
f_A	alternative distribution
z	summary statistic
η_0	mixing constant

$N(\boldsymbol{\mu}, \Sigma)$	normal or gaussian distribution with $\boldsymbol{\mu}$ as vector of expectations Σ as covariance matrix
$t(\text{df})$	Student t distribution with df degrees of freedom
$\chi^2(\text{df})$	Chi squared distribution with df degrees of freedom
$F(\text{df}_1, \text{df}_2)$	F-distribution with df_1 and df_2 degrees of freedom
$\text{Beta}(\text{df}_1/2, \text{df}_2/2)$	Beta distribution with df_1 and df_2 degrees of freedom

A.2 Abbreviations

CAT	correlation-adjusted t-score
CAR	correlation-adjusted correlation
CV	cross-validation
DDA	diagonal discriminant analysis
FDR	false discovery rate
FNDR	false non discovery rate
FN	false negative
FP	false positive
GWAS	genome-wide association study
LDA	linear discriminant analysis
ME	model error
RSS	residual sums of squares
SE	standard error
SNP	single-nucleotide polymorphism
TDR	true discovery rate
TN	true negative
TP	true positive

Appendix B

Software

B.1 Implementation in `R`

We implemented CAT and CAR score in the free statistical computing language `R` (R Development Core Team, 2012) under the GNU General Public License. Precisely we implemented the CAT score in the package `st`, and the CAR score in the `care`-package. Additionally, the CAT score is used for selecting variables in the package `sda` by Ahdesmäki and Strimmer (2010) and the CAR score is listed as one of several measures for variable importance in the `relaimpo` package by Grömping (2006).

To name the most important functions from the `st`, `sda` and the `care` package:

- `shrinkcat.stat(X, L, verbose=TRUE)`
  returns the shrinkage CAT score, where
  - `X` is the data matrix of size $n \times d$ where the columns represent the variables and the rows the observations
  - `L` is a binary factor variable to designate the membership to one of the two groups
  - `verbose=TRUE` prints additional information on the shrinkage parameters while computing
- `sda(Xtrain, L, diagonal=FALSE, verbose=TRUE)`
  fits a classification rule for shrinkage discriminant analysis, where variable selection is performed using CAT scores (LDA) or shrinkage *t*-scores (DDA)
  - `X` is the data matrix of size $n \times d$ where the columns represent the variables and the rows the observations
  - `L` is a binary factor variable to designate the membership to one of the two groups

  - `diagonal=FALSE` adjusts for correlation among predictor variables (LDA), `diagonal=TRUE` discards the correlation among predictor variables (DDA)
  - `verbose=TRUE` prints additional information on the shrinkage parameters while computing

- `carscore(Xtrain, Ytrain, lambda, diagonal=FALSE, verbose=TRUE)` computes the shrinkage CAR score, where
  - `Xtrain` is the data matrix of size $n \times d$ where the columns represent the variables and the rows the observations
  - `Ytrain` is the metric variable of interest
  - `lambda` quantifies the regularization (ranging from `lambda=0` for empirical estimates to `lambda=1` for shrinking all off-diagonal elements to zero). If not specified `lambda` is estimated by the analytic formula in Opgen-Rhein and Strimmer (2007b).
  - `diagonal=FALSE` adjusts for correlation among predictor variables, otherwise `diagonal=TRUE` returns the shrinkage marginal correlation
  - `verbose=TRUE` prints additional information on the shrinkage parameters while computing

- `slm(Xtrain, Ytrain, lambda, lambda.var, diagonal=FALSE)` provides the most important quantities of the (shrinkage) linear model
  - `Xtrain` is the data matrix of size $n \times d$ where the columns represent the variables and the rows the observations
  - `Ytrain` is the metric variable of interest
  - `lambda` quantifies the regularization (ranging from `lambda=0` for empirical estimates to `lambda=1` for shrinking all off-diagonal elements to zero). If not specified `lambda` is estimated by the analytic formula in Opgen-Rhein and Strimmer (2007b).
  - `lambda.var` quantifies the regularization of the variance. If not specified `lambda.var` is estimated by the analytic formula in Schäfer and Strimmer (2005).
  - `diagonal=FALSE` adjusts for correlation among predictor variables, otherwise `diagonal=TRUE` returns the shrinkage marginal correlation

The functions are freely available under the GNU license from the CRAN archive `http://cran.r-project.org/web/packages/care/`, respectively (`http://cran.r-project.org/web/packages/st/`).

B.2 Step by step analysis of the benchmark diabetes data

For illustration we sketch an analysis of the benchmark data on progression of diabetes by Efron et al. (2004). We provide the dataset in the `care` package. The data is loaded into the workspace by

```
library(care)
data(efron2004)

Y = efron2004$y
X = efron2004$x
cn = colnames(efron2004$x)
n = dim(X)[1]      # 442 observations
d = dim(X)[2]      # 10  variables
```

where `X` is the matrix of $d = 10$ predictor variables on $n = 442$ observations and `Y` the metric variable of interest that describes the progression of diabetes of the $n = 442$ patients. The covariates include age (`age`), sex (`sex`), body mass index (`bmi`), blood pressure (`bp`) and six blood serum measurements (`s1`, `s1`, `s2` `s3` , `s4`, `s5`, `s6`). The data were centered and standardized beforehand.

As $d < n$ we use empirical estimates of CAR scores by setting the parameter `lambda=0` and compute ordinary least squares regression using the function `slm` with the regularization parameters set to zero.

```
car.out = carscore(X,Y, lambda=0)
film.out=slm(X, Y, lambda=0, lambda.var=0)
```

To begin with we order the squared CAR scores and compute the cumulative sum of squared CAR score. First we can validate that the sum of squared CAR scores adds up to the proportion of variance explained or `r.squared` from the standard `lm` command. Moreover, as illustrated in Section **4.4.5** the accumulated CAR score gives an intuitive illustration how much a variable can contribute to the proportion of variance explained. A plot of accumulated squared CAR score is shown on the right in Figure **B.1**. Here, it is evident that the largest shares of the proportion of variance explained are achieved by the first six variables. Then, the gain in proportion of variance explained diminishes and finally stagnates at the level of the overall coefficient of determination.

```
ocar = order(abs(car.out), decreasing=TRUE)
cumsum( car.out[ocar]^2 )
#      bmi        s5        bp        s3        s4        s6
#0.1704344 0.3182504 0.3971500 0.4401129 0.4774330 0.5066587
#      sex       age        s2        s1
#0.5130462 0.5167616 0.5176753 0.5177494

film.out$R2                  # 0.5177494
summary(lm(Y~X))$r.squared   # 0.5177494

plot(1:d, cm$R2, type="p",
  ylab="Cumulative sum of squared CAR scores",
  xlab="Number of included predictors",
  main="Proportion of Explained Variance", ylim=c(0,1))
R2max = max(cm$R2)
lines(1:d, cm$R2, type="l",cex=1.4)
lines(c(1,d), c(R2max, R2max), lty=2)
```

Next we analyze the different approaches to determine the model size. Following Section **5.3.1** we use the Beta distribution with shape parameters $1/2$ and $(n-2)/2$ as distribution under the null hypotheses by `pbeta` to compute the p-values for the ten predictor variables. Here, six variables pass a significance level of $\alpha = 0.05$. Subsequently, we use the intrinsic connection of CAR scores with information criteria (see Section **4.4.5**). Table **4.3** lists the specific cut-offs for the CP rule, RIC, and BIC. Using the C_p / AIC rule on the empirical CAR scores results in 8 included variables, RIC leads to 7 variables, and BIC to the same 6 variables as the computation over the empirical null distribution.

```
pval = 1-pbeta(car.out^2, shape1=1/2, shape2=(n-2)/2)
sum(pval <= 0.05)                            # 6 included variables
# AIC/CP rule
sum(car.out^2 > 2*(1-film.out$R2)/n)         # 8 included variables
# RIC
sum(car.out^2 > 2*log(d)*(1-film.out$R2)/n)  # 7 included variables
# BIC
sum(car.out^2 > log(n)*(1-film.out$R2)/n)    # 6 included variables
```

A particular challenge of the diabetes data set is that it contains two variables (`s1` and `s2`) that are highly positively correlated ($r = 0.897$) but behave in an antagonistic fashion. Specifically, their regression coefficients have the opposite signs so that in prediction the two variables cancel each other out. Figure **B.1** shows all regression models that arise when covariates are added to the model in the order of decreasing variable importance given by $\phi^{\text{CAR}}(X_j)$. As can be seen from this plot, the variables `s1` and `s2` are ranked least important and included only in the two last steps. The plot of the coefficient path in Figure **B.1** is generated by the following commands.

```
car.predlist = make.predlist(ocar, numpred = 1:d, name="CAR")
cm = slm.models(X, Y, car.predlist, lambda=0, lambda.var=0)
bmat = cm$coefficients[,-1]

plot(1:d, bmat[,1], type="l", ylab="Estimated regression coefficients",
  xlab="Number of included predictors",
  xlim=c(1,d+1), ylim=c(min(bmat), max(bmat)))
for (i in 2:d) lines(1:d, bmat[,i], col=i, lty=i)
for (i in 1:d) points(1:d, bmat[,i], col=i)
for (i in 1:d) text(d+0.5, bmat[d,i], cn[i])
```

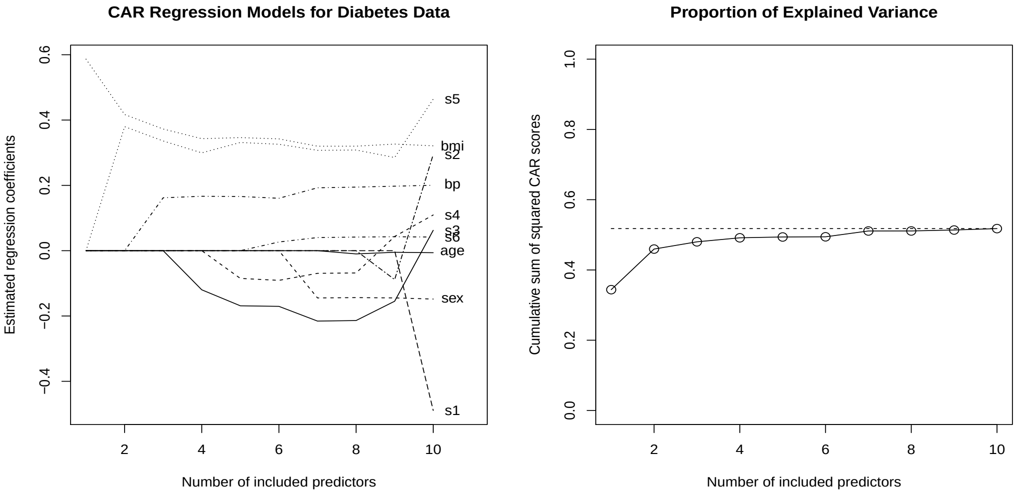


Figure B.1: Estimates of regression coefficients for the diabetes study on the left. Variables are included in the order of empirical squared CAR scores, and the corresponding regression coefficients are estimated by ordinary least squares. The antagonistic correlated variables `s1` and `s2` are included only in the last two steps. Accumulated CAR score on the right.

Appendix C

List of publications underlying this thesis

- **V. Zuber**, P. Duarte Silva, and K. Strimmer. 2012. *A novel algorithm for simultaneous SNP selection in high-dimensional genome-wide association studies.* BMC Bioinformatics 13:284

  Motivation: Identification of causal SNPs in most genome wide association studies relies on approaches that consider each SNP individually. However, there is a strong correlation structure among SNPs that need to be taken into account. Hence, increasingly modern computationally expensive regression methods are employed for SNP selection that consider all markers simultaneously and thus incorporate dependencies among SNPs.

  Results: We develop a novel multivariate algorithm for large scale SNP selection using CAR score regression, a promising new approach for prioritizing biomarkers. Specifically, we propose a computationally efficient procedure for shrinkage estimation of CAR scores from high-dimensional data. Subsequently, we conduct a comprehensive comparison study including five advanced regression approaches (boosting, lasso, NEG, MCP, and CAR score) and a univariate approach (marginal correlation) to determine the effectiveness in finding true causal SNPs.

  Conclusions: Simultaneous SNP selection is a challenging task. We demonstrate that our CAR score-based algorithm consistently outperforms all competing approaches, both uni- and multivariate, in terms of correctly recovered causal SNPs and SNP ranking. An R package implementing the approach as well as R code to reproduce the complete study presented here is available from `http://strimmerlab.org/software/care/`.

- **V. Zuber** and K. Strimmer. 2011. *High-Dimensional Regression and Variable Selection Using CAR Scores.* Statistical Applications in Genetics and Molecular Biology 10: 34

  Variable selection is a difficult problem that is particularly challenging in the analysis of high-dimensional genomic data. Here, we introduce the CAR score, a novel and highly effective criterion for variable ranking in linear regression based on Mahalanobis-decorrelation of the explanatory variables. The CAR score provides a canonical ordering that encourages grouping of correlated predictors and down-weights antagonistic variables. It decomposes the proportion of variance explained and it is an intermediate between marginal correlation and the standardized regression coefficient. As a population quantity, any preferred inference scheme can be applied for its estimation. Using simulations we demonstrate that variable selection by CAR scores is very effective and yields prediction errors and true and false positive rates that compare favorably with modern regression techniques such as elastic net and boosting. We illustrate our approach by analyzing data concerned with diabetes progression and with the effect of aging on gene expression in the human brain. The R package "care" implementing CAR score regression is available from CRAN.

- **V. Zuber** and K. Strimmer. 2009. Correlation-adjusted t-scores in application to functional magnetic resonance imaging data. Proceedings of the 6th International Workshop on Computational Systems Biology, WCSB 2009 (June 10-12, 2009, Aarhus, Denmark). pp. 163-166.

  The correlation-adjusted t-score (CAT score) is a modification of the ordinary t-statistic to account for dependencies among variables. Recently, we have shown (Zuber and Strimmer 2009) that the CAT score improves ranking of genes to detect differential expression in the presence of correlation. Noting the similarity of structure between high-dimensional gene expression and image analysis data, here we apply the CAT score using shrinkage estimation to functional magnetic resonance imaging (fMRI) data. We show that the cat score is a simple, yet effective, means to accommodate correlation among voxels and to improve standard t-type tests of neural activation.

- **V. Zuber** and K. Strimmer. 2009. *Gene ranking and biomarker discovery under correlation.* Bioinformatics 25 (20): 2700-2707

  Motivation: Biomarker discovery and gene ranking is a standard task in genomic high throughput analysis. Typically, the ordering of markers is based on a stabilized variant of the *t*-score, such as the moderated *t* or the SAM statistic. However, these procedures ignore gene-gene correlations, which may have a profound impact on the gene orderings and on the power of the subsequent tests.

  Results: We propose a simple procedure that adjusts gene-wise *t*-statistics to take account of correlations among genes. The resulting correlation-adjusted *t*-scores ("cat" scores) are derived from a predictive perspective, i.e. as a score for variable selection to discriminate group membership in two-class linear discriminant analysis. In the absence of correlation the cat score reduces to the standard *t*-score. Moreover, using the cat score it is straightforward to evaluate groups of features (i.e. gene sets). For computation of the cat score from small sample data we propose a shrinkage procedure. In a comparative study comprising six different synthetic and empirical correlation structures we show that the cat score improves estimation of gene orderings and leads to higher power for fixed true discovery rate, and vice versa. Finally, we also illustrate the cat score by analyzing metabolomic data.

  Availability: The shrinkage cat score is implemented in the R package "st" available from URL `http://cran.r-project.org/web/packages/st/`.