Using Survival to Predict Survival

Goeman, J.J.

Statistical Methods for Microarray Data

THOMAS STIELTJES INSTITUTE
FOR MATHEMATICS

De uitgave van dit proefschrift werd ondersteund door het Fonds Medische Statistiek en door het Thomas Stieltjes Institute for Mathematics.

ISBN-10: 90-9020372-9
ISBN-13: 978-90-9020372-0

Statistical Methods for Microarray Data

Pathway Analysis, Prediction Methods and Visualization Tools

PROEFSCHRIFT

ter verkrijging van de graad van Doctor
aan de Universiteit Leiden,
op gezag van de Rector Magnificus Dr. D. D. Breimer,
hoogleraar in de faculteit der Wiskunde en
Natuurwetenschappen en die der Geneeskunde,
volgens besluit van het College voor Promoties
te verdedigen op woensdag 8 maart 2006
te klokke 15.15 uur

door

Jelle Jurjen Goeman

geboren te Leiderdorp in 1976

PROMOTIECOMMISSIE

PROMOTORES:	Prof. dr. J. C. van Houwelingen Prof. dr. S. A. van de Geer · *Eidgenössische Technische Hochschule, Zürich*
REFERENT:	Prof. dr. S. Richardson · *Imperial College, Londen*
OVERIGE LEDEN:	Prof. dr. C. Kooperberg · *Fred Hutchinson Cancer Research Center, Seattle* Prof. dr. G. J. B. van Ommen Dr. E. W. van Zwet

Contents

Published and submitted chapters

The folowing chapters have been published in scientific journals or have been submitted for publication:

Chapter 2:
J. J. Goeman, S. A. van de Geer, F. de Kort, and J. C. van Houwelingen (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* **20** (1), 93–99.

Chapter 3:
J. J. Goeman, J. Oosting, A. M. Cleton-Jansen, J. Anninga, and J. C. van Houwelingen (2005). Testing association of a pathway with survival using gene expression data. *Bioinformatics* **21** (9), 1950–1957.

Chapter 4:
J. J. Goeman and S. le Cessie. A goodness-of-fit test for multinomial logistic regression. submitted.

Chapter 5:
J. J. Goeman, S. A. van de Geer and J. C. van Houwelingen (2006) Testing against a high-dimensional alternative. *Journal of the Royal Statistical Society, Series B* **68**, in press.

Chapter 6:
J. J. Goeman and J. C. van Houwelingen. Model-based dimension reduction for high-dimensional regression. submitted.

Chapter 7:
P. H. C. Eilers and J. J. Goeman (2004). Enhancing scatterplots with smoothed densities. *Bioinformatics* **20** (5), 623–628.

Appendix:
J. J. Goeman and J. Oosting (2005). Globaltest: testing association of a group of genes with a clinical variable. R package, version 3.2.0. *www.bioconductor.org*.

CHAPTER 1

Introduction and overview

The subject of this thesis is the statistical analysis of high-dimensional data. It is motivated by (and primarily focussed on) problems arising from microarray gene expression data, a new type of high-dimensional data, which has become important in many areas of biology and medicine in the last decade.

The thesis is a collection of six articles and a software manual. The articles are self-contained and they can in principle be read in any order. However, such random reading of the chapters would not do justice to the close connections that exist between them, which are partly obscured by the fact that the articles were written for different journals and therefore for different audiences with different backgrounds and interests.

The objective of this introduction is to provide the context in which the papers should be read and to make the connections between the different chapters more explicit. It is not meant to be a full review of microarray data and their analysis. These can be found for example in Díaz-Uriarte (2005), Speed (2003) and Simon et al. (2003). I give a short introduction to gene expression data and the biological and clinical questions arising from them in section 1.1. The next section 1.2 reviews some of the statistical methods that have been developed in recent years to address these questions. Section 1.3 examines the contribution of this thesis to the field.

1.1 Biological context

The microarray is a recent technology from molecular biology which was first developed around 1995 (Schena et al., 1995), and became widely available around the turn of the century (see Ewis et al., 2005, for a short history). The microarray is designed to measure the activity of gene expression, which is the process by which the genetic information in DNA is used to make proteins. The microarray technology gives the biologist the potential to greatly increase knowledge on the functions of genes and on the biology of disease, as well as allowing improvements in diagnosis and prognosis of patients.

Technology

The central dogma of molecular biology says that the information encoded in the DNA is first transcribed into multiple copies of RNA, which is then translated into a protein. Graphically:

$$\text{DNA} \rightarrow \text{RNA} \rightarrow \text{Protein}$$

The genetic information in the DNA is generally the same in all cells of the same individual and also for the major part between individuals of the same species. However, the need for proteins varies depending on the type of tissue the cell belongs to and depending on the condition of that tissue. Therefore the rate of transcription of each gene also varies and so does the concentration of RNA.

A microarray is a high-throughput measurement device that can simultaneously measure RNA concentrations of tens of thousands of genes in a single biological sample (tissue or cell line). It consists of a small glass slide on which there are fixed spots, at least one for each gene, consisting of single-strand DNA from part of the gene. When using the microarray to measure gene expression in a tissue sample, one extracts the RNA from the tissue, labels it with fluorescent dye, and brings it in contact with the glass slide. By the properties of RNA and DNA, the RNA mostly binds to the DNA of the gene it belongs to. Therefore, the RNA concentrations of different genes can be compared by comparing the colour intensity of the spots after the experiment.

Colour intensities for each gene are read off by a laser scanner using specialized scanning software for microarrays. This produces a single quantitative measurement per spot on the array, which can be used for statistical analysis. The result of the experiment is a measure of the RNA concentration of the genes spotted on the microarray.

These measurements are only indirect, because the colour intensity does not have a simple and immediate relationship to RNA concentration. Furthermore, the measures are only relative. As there are factors in each microarray slide and each tissue sample that cannot be completely controlled, the relationship between colour intensity and RNA concentration is different for every microarray slide. Therefore, direct comparison of measured values is only possible between spots on the same slide. To make different slides comparable, the slides have to be 'normalized'. The process of normalization also includes removing systematic biasses and artifacts of the technology and possibly removing genes whose signal is considered unreliable. The choice of a technique for normalization depends very much on the precise microarray technology used, and many different methods are in use (Bolstad et al., 2003; Huber et al., 2002; Irizarry et al., 2003; Kerr et al., 2000; Wu et al., 2004; Yang et al., 2002).

Microarray technology is valuable because it can generate large amounts of useful data, but the measurements are not extremely accurate. Individual measurements of the gene expression level of a specific gene in a specific tissue can be noisy and the arrays suffer from many systematic errors, which have to be carefully filtered out before the analysis can start. Accuracy has improved, however, over the few years that the technology has matured.

Research questions

The vast amount of data that can be generated using microarray technology has proved very attractive to researchers, leading to an explosion of publications using microarray technology (Ewis et al., 2005). Microarray data are used in very diverse areas of biology and medicine to answer very diverse research questions. Most research tries to find relationships between gene expression measurements and external data. These external data may be different experimental conditions if the microarrays were done on cell lines or tissue from model organisms. More often they are phenotypic data, especially when the microarrays were done on tissue samples of patients (e.g. dissected tumours). Research questions can then be loosely divided into three classes by the three basic ingredients of a microarray experiment: research questions can be about genes, about patients or about a phenotype.

Studies with research questions relating to genes are most common in microarray research in biology. These studies aim at increasing the understanding of the function of genes. This is usually done by searching for genes whose expression is correlated to the experimental conditions or to important phenotypes.

Clinically motivated microarray studies often have research questions relating to patients. Such studies aim at improving diagnosis and prognosis of patients, for which microarrays are expected to have great potential (Golub et al., 1999; Van 't Veer et al., 2002). Microarrays can be especially valuable in prognosis, as future events such as metastasis which are not yet clinically detectable may already be detectable in the gene expression activity. Microarrays can also be used in diagnosis to replace older methods which are more costly or more damaging to the patient. Patient-oriented studies usually try to find a prediction rule for predicting patient phenotypes from the microarray data.

A third kind of studies has research questions relating to the phenotype studied, which is usually some aspect of disease. These studies aim at increasing the understanding the biology of disease. They try to find out which biological processes are related to disease or to certain aspects of a disease. This information can be used to unravel the biological mechanisms involved in the disease. This type of phenotype-oriented research can be seen as the inverse of

gene-oriented research: instead of inferring information on gene function from knowledge about the phenotype, information about the phenotype is inferred from knowledge about gene functions. This kind of research relies heavily on gene annotation, which is used to link genes to biological processes and other gene functions.

These three types of research questions are not so neatly distinguished in practice. Many actual experiments are set up with a mixture of research aims, which may be of all three kinds. The research aims are also strongly related. Knowing which genes are related to a phenotype can be a step in the direction of predicting the phenotype and also, by studying known functions of the genes, a step towards knowing which biological processes are involved. It is important, however, to distinguish the three types of research questions, as they are different on a fundamental level and require different statistical methods to solve.

Annotation and pathways

The vast amount of data generated by microarray experiments is not only challenging from a statistical point of view. It is also a challenge for the biologist to interpret the results and to compare the results of his or her experiments to the literature and to the results of earlier experiments. An important aid for interpretation is given by the annotation tools that are available in the world of bioinformatics. Annotation tools link genes to knowledge that has so far been accumulated about that gene. I will mention a few which are important for this thesis.

The most widely used gene annotation system is the Gene Ontology (GO, Ashburner et al., 2000, www.geneontology.org). GO is a structured vocabulary that can be used to systematically describe genes, gene products, biological processes, cellular components and molecular functions. The GO ontology can be used as a tool for annotation, because it includes a database that links genes to terms from the ontology. Evidence for these links is collected in various ways. The database is generated by automated scanning of the literature, but part of the database is managed by experts. GO is by far the largest annotation tool, containing over 18,000 annotation terms.

Alternatives to GO tend to be smaller but more reliable, as they are usually not automated but fully managed by experts. The Kyoto Encyclopedia of Genes and Genomes (KEGG, Ogata et al., 1999, www.genome.jp/kegg) is a database of 267 pathways (July 2005). A pathway is a set of genes related to a specific function or biochemical process. The pathways in the KEGG database are predominantly metabolic pathways. They include diagrams of the relationships between genes in the pathway. The KEGG pathways are created and kept up-

to-date by experts.

1.2 Statistical context

It has taken some time before biologists and bioinformaticians working with microarray data realized that they needed statistical methods (Vingron, 2001) and it has taken some more time before statisticians responded to the call for good methodology. Since then, however, many interesting new statistical methods have been developed to answer the various research questions mentioned above.

Cluster analysis

The first statistical method to become popular in microarray research was hierarchical cluster analysis (Eisen et al., 1998). Cluster analysis is a visualization tool which allows biologists to inspect the results of the experiment by dividing the samples and the genes into 'clusters' which have similar expression patterns. Cluster analysis is sometimes also used to infer that patients in the same cluster have the same subtype of a disease or to infer that genes in the same cluster have the same function.

Cluster analysis and its associated heat map display are very useful for visualization. They provide a well-ordered display of data which are otherwise very difficult to survey. However, hierarchical cluster analysis has two important drawbacks when used as a method of statistical analysis. In the first place it is unsupervised, i.e. it does not make use of the phenotype information of the experiment. It cannot, therefore, answer any of the research questions listed above, which are all related to a phenotype. Secondly, it is very weak as a tool for inference. Hierarchical cluster analysis is not based on any model and has no statistical or probabilistic motivation (although model-based alternatives have been developed: McLachlan et al., 2002). There are no accepted estimation procedures which can be used to determine the number of clusters and there are no accepted testing procedures which allow one to show that there is actually more than one cluster. Because of these drawbacks, the popularity of cluster analysis as the primary analysis tool is slowly declining in microarray research.

Finding differentially expressed genes

Finding genes that are associated with a phenotype is the most common goal of a microarray experiment. It is usually solved by multiple hypothesis testing: testing association of the gene expression measurements with the phenotype separately for each gene. If the phenotype takes two values, the classical so-

lution to this multiple testing problem would be to do tens of thousands of t-tests, one for each gene, and to correct for multiple testing using Bonferroni. This classical setup has to be amended for the microarray context in several ways.

The Bonferroni correction for multiple testing controls the family-wise error rate, which is the probability of making at least one false rejection, or 'false discovery'. The 2×2 table in table 1.1 shows the possible outcome of a multiple testing procedure. Of m null hypotheses, an unknown number m_0 are true. Of

TABLE 1.1: *Two-by-two table showing the outcome of a multiple testing procedure of m null hypotheses of which m_0 are true.*

	not significant	significant	total
true null	U	V	m_0
false null	T	S	$m - m_0$
total	$m - R$	R	m

these true null hypotheses V are rejected, while $U = m_0 - V$ are not. Similarly of the $m - m_0$ false null hypotheses, T are rejected, while S are not. The family-wise error, that the Bonferroni procedure controls, is $\mathrm{P}(V \geq 1)$.

The Bonferroni criterion for the family-wise error criterion is considered too conservative for microarray data analysis. One reason for this conservatism is that the Bonferroni procedure is not efficient, especially because it does not take into account dependency between the test statistics. A review of many proposed improvements is given by Dudoit et al. (2003), who argue for the resampling-based procedure of Westfall and Young (1989) to control the family-wise error rate, which takes into account the dependency structure between the test statistics. A second reason why the Bonferroni adjustment is felt to be too conservative is because the family-wise error criterion is considered too strict for use in microarray data analysis. Microarray research that aims at finding genes associated with a certain phenotype is often largely exploratory. The experiments are meant to generate hypotheses that can later be tested using more accurate conventional biological techniques. It is therefore not so important that every discovery is absolutely reliable, but it is more important that a good proportion of the discoveries can be trusted.

For this reason one often does not control the family-wise error rate in microarray research, but the false discovery rate (FDR). This concept was formulated by Benjamini and Hochberg (1995) as the expected proportion of false discoveries V among the discoveries $R = V + S$. It can be interpreted as the proportion of genes in the list of declared significant genes that is reliable. This

proportion is taken to be zero if V and R are both zero. Benjamini and Hochberg (1995) also provide a procedure that controls the FDR in a multiple testing procedure if all test statistics are independent, as well as when they are positively correlated (Benjamini and Yekutieli, 2001). Reiner et al. (2003) provide resampling-based methods to control the FDR more efficiently when the test statistics are dependent. Other authors prefer to estimate the proportion of false discoveries V/R instead of controlling its expectation a priori (Efron et al., 2001; Storey, 2002; Storey and Tibshirani, 2003; Tusher et al., 2001). The different approaches to the false discovery rate are compared in detail by Ge et al. (2003).

When doing multiple testing, some power can be gained by recognizing that the thousands of tests of a single microarray experiment are in fact highly comparable. Therefore, information from all other genes can be used to improve the power of each individual test. In a multiple t-test procedure, for example, the genes with the smallest estimated standard errors will probably have underestimated standard errors; any significant t-statistics that result are, therefore, likely to be false positives. Shrinking the estimated standard errors towards each other can prevent such false positives and therefore gain power. A primitive implementation of this idea can be found in the popular method SAM (Tusher et al., 2001). More sophisticated methods use empirical Bayes methods to shrink the estimates of quantities for different genes toward each other in order to 'borrow strength' across the genes (De Menezes et al., 2004; Smyth, 2004; Wright and Simon, 2003).

Prediction methods

Prediction methods aim at predicting a patient phenotype (e.g. a class membership or a survival time) from the gene expression measurements. The primary goal is to let the resulting prediction rule help diagnosis or prognosis of patients. However, there is often a secondary goal of interpreting the resulting prediction rule in terms of the genes involved and/or their function. Prediction methods, which are designed to answer the patient-related research questions described in section 1.1, are typically also used to answer gene- or phenotype-related research questions.

Prediction methods are perhaps the most active area of statistical microarray research, as the inability of classical statistics to handle high-dimensional data is most clearly visible in the prediction context. Many new statistical methods have been proposed (see Hastie et al., 2001, for an overview). Some of these methods have their origins in statistics, usually from methods designed to deal with collinearity in regression, such as principal components regression (Jolliffe, 2002), Ridge Regression (Hoerl and Kennard, 1970), the LASSO

(Tibshirani, 1996). Other methods are taken from chemometrics, where high-dimensional prediction problems are well-known, e.g. in spectroscopy (see Brown, 1993, for an overview). Most notable among the methods from chemometrics is Partial Least Squares (Wold et al., 1984), but also, more recently, the model-based maximum likelihood method of Burnham et al. (1999b). Many other methods have come from machine learning, which is also a very active field that develops methods for microarray research and slowly becomes more integrated with statistical methodology. Important methods from machine learning include *k*-nearest neighbours and support vector machines (Hastie et al., 2001). Unlike the methods taken from statistics, which are typically formulated in a regression context, methods from machine learning are typically exclusively useable in classification problems.

The main problem that prediction methods for high-dimensional data have to cope with is overfit, because the number of parameters of most prediction methods grows with the number of predictors. There are various strategies to reduce this overfit. Each of these strategies effectively reduces the parameter space to reduce the possibility of overfit. This decreases the prediction variance at the cost of introducing bias.

One important strategy for reducing overfit is shrinkage. Shrinkage methods in regression reduce the estimated regression coefficients toward zero. A typical shrinkage method is Ridge Regression. This has been applied on microarray data for example by Eilers et al. (2001) for the classification problem and by Pawitan et al. (2004) and Van Houwelingen et al. (2005) for predicting survival. Ridge regression shrinks all regression coefficients towards zero, without making them vanish. An alternative shrinkage method is the LASSO (Tibshirani, 1996) and its generalization Least Angle Regression (Efron et al., 2004), which sets most of the regression coefficients to zero. This has the additional advantage of leading to a sparse predictor. The LASSO has been applied in microarray data by Shevade and Keerthi (2003). All these shrinkage methods can be described as empirical Bayesian models (Van Houwelingen, 2001). A LASSO-like shrinkage can also be applied in discriminant analysis, as in the popular Nearest Shrunken Centroids method (also known as PAM), which shrinks the centroids of gene expression in each class towards the overall centroid (Tibshirani et al., 2002).

A second strategy for reducing overfit is to use dimension reduction methods such as principal components or Partial Least Squares, which reduce the gene expression measurements to a small number of orthogonal linear combinations, which are subsequently used to predict the phenotype. Partial Least Squares (PLS) is the standard method for high-dimensional prediction problems in chemometrics and has quite quickly found its way into microarray

data analysis. PLS has been applied on microarray data by Nguyen and Rocke (2002a,b). A model-based variant of PLS by Burnham et al. (1999b) has been applied by Tan et al. (2005). Bair et al. (2004) proposed 'Supervised Principal Components': principal components regression after a pre-selection of genes based on their association with the phenotype.

A third strategy for reducing overfit is variable selection. Pure variable selection is not so popular in microarray data analysis, but in combination with other methods such as shrinkage or dimension reduction it is very popular. Methods such as the LASSO, Supervised Principal Components and Nearest Shrunken Centroids all return a sparse prediction rule. For a major part, the popularity of sparse prediction rules stems from the biological belief that most of the genes have no predictive value for the phenotype, so that a good prediction rule is expected to be sparse.

Another reason why there is a desire for a sparse prediction rule, is for the secondary aim of prediction modelling: interpretation. Prediction rules with thousands of regression coefficients are very difficult to interpret, while a sparse prediction rule is relatively easily given a causal interpretation. However, it is highly dangerous to give too much causal interpretation to the genes selected in a resulting prediction rule, as the set of selected genes is often highly variable (Ein-Dor et al., 2005). This is a general problem in interpretation of high-dimensional prediction rules: there are invariably many prediction rules which are very different but still have about the same quality of prediction.

Analysis of pathways

The goal of phenotype-related research is to infer the mechanisms underlying disease from knowledge about the function of genes whose expression is associated with the phenotype. For example, if the genes which are known to be involved in programmed cell death tend to be differentially expressed between metastasizing tumours and non-metastasizing tumours, one can infer that the mechanism of metastasis is related to the mechanism of programmed cell death. Methods that address such phenotype-related research questions were slow to develop (Díaz-Uriarte, 2005). One reason for this is that the type of research question does not have a direct similarity to research questions that statisticians are familiar with.

Phenotype-related research questions usually focus on pathways, or, more generally, on (GO) annotation terms. The question is to test the hypothesis that the expression pattern of a certain pathway is associated with the phenotype. It can also be more exploratory, asking which pathways (out of a library of pathways or annotation terms) are associated with the phenotype.

The analysis of such questions is usually performed as a second step after

finding single genes which are associated with the phenotype. Once such a list is obtained, the researcher searches for significant overlap between the list of significant genes and a list of genes belonging to a certain pathway. Several authors have described 'enriched gene set' methods to assess whether an overlap is significant (Al-Shahrour et al., 2004; Beissbarth and Speed, 2004; Boyle et al., 2004; Smid and Dorssers, 2004; Zeeberg et al., 2003; Zhang et al., 2004). These methods create a 2×2 table for each pathway as shown in Table 1.2. Based on this table, one tests for independence of "being significant" and "being in the pathway", typically using Fisher's exact test because the expected count in the upper left cell is usually very small. If the test is significant, it means that, in this experiment, the genes in the pathway have a different (usually higher) probability of being significant than the genes that are not in the pathway.

TABLE 1.2: *Two-by-two table for enriched gene set analysis. The table is filled with counts of the number genes on the microarray chip based on whether they come out as significantly associated with the phenotype and whether they belong to a certain pathway.*

	significant	non-significant
in pathway		
not in pathway		

Enriched gene set analysis is a simple, elegant and useful procedure. However, it is easy to misinterpret the results. The p-values that come out of the method are with respect to the experiment of drawing a gene at random from the genes on the microarray *given the data*. The p-value does therefore not say anything directly about whether a result can be expected to be replicated in future experiments, as an ordinary p-value would. The enriched gene set procedure is therefore not a method that generalizes testing whether a gene is associated with the phenotype to testing whether the pathway is associated with the phenotype.

Some interesting variants of the enriched gene set method are given by Sohler et al. (2004), who make explicit use of network structures between genes to find interesting subnetworks. Mootha et al. (2003) do not use a cut-off between significant and non-significant, but use the p-values as a ranking. They look for enriched gene sets by testing for departure from a uniform distribution with a Kolmogorov-Smirnov statistic.

A very different method for testing whether a gene set is associated with a phenotype was recently proposed by Mansmann and Meister (2005). Unlike the enriched gene set methods, this method is in line with classical statistical practice. It is based on a ANCOVA model and can be used for a phenotype that

takes two values. It tests whether the joint distribution of the gene expression measurements is the same for both values of the phenotype. Their approach is closely related to the GlobalTest method proposed in Goeman et al. (2005, 2004) and in this thesis (Chapters 2 and 3), although the analysis is quite different. The model of Mansmann and Meister considers the distribution of the gene expressions given the phenotype, where Goeman et al. study the distribution of the phenotype given the gene expression. Mansmann and Meister show that under some conditions their ANCOVA approach has more power than the GlobalTest, but more research is needed to compare the two methods.

The BioConductor project

To make the new statistical methods for normalization and analysis available to statisticians, biologists and computer scientists, a software project called BioConductor (www.bioconductor.org) has been set up by Gentleman et al. (2004). BioConductor is a collection of software packages written for R (www.r-project.org), a general language and environment for statistical computing (R Development Core Team, 2005), similar to S.

Since its start in 2001, BioConductor has quickly grown with the addition of numerous packages from the statistical and bioinformatics community. Part of the popularity of BioConductor can be explained by the fact that both BioConductor and R are free and open-source distributions. BioConductor nowadays has an enormous impact on the way microarray data are analyzed. Only very few current statistical methods for microarray data analysis are not available on BioConductor.

1.3 This thesis

This thesis consists of three parts. The first part is the most important one and consists of Chapters 2 up to 5 and the appendix. It develops a new way of answering phenotype-related questions, which is very different from the enriched gene set methods. The second part consists of Chapter 6, which presents a model-based way to motivate dimension reduction techniques for prediction methods. Finally, the third part, Chapter 7, addresses the important subject of visualization of microarray data. The contributions of the three parts to the field of microarray data analysis will be briefly reviewed below.

Global Testing of pathways

The first part of this thesis explores how best to answer the phenotype-related research questions that try to find relationships between a phenotype and known pathways or gene annotation terms. It presents various aspects the

GlobalTest methodology designed to answer such questions (see Díaz-Uriarte, 2005; Mansmann and Meister, 2005, for reviews).

Just like in the enriched gene set method, we define a pathway in the simplest possible way as any pre-defined set of genes. This has the additional advantage that the resulting methods for the analysis of pathways are quite generally applicable. They can also be used to test association between a phenotype and a set of genes with the same chromosomal location or with a set of genes that have been marked as interesting by another experimenter.

Otherwise, the approach underlying the GlobalTest method is fundamentally different from the approach of the enriched gene set methods described above. The test we present generalizes testing for association of a single gene with a phenotype (e.g. a t-test) to testing for association of a set of genes with a phenotype. Unlike the enriched gene set methods, this test is based on a classical statistical model in which patients are the units of observation, not the genes. The p-values that come out of the test have the regular statistical interpretation.

Although a method for testing, the GlobalTest is closely related to prediction methods. The basic idea behind the method is very simple. The method is based on an empirical Bayesian regression model for predicting the phenotype from the gene expression measurements of the genes in the pathway. This is the same type of model that can be used to motivate prediction methods like ridge regression or the LASSO. If the pathway is associated with the phenotype, then the gene expression measurements of the pathway should give some information for predicting the phenotype. To test for this association, the GlobalTest therefore tests whether the gene expression measurements have any predictive potential for predicting the phenotype. The details of this testing methodology and its application to microarray data are worked out in Chapters 2 using the linear and logistic regression models and in Chapter 3 using the Cox proportional hazards model for censored survival data.

It is shown in these chapters that the resulting test is mathematically equivalent to a goodness-of-fit test for regression models. Part of the technical details of Chapter 2 therefore rely on the goodness-of-fit test developed by Le Cessie and van Houwelingen (1995) and its improvements by Houwing-Duistermaat et al. (1995). Similarly, the test of Chapter 3 makes use of the goodness-of-fit test for the Cox model by Verweij et al. (1998). Many phenotypes of interest in microarray data are multi-category of nominal scale, however, and would have to be modelled using a prediction model based on the multinomial logistic regression model. As there was no goodness-of-fit test available that could be used in the same way as Le Cessie and van Houwelingen (1995), such a goodness-of-fit test for multinomial logistic regression has been developed in Chapter 4.

Finally, Chapter 5 views all the tests of chapters 2, 3 and 4 in a more abstract way as examples of a general type of locally most powerful test. It creates a general framework for this type of test, which tests a simple null hypothesis against a high-dimensional alternative. The power properties of this type of test are investigated from a purely frequentist point of view.

Model-based Prediction

In Chapter 6 we consider the problem of predicting the phenotype of patients. As shown in Section 1.2 there are many different prediction methods available. It is a big practical problem which method to choose for the analysis of a specific data set.

Ideally, the choice of the prediction method should make use of knowledge about the data. Each of the methods has restrictions on the parameter space that reduce overfit, but introduce bias. Whether this bias is serious or not, depends on the nature of the data. For example, a LASSO that sets most regression coefficients to zero can be expected to do especially well when most of the genes are not associated with the phenotype. However, problems arise when trying to use knowledge of the data when choosing a prediction method, because most of these methods are not model-based. Therefore, the type of data for which they perform well is not well-defined.

Chapter 6 attempts to fill this gap by approaching dimension reduction in a model-based way. It builds a model of the joint distribution of the phenotype and the gene expression values, in which their distribution depends on a set of latent variables. It will be shown that the assumption of such a model leads to a method similar to Supervised Principal Components (Bair et al., 2004) in a natural way.

Smooth visualization

The sheer quantity of microarray data poses problems by itself. In smaller data sets, researchers are easily able to inspect the data, observing interesting patterns, searching for outlying observations and formulating hypotheses. All these things are not possible in a data matrix resulting from a microarray experiment, which easily contains a million entries. However, careful inspection of the data is all the more important in microarray data, as disturbing outliers occur frequently and hypothesis generation is a primary goal of many experiments. Good visualization is therefore felt to be important, as can be seen from the immense popularity of cluster analysis, which is much more an exploratory visualization tool than an inferential statistical method.

Chapter 7 of this thesis presents a tool for better visualization of scatterplots of thousands of dots. Drawn in the conventional way, such plots tend to give

a distorted impression of the true density of points. To amend this, Chapter 7 proposes a simple way of representing the density of the dots as a colour representation, constructed by smoothing a two-dimensional histogram. An important advantage of the smoothed histogram method to calculate the density over other, more sophisticated density estimates is that it can be calculated very fast, even for millions of dots.

CHAPTER 2

Testing Association of a Pathway with a Clinical Variable

Abstract

This paper presents a global test to be used for the analysis of microarray data. Using this test it can be determined whether the global expression pattern of a group of genes is significantly related to some clinical outcome of interest. Groups of genes may be any size from a single gene to all genes on the chip (e.g. known pathways, specific areas of the genome or clusters from a cluster analysis). The test allows groups of genes of different size to be compared, because the test gives one p-value for the group, not a p-value for each gene. Researchers can use the test to investigate hypotheses based on theory or past research or to mine gene ontology databases for interesting pathways. Multiple testing problems do not occur unless many groups are tested. Special attention is given to visualizations of the test result, focussing on the associations between samples and showing the impact of individual genes on the test result. An R-package GlobalTest is available from http://www.bioconductor.org.

2.1 Introduction

The popularity of microarray technology has led to a surge of new statistical methods aimed at finding differentially expressed genes. A sophisticated methodology has been developed to counter the multiple testing problem that occurs when testing thousands of genes simultaneously (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001; Dudoit et al., 2003; Storey, 2002; Tusher et al., 2001).

This is a pre-copy-editing, author-produced version of an article accepted for publication in *Bioinformatics* following peer review. The definitive publisher-authenticated version: J. J. Goeman, S. A. van de Geer, F. de Kort, and J. C. van Houwelingen (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics 20*(1), 93–99 is available online at: `http://dx.doi.org/10.1093/bioinformatics/btg382`

This paper looks at the problem of finding differentially expressed genes from a different point of view. It presents a global test that can be used to determine whether some pre-specified *group of genes* is differentially expressed. This allows the unit of analysis to be shifted from individual genes to groupings of genes. The question addressed is whether the gene expression pattern over the whole group of genes is related to a clinical outcome. It does not matter for the test whether the group consists of up- or downregulated genes or is a mixture of both. The clinical outcome may be a group label or a continuous measurement.

Often researchers who conduct microarray experiments have one or more specific groups of genes that they are especially interested in, e.g. certain pathways or areas of the genome. Even if this is not the case, many pathways are at least partially known from the scientific literature and it could sometimes be more worthwhile to test a limited number of pathways or gene ontology classes than an enormous number of individual genes. Other potentially interesting groups of genes to be tested include the clusters from a cluster analysis or all genes on the chip.

The first part of the paper presents the mathematical details, starting with the empirical Bayesian generalized linear model on which the test is based. Connections to other methods (especially prediction methods) are elaborated.

In the second part two elaborate applications are presented, showing different aspects of the test. One is the well-known public dataset by Golub et al. (1999) with Affymetrix arrays of patients with Acute Lymphoic Leukemia (ALL) and Acute Myeloid Leukemia (AML). Here the test is applied to the set of all genes to show an enormous difference in overall expression pattern. The second is a smaller in-house dataset with oligonucleotide arrays of cell lines, of which some were exposed to a heat shock. The test is applied to two groups of genes associated with heat shock.

In the applications, special attention is given to visualizations of the test result which make the results easier to interpret for the researcher. These include graphs to search for outlying samples and diagnostic plots to judge how much each individual gene contributes to a significant test result for the group.

2.2 The data

Proper normalization is very important for a meaningful analysis of microarray data. The problem of normalization generates an enormous amount of literature (e.g. Dudoit et al., 2002; Huber et al., 2002; Kerr et al., 2000) and is fast becoming a statistical specialization by itself. In this paper we will simply assume that the data have been normalized beforehand in a way that fits the ex-

perimental design and that possible confounding effects of array, dye etc. have been removed as well as the experimental design allows. However, missing values are allowed (see section 2.8).

We assume we have normalized gene expression measurements of n samples for p genes. Of these p genes, there is a subgroup of m ($1 \leq m \leq p$) genes, which we want to test. It is important that the clinical outcome was not used in the selection of these m genes. Define $X = (x_{ij})$ as the $n \times m$ data matrix containing only of the m genes of interest. Note that we follow the statistical convention to use the rows for the samples and the columns for the genes, instead of the transposed notation which is common in microarray literature. Define Y as the clinical outcome (an $n \times 1$ vector). Usually Y will be a $0/1$ group label (e.g. AML vs. ALL), but it may also be a continuous measurement.

2.3 The model

There is a close connection between finding differentially expressed genes and predicting the clinical outcome. If a group of genes can be used to predict the clinical outcome, the gene expression patterns must differ for different clinical outcomes. This duality will be used to derive the test.

Modelling the way in which Y depends on X, we adopt the framework of the generalized linear model (McCullagh and Nelder, 1989), which includes linear regression and logistic regression as special cases. In this model there is an intercept α, a length p vector of regression coefficients $\boldsymbol{\beta}$ and a link function h (e.g. the logit function), such that

$$E(Y_i|\boldsymbol{\beta}) = h^{-1}\big(\alpha + \sum_{j=1}^{m} x_{ij}\beta_j\big). \tag{2.1}$$

Here β_j is the regression coefficient for gene j ($j = 1, \ldots, m$).

Testing whether there is a predictive effect of the gene expressions on the clinical outcome is equivalent to testing the hypothesis

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_m = 0,$$

that all regression coefficients are zero. It is not possible to test this hypothesis in a classical way (with $\boldsymbol{\beta}$ non-stochastic) because m may be large relative to n. In this case there are too few degrees of freedom.

However, it is possible to test H_0 if it is assumed that $\beta_1, \ldots, \beta_m$ are a sample from some common distribution with expectation zero and variance τ^2. Then a single unknown parameter τ^2 determines how much the regression coefficients are allowed to deviate from zero. The null hypothesis becomes simply

$$H_0 : \tau^2 = 0.$$

Note that the choice of $\tau^2 I_m$ (with I_m the $m \times m$ identity matrix) as the covariance matrix of the stochastic vector $\boldsymbol{\beta}$ is not imperative. It is the most convenient choice which will yield a test that treats all genes on an equal footing. Any other $m \times m$ covariance matrix may be used to replace I_m, if desired, resulting in a different test with power against different alternatives. For example a different diagonal matrix can be taken to reflect prior beliefs in the greater reliability of certain genes. Assuming positive correlations between the elements of $\boldsymbol{\beta}$ results in more power against alternatives where they all coefficients of $\boldsymbol{\beta}$ have the same sign.

The model (2.1) with $\boldsymbol{\beta}$ random may be looked at in various ways. Firstly, the distribution of $\boldsymbol{\beta}$ can be seen as a prior, with unknown shape and with a variance depending on an unknown parameter. Viewed in this way the model (2.1) is an empirical Bayesian model.

A second interpretation is to view the model as a penalized regression model, in which the estimated coefficients are shrunk towards a common mean. The loglikelihood of Y can be written

$$\text{loglik}(Y, \boldsymbol{\beta}) = \text{loglik}(Y|\boldsymbol{\beta}) + \text{loglik}(\boldsymbol{\beta}),$$

where the first term on the right is the likelihood of the ordinary generalized linear model and the second term is known as the *penalty*. Well-known examples of penalized regression include ridge regression (Hoerl and Kennard, 1970), which arises when $\boldsymbol{\beta}$ is normally distributed and the LASSO (Tibshirani, 1996), which is a variant where $\boldsymbol{\beta}$ has a double exponential distribution. Ridge regression with a logistic link function has been described by Le Cessie and van Houwelingen (1992) and applied on microarray data by Eilers et al. (2001) with promising results.

There is a third interpretation which will be the basis for the test in the next section. For this we write $r_i = \sum_j x_{ij}\beta_j$, $i = 1, \ldots, n$. Then r_i is the linear predictor, the total effect of all covariates for person i. Let $\mathbf{r} = (r_1, \ldots, r_n)$, then $\mathbf{r}$ is a random vector with $\text{E}(\mathbf{r}) = 0$ and $\text{Cov}(\mathbf{r}) = \tau^2 XX'$. The model (2.1) simplifies to

$$\text{E}(Y_i|r_i) = h^{-1}(\alpha + r_i). \tag{2.2}$$

This is a simple random effects model, in which each sample has a random effect that influences its outcome. The covariance matrix between the random effects is known and is determined by the gene expression levels. If $\tau^2 > 0$, two samples i and j with similar gene expression patterns have correlated random effects r_i and r_j and therefore have a greater probability of having similar outcomes Y_i and Y_j than samples with less similar expression patterns.

2.4 The score test

A test for testing H_0 in the model (2.2) is discussed in Le Cessie and van Houwelingen (1995) and Houwing-Duistermaat et al. (1995). The marginal likelihood of Y in this model depends on only two or three parameters. These are α and τ^2 and sometimes, depending on the specific model, an extra dispersion parameter (e.g. the residual variance σ^2 of the outcome Y in an ordinary linear regression model).

In this section we first suppose that α and the dispersion parameter are known (the case where they are unknown is dealt with in section 2.6). In this case a score test for $\tau^2 = 0$ can be calculated by taking the derivative of the loglikelihood with respect to τ^2 at $\tau^2 = 0$, divided by the standard deviation of this derivative under H_0. This yields the test statistic

$$T = \frac{(Y-\mu)'R(Y-\mu) - \mu_2 \text{trace}(R)}{\left[2\mu_2^2 \text{trace}(R^2) + (\mu_4 - 3\mu_2^2)\sum_i R_{ii}^2\right]^{1/2}},$$

where $R = \frac{1}{m}XX'$ is an $n \times n$ matrix proportional to the covariance matrix of the random effects $\boldsymbol{r}$, $\mu = h^{-1}(\alpha)$ is the expectation of Y under H_0 and μ_2 and μ_4 the second and fourth central moments of Y under H_0.

It will be more convenient to use the equivalent, much simpler test statistic

$$Q = \frac{(Y-\mu)'R(Y-\mu)}{\mu_2}$$

which has expectation

$$\text{E}(Q) = \text{trace}(R) \tag{2.3}$$

and variance

$$\text{Var}(Q) = 2\text{trace}(R^2) + \left(\frac{\mu_4}{\mu_2^2} - 3\right)\sum_i R_{ii}^2. \tag{2.4}$$

The statistic Q is a quadratic form which is non-negative, because R is non-negative definite. It has been argued by Le Cessie and van Houwelingen (1995) that for a good asymptotic approximation to the distribution of Q is a scaled chi-squared distribution $c\chi_\nu^2$, where c is a scaling factor and ν is the number of degrees of freedom. This has been corroborated using simulations in Le Cessie and van Houwelingen (1995). Equating the mean and variance of $c\chi_\nu^2$ and Q yields $c = \text{var}(Q)/[2E(Q)]$ and $\nu = 2[E(Q)]^2/\text{var}(Q)$.

Note that the statistic Q and its distribution are easy to calculate for high-dimensional data because they only involve the small $n \times n$ "covariance" matrix $R = \frac{1}{m}XX'$ between the samples and never the potentially large $m \times m$ covariance matrix $\frac{1}{n}X'X$ between the genes. Testing a large number of genes therefore never gives computational problems.

2.5 Properties of the test

There are two ways of rewriting the test statistic Q to gain a better intuitive understanding of the test. The first can be used to show the influences of the genes, the second the influence of the samples. These two decompositions of Q will be the basis of various illustrative graphs in sections 2.9 and 2.10. Furthermore, the fact that the test is a score test also gives the test a nice optimality property.

For the first interpretation rewrite

$$Q = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{\mu_2}[X_i'(Y-\mu)]^2$$

where X_i ($i = 1, \ldots, m$) is the $n \times 1$ vector of the gene expressions of gene i. Note however that the expression $Q_i = \frac{1}{\mu_2}[X_i'(Y-\mu)]^2$ is exactly the test statistic that would have been calculated for a group of genes consisting only of the i-th single gene in the group of interest. Therefore the test statistic Q for a group of m genes is just the average of the statistics $Q_1, \ldots, Q_m$, calculated for the m single genes that the group consists of.

Each Q_i can be written as (a multiple of) the squared covariance between the expression pattern of the gene and the clinical outcome. Because the averaging is done at this squared covariance level, genes with large variance have much more influence on the outcome of the test statistic Q than genes with small variance. This is a nice property in the context of microarray analysis, because low variance genes are generally seen as uninteresting. This low variance usually implies that there is little biological variation in these genes.

For a different look at the test, the statistic Q can be written in the following way

$$Q = \frac{1}{\mu_2}\sum_{i=1}^{n}\sum_{j=1}^{n}R_{ij}(Y_i-\mu)(Y_j-\mu) \tag{2.5}$$

as the sum over all terms of the Hadamard (term-by-term) product of the matrices R and $(Y-\mu)(Y-\mu)'$. The matrix $R = \frac{1}{m}XX'$ is the "covariance" of the gene-expression patterns between the samples, and the matrix $(Y-\mu)(Y-\mu)'$ is the "covariance" matrix of the clinical outcomes of the samples. The statistic Q has a high value whenever the terms of these two matrices are correlated. This happens when the covariance structure of the gene-expressions between samples resembles the covariance structure between their outcomes. The score test can therefore be viewed as a test to see whether samples with similar gene-expression patterns also have similar outcomes.

An interesting property of a score test in general is that it maximizes the average power against all alternatives where the true value of the parameter is

small. Equivalently, in this case it has optimal power against the range of alternatives $R_t = \{\|\boldsymbol{\beta}\|^2 \leq t^2\}$ as $t^2 \to 0$. As R_t is an m-ball it contains relatively many alternatives with all β's nonzero but small, therefore the test is focussed mostly on detecting alternatives where many genes play a part. This is a desirable property, because the test is designed to say something about the group of genes as a whole.

2.6 Some technical adjustments

In the previous section it was assumed that α (and therefore μ) was known and that the dispersion parameter (if any) was also known. In practice this is never true. In this section some adjustments of the test are presented which correct for using estimated parameters.

First suppose that μ is unknown, but μ_2 and μ_4 are known. It is easily verified that

$$Y - \hat{\mu} = (I - H)(Y - \mu),$$

where $H = \frac{1}{n}\mathbf{1}\mathbf{1}'$ is the hat matrix for estimation of the mean μ of Y and $\mathbf{1}$ is a length n column vector of ones. Therefore calculating Q using $\hat{\mu}$ in stead of μ results in calculating

$$\begin{aligned} Q &= \frac{1}{\mu_2}(Y - \hat{\mu})'R(Y - \hat{\mu}) \\ &= \frac{1}{\mu_2}(Y - \mu)'(I - H)R(I - H)(Y - \mu). \end{aligned}$$

The mean and variance of Q are therefore simply given by (2.3) and (2.4) with R replaced by $\tilde{R} = (I - H)R(I - H)$. This is equivalent to centering the genes so that the average value of each gene over the samples is set to zero.

Correction for estimation of μ_2 is not so easy. Simply replacing μ_2 by its estimate $\hat{\mu}_2$ would generally lead to a test that is too conservative, because the numerator $(Y - \hat{\mu})'R(Y - \hat{\mu})$ and the denominator $\hat{\mu}_2 = \frac{1}{n}(Y - \hat{\mu})'(Y - \hat{\mu})$ of Q are positively correlated, so that the variance of Q is overestimated if this dependency is not taken into account.

Corrections for the variance of Q are available from Houwing-Duistermaat et al. (1995) for a the linear regression model (continuous clinical outcome) and for the logistic regression model (two groups). For a linear regression $Q = (Y - \hat{\mu})'R(Y - \hat{\mu})/\hat{\sigma}^2$, which has $E(Q) = \text{trace}(\tilde{R})$ and variance

$$\text{Var}(Q) = \frac{2}{n+1}\left[(n-1)\text{trace}(\tilde{R}^2) - \text{trace}^2(\tilde{R})\right].$$

For the logistic regression model $Q = (Y - \hat{\mu})' R (Y - \hat{\mu}) / [\hat{\mu}(1 - \hat{\mu})]$. This also has $E(Q) = \text{trace}(\tilde{R})$ and its variance can be approximated by

$$\begin{aligned} \text{Var}(Q) &\approx \frac{1 - 6\mu + 6\mu^2}{\mu(1-\mu)} \left[\sum_{i=1}^{n} \tilde{R}_{ii}^2 - \frac{1}{n} \text{trace}^2(\tilde{R}) \right] \\ &\quad + 2\text{trace}(\tilde{R}^2) - \frac{2}{n-1} \text{trace}^2(\tilde{R}). \end{aligned} \tag{2.6}$$

2.7 Handling small sample size

If the sample size is small the asymptotic formulae used to calculate the p-value may not be accurate. In this case a different approach could be to find the p-value using a permutation method. The empirical distribution of Q can be found by calculating Q for all permutations of the outcome Y or a random sample from these. The permutation method also works for other distributions of Y than normal or Bernoulli.

A drawback of the permutation method is that it is hard to demonstrate low p-values. Showing that a p-value is lower than 10^{-7} for example, needs at least 10^7 permutations. Often if the sample size is small, the total number of permutations is not large enough to attain very low significance levels. The minimum sample size needed to attain $\alpha = 0.05$ can be calculated as 2×4 samples if Y takes two values and 5 samples if Y is continuous. The permutation method is illustrated in section 2.9.

It is important to note that using permutations one calculates the distribution of Q under H_0 conditional on the set of observed outcomes in Y. For Y a group label this means that the sizes of the groups are taken as fixed; for a continuous outcome each value in the observed vector Y is assumed to occur exactly once. Therefore the permutation version is strictly speaking a different test (although asymptotically equivalent). The expectation and variance of Q under the null hypothesis and the p-value can therefore be systematically different, although in practice the difference is usually small except for very small samples.

2.8 Handling missing values

Missing values for some genes in the data set are not a problem. If some genes with missing values are too important to be left out of the analysis, the missing values can be handled by simply imputing the mean expression value of the same gene from the other samples (or the K-nearest samples). This allows the matrix $\tilde{R}$ of covariance between the gene expression patterns of the samples to be calculated using all available information. A nice property of this imputation

is that genes or samples with many missing values get a small variance and are therefore automatically given less weight in the analysis.

2.9 Application: AML/ALL

The first application is the well-known data set by Golub et al. (1999). These data were collected to for the purpose of distinguishing between *Acute Myeloid Leukemia* (AML) and *Acute Lymphoic Leukemia* (ALL) on the basis of gene expression. There are microarray data of 7,129 genes from 27 ALL and 11 AML patients. A preselection of genes was made in the same manner as in earlier publications on this data set (Eilers et al., 2001; Golub et al., 1999), truncating very high and very low expression levels and removing genes whose truncated expression showed no variation. This left 3,571 genes. There were no missing values. This data set will be used here to illustrate the use of the score test on all genes. The null hypothesis to be tested here is whether AML and ALL patients are different with respect to their overall gene expression pattern.

Test result The ALL patients were coded 0 and the AML patients 1. Now $\hat{\mu} = 11/38$, which was used to calculate

$$Q \approx 13.2.$$

Under the null hypothesis H_0 the distribution has $E(Q) \approx 2.88$ and s.e.$(Q) \approx 0.78$, calculated using (2.6). This results in a rejection of H_0 with a p-value $\approx 1.6 \times 10^{-14}$, calculated on the $c\chi^2_\nu$-distribution with $c \approx 0.11$ and $\nu \approx 27.0$.

This shows that AML and ALL patients do indeed differ enormously with respect to their overall gene expression signature. The extremely low p-value here can be seen as an explanation why many people using many different methods have been able to find good discriminating rules between AML and ALL on the basis of these data.

The permutation method Because the p-value is so extreme, it is prudent to check the distribution of Q empirically. We do this by randomly taking 100,000 permutations of the vector Y of outcomes, calculating Q and making a histogram. The result is given in figure 2.1, with the observed value of Q in the real data set indicated by an arrow. The empirical mean and standard deviation are $\bar{Q} \approx 2.96$ and s.e.$(Q) \approx 0.80$, which are not very far from the theoretical values.

The empirical p-value is the number of times the Q for the permuted Y is as least as large as the 'true' Q, divided by the number of permutations. In principle, because there are about 3.3×10^{29} possible permutations of Y, this can

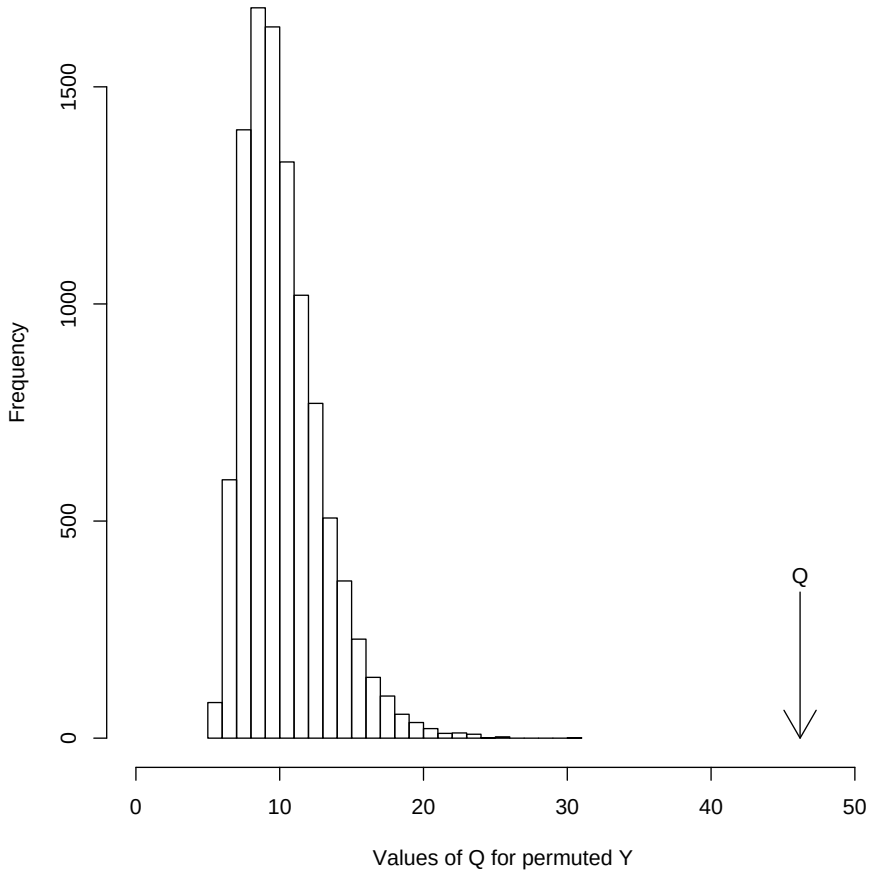

FIGURE 2.1: *Histogram of values of the test statistic Q for 100,000 permutations of Y, compared with the observed value.*

be calculated to almost any desired accuracy. But taking only 10^5 permutations (about 10 seconds on a normal computer) we can only say that the p-value is most probably below 10^{-5}, although figure 2.1 suggests that it is much lower than that.

The Regression and Checkerboard Plots It has already been explained using (2.5) that the test statistic Q evaluates the resemblance between the covariance between the gene expressions of all pairs of samples and the covariance between their clinical outcomes. This comparison might also be done by inspec-

tion. Figure 2.2 is an image of the symmetric matrix $\tilde{R}$, with white denoting that an entry is larger than the median off-diagonal element and black that it is smaller.

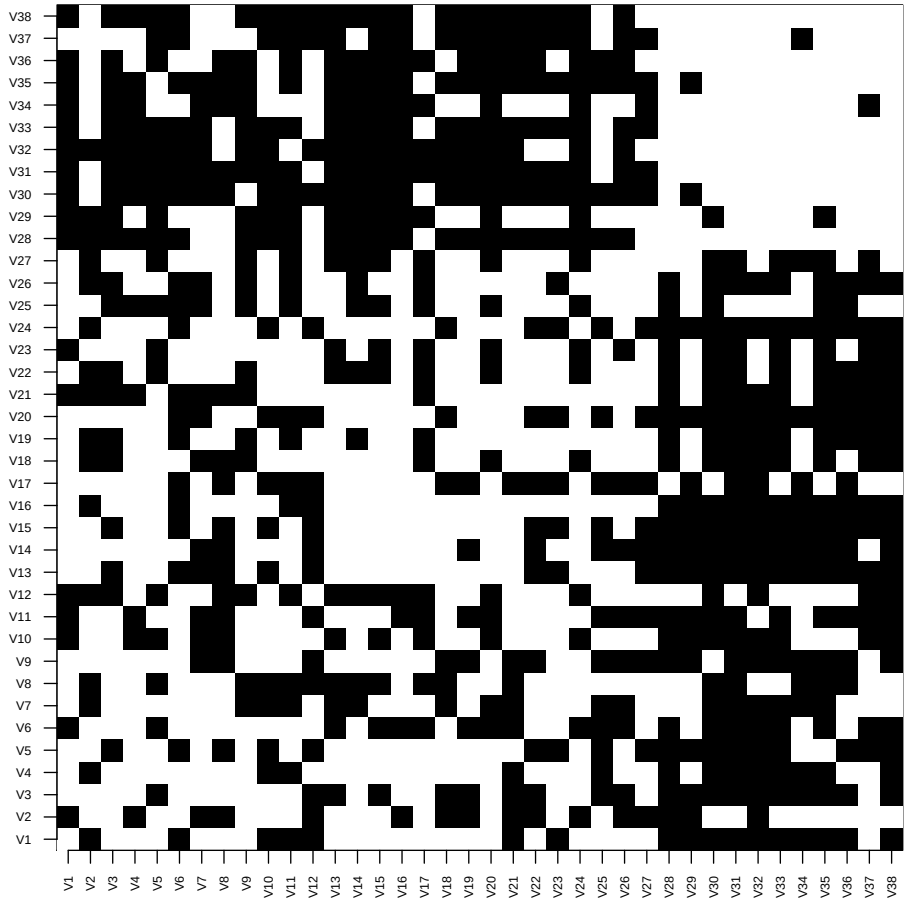

FIGURE 2.2: *Checkerboard plot for the AML/ALL dataset, showing the matrix $\tilde{R}$ of covariance between the gene expressions of all pairs of samples. White = above median; black = below median.*

From this image it is easy to recognize that the true outcomes Y have been sorted, starting with the 27 ALL patients and continuing with the 11 AML patients. The block-like structure of the matrix $\tilde{R}$ strongly resembles the block structure of the covariance matrix between the outcomes Y. This can be used as an illustration of the low p-value that was found.

This method of visualization works best when the outcome is a group indicator. For continuous outcomes, two images of $\tilde{R}$ and $S = (Y - \hat{\mu})(Y - \hat{\mu})'$ might be placed side by side for comparison, perhaps with the samples sorted by their outcomes to simplify the structure of the two matrices. In that case a multi-color plot might be preferred over a black and white one.

Some interesting things can be learned from the plot in figure 2.2. In the first place it can be seen from the image that the AML group is much more homogeneous than the ALL group. Another thing that can be noticed is that some arrays do not seem to fit very well into the block-like structure. The ALL arrays #2 and #12 for example (second and twelfth row/column) seem at least as similar to the AML group as to the ALL group. These arrays could have been wrongly classified or be of poor quality.

A second way of visualizing the test is by plotting the off-diagonal entries of R against those of $S = (Y - \hat{\mu})(Y - \hat{\mu})'$. This is a way of representing Q, because Q is proportional to the covariance between the plotted entries and can therefore be represented by the slope of the regression line of the off-diagonal entries of R on those of S. This type of plot is also very useful when the outcome Y is continuous.

For the AML/ALL dataset, the plot is shown in figure 2.3. Because Y only takes the values 0 and 1, the matrix S takes only three values. From left to right on the x-axis, these are ALL versus AML , ALL versus ALL and AML versus AML. The AML/AML comparisons have a higher covariance between outcomes than the ALL/ALL comparisons because there are fewer AML (so that $Y_i - \hat{\mu} = \frac{27}{38}$ for the AML and $Y_i - \hat{\mu} = -\frac{11}{38}$ for the ALL). The large value of Q is seen from the steep slope of the regression line.

Using this type of plot the possibly outlying arrays can be investigated further. In figure 2.4 all points corresponding to pairs of arrays that involve array #12 have been replaced by crosses. An extra dotted regression line is drawn for reference, which is the least squares fit only through the crosses. This way it can be seen that ALL array #12 actually resembles the AML arrays more than it resembles the other ALL arrays. This is not suggestive of bad data quality (in which case #12 would resemble none of the arrays very well) so it either indicates a misclassification of #12, or perhaps it might be that ALL is quite diverse and some forms are genetically closer to AML.

2.10 Application: Heat Shock

The second dataset contains six replicates each of a cell line treated with a heat shock (hs+) and untreated (hs−). These samples were labelled with two different fluorescent dyes and co-hybridized in hs+/hs− pairs on six spotted

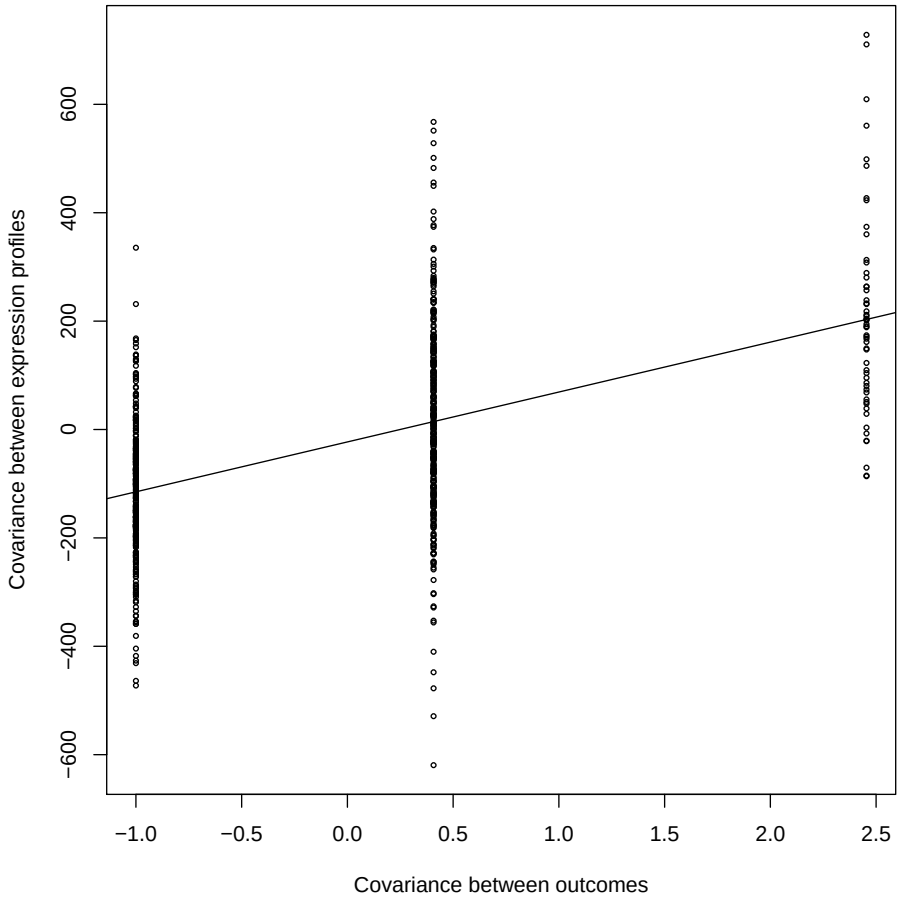

FIGURE 2.3: *Regression plot I: visualization of Q as a regression between off-diagonal entries of S and $\tilde{R}$.*

oligonucleotide microarrays containing 20,160 genes. Normalization on the 12 samples was carried out using the variance stabilizing method VSN (Huber et al., 2002).

In this dataset two groups of genes were of specific interest. One was a group of 27 genes which were classified for biological process as heat shock response genes by the Gene Ontology Consortium (Ashburner et al., 2000, www.geneontology.org). Another group of 17 genes belonged to different biological processes but their gene names referred to heat shock.

The test on the total group of all 20,160 genes gave a non-significant result

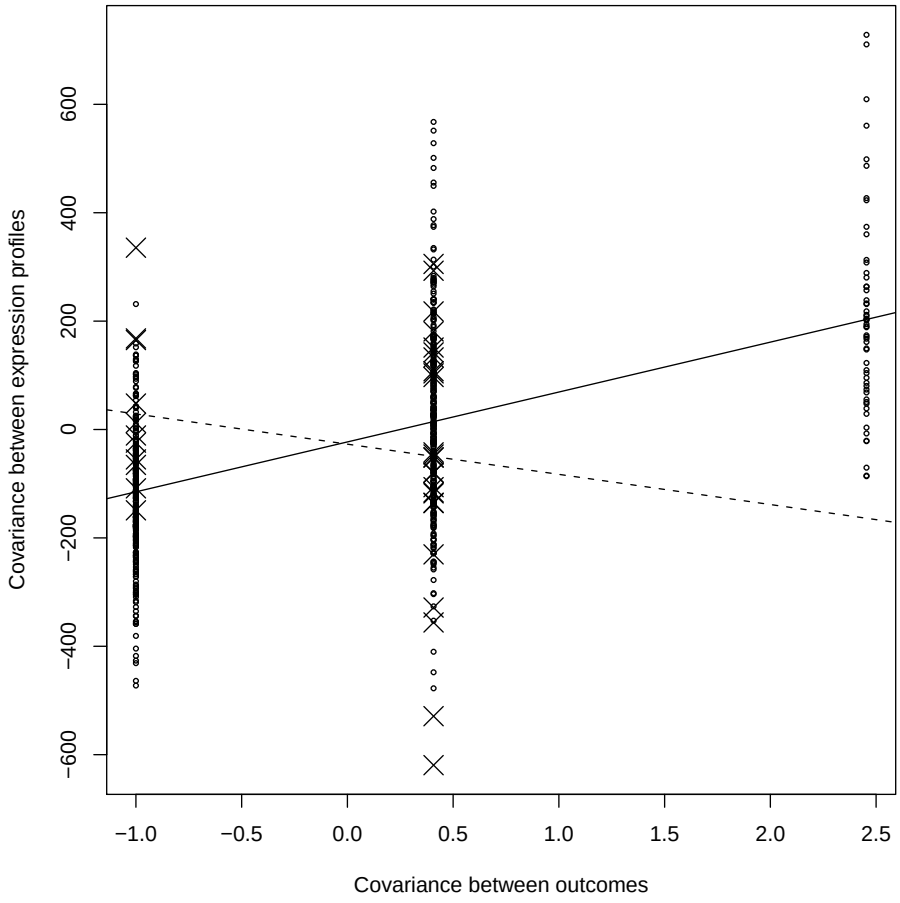

FIGURE 2.4: *Regression plot II: visualization of Q as a regression between off-diagonal entries of S and $\tilde{R}$. Crosses involve array* #12

($p = 0.94$). Looking at all genes, it could not be proved that any gene was affected: the overall expression pattern was not notably different between the hs+ and hs− groups. However, using the global test on the selected genes gave a different picture. The global test on the 27 genes known to function in heat shock response had an empirical p-value of 0.017. The expression pattern of this group of genes was therefore different between the two experimental conditions. The other group of 17 genes with heat shock' in the name only had a non-significant p-value of 0.25.

As an informal comparison, we did an analysis using SAM (Tusher et al.,

2001). On the optimal false discovery rate, which was 11%, we could only find a small set of nine differentially expressed genes. This set contained only a single gene from the group of 27 heat shock genes (Gene NM 002155 in figure 2.5).

A gene diagnostics plot When testing a small group of genes for differential expression of the group, it is often interesting to look at the single genes, even if the group is the main focus of interest. A group of genes can yield a significant test result because a few genes are very much differentially expressed or because most genes are a little differentially expressed. This can be an interesting biological difference. In other cases single genes within the group may be of interest in themselves.

The influence of single genes on the test result can be evaluated in a Gene Influence Plot, as shown for the group of 27 genes in figure 2.5. The bars in the figure indicate the values of Q_i for each gene (see section 2.5). Each Q_i gives the value of the test statistic for a group of genes consisting only of that single gene. A line is drawn for reference to indicate the expected length $\mathrm{E}(Q_i)$ of the bar under the null hypothesis.

From the figure it can be seen which genes contribute positively to a high value of the test statistic and which do not contribute. The difference in expected contribution arises because genes which have greater variance among all arrays are naturally expected to also have a greater discriminating power. In this data set we can see that really only a minority of 5 or 6 genes out of 27 is clearly above the reference line and that the majority of the genes do not show any effect.

2.11 Discussion

The test presented in this paper is a useful new tool for the analysis of microarray data. It allows researchers to use prior information on groupings of genes and to specifically investigate groups of genes that interest them from a biological point of view.

In cases where there is a single candidate group of interest, the global test opens the door to real inference: testing hypotheses about biological mechanisms based on theory or past research. In other cases, when researchers have many candidate pathways available for example from the Gene Ontology database (Ashburner et al., 2000, *www.geneontology.org*) or programs like GenMAPP (*www.genmapp.org*), the global test can be used to find promising pathways. Alternatively, the clusters from a cluster analysis can be assigned a p-value to mark how much the cluster is co-regulated with the clinical outcome.

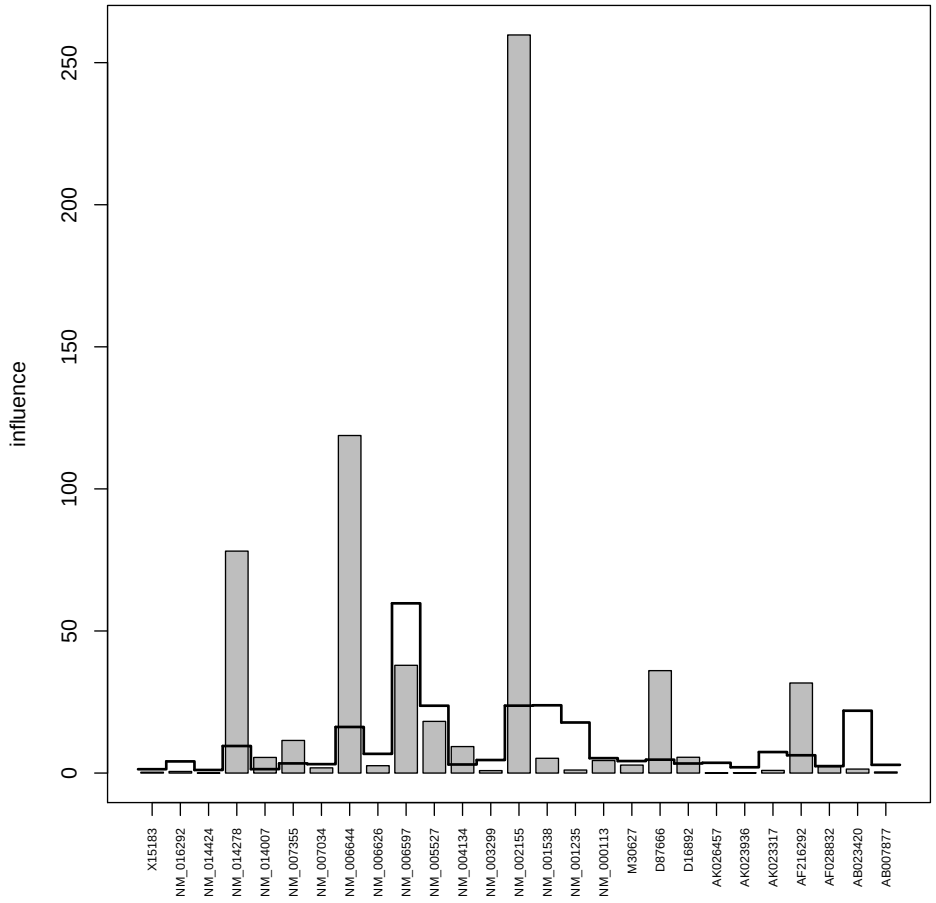

FIGURE 2.5: *Gene influence plot for the Heat Shock data. High bars indicate influential genes. Reference line is the expected influence under the null hypothesis.*

Test results for groups of different sizes are fully comparable. However, when many groups of genes are to be tested, multiple testing procedures come back into play (Benjamini and Hochberg, 1995). Nested groups may be tested without adjustments to the α-level. Always keep in mind that groups of genes may never be chosen with reference to the clinical outcome.

Furthermore, using the test on all genes could be a useful preliminary data quality check. If the test is not significant, samples with a similar clinical outcomes do not have very similar gene expression patterns. In this case it is unlikely that there are many genes highly differentially expressed and it is un-

likely that a good classification rule can be found on the basis of all genes. Because of the close connection of the global test to penalized regression methods, the p-value that results from the test can be used as a quality label for the classification rule found with these methods.

CHAPTER 3

Testing Association of a Pathway with Survival

Abstract

A recent surge of interest in survival as the primary clinical endpoint of microarray studies has called for an extension of the Global Test methodology (Goeman et al., 2004) to survival. We present a score test for association of the expression profile of one or more groups of genes with a (possibly censored) survival time. Groups of genes may be pathways, areas of the genome, clusters from a cluster analysis or all genes on a chip. The test allows one to test hypotheses about the influence of these groups of genes on survival directly, without the intermediary of single gene testing. The test is based on the Cox proportional hazards model and is calculated using martingale residuals. It is possible to adjust the test for the presence of covariates. We also present a diagnostic graph to assist in the interpretation of the test result, visualizing the influence of genes. The test is applied to a tumour data set, revealing pathways from the Gene Ontology database that are associated with survival of patients. The global test for survival has been incorporated into the R-package globaltest (from version 3.0), available from http://www.bioconductor.org.

3.1 Introduction

A microarray experiment typically results in many thousands of measurements, each relating to the expression level of a single gene. Single genes, however, are often not the primary theoretical focus of the researcher, who might be

This is a pre-copy-editing, author-produced version of an article accepted for publication in *Bioinformatics* following peer review. The definitive publisher-authenticated version : J. J. Goeman, J. Oosting, A. M. Cleton-Jansen, J. Anninga, and J. C. van Houwelingen (2005). Testing association of a pathway with survival using gene expression data. *Bioinformatics 21*(9), 1950–1957 is available online at: `http://dx.doi.org/10.1093/bioinformatics/bti267`

more interested in certain pathways or genomic regions that are suspected to be biologically relevant.

For this reason we have introduced the Global Test for groups of genes (Goeman et al., 2004), which allows the unit of analysis of the microarray experiment to be shifted from the single gene level to the pathway level, where a "pathway" may be any set of genes, e.g. chosen using the *Gene Ontology* database or from earlier experiments. For every pathway, the Global Test can test (with a single test) whether the expression profile of that pathway is significantly associated with a clinical variable of interest. This allows researchers immediately to test theoretical hypotheses on the clinical importance of certain pathways. Even when such hypotheses are not directly available from biological theory or past research, the Global Test can significantly reduce the multiple testing problem, because there are typically much fewer pathways than genes.

In the original publication of the Global Test, the clinical variable could be either a continuous measurement or a $0/1$ group indicator. Recently, however, there has been a surge of interest in survival time of patients as the primary clinical outcome in a microarray experiment. Many studies focus on prediction of survival, e.g. in breast cancer Van 't Veer et al. (2002), Van de Vijver et al. (2002) and Pawitan et al. (2004) and in lung cancer Wigle et al. (2002) and Beer et al. (2002). Other studies use multiple testing methods to find genes which are associated with survival (Jenssen et al., 2002).

The present paper extends the Global Test methodology to survival outcomes. It allows the researcher to test whether the expression profile of a given set of genes is associated with survival. More precisely, it tests whether individuals with a similar gene expression profile tend to have similar survival times. A significant pathway may be a mix of genes which are upregulated for patients with short survival time, genes which are downregulated for the same patients, and other genes that show no association with survival at all.

The test of the present paper is based on the Cox proportional hazards model. Therefore, it avoids the requirement of many analysis strategies to choose an arbitrary cut-off (e.g. five years survival), but uses all survival information that is present in the data. Technically, the test is derived from the goodness-of-fit test of the Cox model by Verweij et al. (1998). The original Global Test was derived in a similar way from a goodness-of-fit test for generalized linear models (Le Cessie and van Houwelingen, 1995). The two Global Tests are therefore highly comparable and allow quite similar interpretations.

In this paper we also show how the test can be adjusted for the presence of covariates (possible confounders or competing risk factors). This allows better use of the Global Test in observational studies. Furthermore, it allows the researcher to establish that the microarray really adds something to the predictive

performance of known risk factors, showing that it is not simply an expensive way to measure risk factors already known. It also allows the test to be used on more complex designs than a simple one-sample follow up study. The approach will be illustrated on a data set of 17 osteosarcoma patients, testing pathways from the Gene Ontology database.

The new Global Test method presented in this paper has been incorporated into the R-package *globaltest*, version 3.0, which is available from BioConductor (Gentleman et al., 2004, www.bioconductor.org).

3.2 The model

The Global Test exploits the duality between association and prediction. By definition, if two things are associated, knowing one improves prediction of the other. Hence, if survival is associated with gene expression profile, this means that knowing the gene expression profile allows a better prediction of survival than not knowing the expression profile.

With this idea in mind we make a prediction model for prediction of survival from the gene expression measurements. The most convenient choice for such a model is the Cox proportional hazards model, which is the most widely used model for survival data in medical research. The Cox model uses the full empirical distribution of the survival times and it can handle censored data, i.e. samples for which the exact survival time is not known, but for which it is only known that the patient is still alive at a certain moment (Klein and Moeschberger, 1997). The use of the Cox model requires a true follow-up study design, meaning that patients were not selected on their survival times in any way. If such a patient selection was made, the methods of this paper may not be appropriate: in Van 't Veer et al. (2002), for example, where a selected group of early metastases was compared to a selected group which was at least five years metastasis-free, the original Global Test for a $0/1$ outcome is preferable (Goeman et al., 2004).

Suppose the matrix of normalized gene expression measurements for the group of genes of interest is given by the $n \times m$ matrix X with elements x_{ij}, where n is the sample size and m the number of genes in the group. Suppose also that there is a number $p \geq 0$ of covariates for each patient, which we put in an $n \times p$ data matrix Z with elements z_{ij}. It will be assumed that $p < n$, but no such restriction is put on m.

Cox's proportional hazards model (Klein and Moeschberger, 1997, chapter 8) assumes the hazard function at time t for individual i to relate to the covariates as

$$h_i(t) = h(t)e^{c_i + r_i}, \tag{3.1}$$

where $h(t)$ is an unknown baseline hazard function and $c_i + r_i$ is a linear function of the covariates, which is split up in our case into $r_i = \sum_{k=1}^{m} \beta_k x_{ik}$, relating to the gene expressions, and $c_i = \sum_{l=1}^{p} \gamma_l z_{il}$, relating to the covariates. The hazard function determines the survival function $S_i(t)$, which gives the probability that individual i survives up to time t, through

$$S_i(t) = e^{-H_i(t)},$$

where $H_i(t) = \int_0^t h_i(s)\,\mathrm{d}s$ is the cumulative hazard up to time t.

In this model showing that the gene expressions are associated with survival is equivalent to rejecting the null hypothesis

$$H_0 : \beta_1 = \ldots = \beta_m = 0,$$

that all regression coefficients relating to the gene expressions are zero. If m were always small, we could test H_0 using classical tests which were developed for the Cox model. These tests do not work for general m, however (for an overview of these classical tests see Klein and Moeschberger, 1997, section 8.2).

To obtain a test that works whatever the value of m, we put an extra assumption on the regression coefficients $\beta_1, \ldots, \beta_m$. We assume that the regression coefficients of the genes are random and a priori independent with mean zero and common variance τ^2. The null hypothesis now becomes simply

$$H_0 : \tau^2 = 0,$$

so that the dimension of H_0 does not depend on m anymore. Note that the coefficients $\gamma_1, \ldots, \gamma_p$ of the covariates are not assumed to be random.

The Cox model with random coefficients is an empirical Bayesian model and is closely linked to penalized likelihood methods. It should be noted that we have not assumed a specific distributional form for the regression coefficients; the derivation of our test is invariant to the choice of the shape of this distribution. Choosing a Gaussian distribution results in a Cox ridge regression model (Pawitan et al., 2004); choosing a double exponential distribution results in a LASSO model (Tibshirani, 1997). Both models can also be used to predict survival times of patients.

In the context of testing it is most insightful to view the prior distribution of the regression coefficients as the *focus of the power* of the test. The test that will be derived in the next section will be a score test, which has the property that it has optimal power against alternatives with small values of the parameter τ^2. This property stems from the fact that the score test is equivalent to the likelihood ratio test in the limit where the alternative $\tau^2 \to 0$ (Cox and Hinkley, 1974). Alternatives with small values of τ^2 tend to have small values of $\sum \beta_i^2$,

so that the test can be said to be optimal on average against alternatives with small values of $\sum \beta_i^2$. These alternatives are mainly alternatives which have all or most regression coefficients non-zero but small. The test can therefore be said to be optimized against alternatives in which all or most genes have some association with the outcome. This alternative is precisely the situation in which we are interested, because we want to say something about the pathway as a whole.

Alternative tests can easily be derived for regression coefficients with a more complex covariance structure. If the vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)'$ is assumed a priori to have mean zero and covariance matrix $\tau^2\Sigma$, the resulting test of H_0 would be optimal against alternative with small values of $\boldsymbol{\beta}'\Sigma\boldsymbol{\beta}$. The standard choice of $\Sigma = I_m$ distributes power equally over all directions of $\boldsymbol{\beta}$, while a different choice will have more power against deviations from H_0 in directions which correspond to the larger eigenvalues of Σ. This property could be exploited in the derivation of a test for a specific purpose or to incorporate prior knowledge. In this paper we shall restrict ourselves to $\Sigma = I_m$.

3.3 Derivation of the test

Testing association of a group of genes with survival can therefore be done by testing H_0 in the empirical Bayesian model (3.1) with random regression coefficients. In this section we will derive the test statistic for this test. A score test for the same model has also been studied by Verweij et al. (1998) in the context of testing the fit of the Cox model. Their derivation was based on the partial likelihood of the Cox model. In this paper we give an alternative derivation based of the full likelihood and a simpler martingale argument.

We derive the test in stages. First suppose that all parameters except τ^2 are known, i.e. the regression coefficients $\gamma_1, \ldots, \gamma_p$ and the baseline hazard function $h(t)$ are known. In this simplified situation it will be relatively easy to derive the score test, which can be generalized to the situation with unknown parameters later in this section.

The basic score test

By definition a score test is based on the derivative of the log-likelihood at the value of the parameter to be tested. Suppose for each individual i we have observed a survival time t_i and a status indicator d_i, where $d_i = 1$ indicates death (the patient died at t_i) and $d_i = 0$ censoring (the patient was lost to follow-up at t_i). The loglikelihood of τ^2 in the model (3.1) is

$$L(\tau^2) = \log\left\{\mathrm{E}_{\mathbf{r}}\left[\exp\left(\sum_{i=1}^{n} f_i(r_i)\right)\right]\right\}, \tag{3.2}$$

where

$$f_i(r_i) = d_i[\log\{h(t_i)\} + c_i + r_i] - H(t_i)e^{c_i+r_i}$$

is the contribution to the loglikelihood of individual i for fixed r_i, and $H(t) = \int_0^t h(s)\,\mathrm{d}s$ is the cumulative baseline hazard.

From the assumptions on the distribution of $\beta_1, \ldots, \beta_m$ we can derive the distribution of $\mathbf{r} = (r_1, \ldots, r_n)'$, the vector of the linear effects of the gene expressions. This $\mathbf{r}$ has mean zero and covariance matrix $\tau^2 R$, where $R = XX'$. For the general likelihood (3.2) and an $\mathbf{r}$ of this form, Le Cessie and van Houwelingen (1995) have used a Taylor approximation to derive that

$$\frac{\partial L(0)}{\partial \tau^2} = \frac{1}{2}\left(\sum_i R_{ii} \frac{\partial^2 f_i(0)}{(\partial r_i)^2} + \sum_{i,j} R_{ij} \frac{\partial f_i(0)}{\partial r_i} \frac{\partial f_j(0)}{\partial r_j} \right).$$

For the Cox model this becomes

$$\frac{\partial L(0)}{\partial \tau^2} = \frac{1}{2}\left(\sum_{i,j} R_{ij}(d_i - u_i)(d_j - u_j) - \sum_i R_{ii} u_i \right), \tag{3.3}$$

where $u_i = e^{c_i} H(t_i)$, $i = 1, \ldots, n$, is the hazard incurred by individual i up to time t_i. Note that $d_i - u_i$ is the martingale residual of individual i at time t_i (Klein and Moeschberger, 1997, section 11.3).

For known $H(t)$ and known $c_1, \ldots, c_n$, the expression (3.3) can be standardized to have unit variance and used as the score test statistic. When these parameters are unknown, we must plug in maximum likelihood estimates for them under the null model in which $\tau^2 = 0$. Standardizing the score test is traditionally done using the Fisher Information, calculated from the second derivatives of the loglikelihood. In this case these calculations are very unpleasant, and it turns out to be simpler to standardize using the estimated variance of the test statistic.

Using estimated baseline hazard

We shall first plug in the estimate for the cumulative hazard $H(t)$, but still assume that $\gamma_1, \ldots, \gamma_p$ and hence $c_1, \ldots, c_n$ are known. As the maximum likelihood estimate of $H(t)$ we can take the Breslow estimator (Klein and Moeschberger, 1997, section 8.6)

$$\hat{H}(t_i) = \sum_{t_j \le t_i} \frac{d_j}{\sum_{t_k \ge t_j} e^{c_k}}, \qquad i = 1, \ldots, n,$$

and write $\hat{u}_i = e^{c_i} \hat{H}(t_i)$, $i = 1, \ldots, n$.

Using twice the estimated derivative of the log-likelihood (3.3) as the test statistic and writing it in matrix notation we get the test statistic

$$T = (\mathbf{d} - \hat{\mathbf{u}})' R (\mathbf{d} - \hat{\mathbf{u}}) - \text{trace}(R\hat{U}) \tag{3.4}$$

where $\mathbf{d} = (d_1, \ldots, d_n)'$, $\hat{\mathbf{u}} = (\hat{u}_1, \ldots, \hat{u}_n)'$ and $\hat{U} = \text{diag}(\hat{\mathbf{u}})$, an $n \times n$ diagonal matrix with $\hat{U}_{ii} = \hat{u}_i$.

The derivation of estimates for the mean and variance of T is quite technical and will be given in the separate subsection on page 41. The estimated mean is

$$\hat{\mathrm{E}}(T) = -\text{trace}(RPP'), \tag{3.5}$$

where P is an $n \times n$ matrix with i, j-th element

$$p_{ij} = \mathbf{1}_{\{t_i \geq t_j\}} \frac{d_j e^{c_i}}{\sum_k \mathbf{1}_{\{t_k \geq t_j\}} e^{c_k}},$$

where $\mathbf{1}_{\{\cdot\}}$ indicates an indicator function. Each p_{ij} is the increment of the cumulative hazard incurred by individual i at time t_j, so that $\sum_i p_{ij} = d_j$ and $\sum_j p_{ij} = \hat{u}_i$.

The estimated variance of T is

$$\widehat{\mathrm{Var}}(T) = \sum_{j=1}^{n} \mathbf{p}_j{}' \, \text{diag}(\mathbf{t}_j \mathbf{t}_j'), \tag{3.6}$$

where $\mathbf{p}_j$ is the j-th column of P and $\mathbf{t}_j = (I - \mathbf{1}\mathbf{p}_j{}')[\text{diag}(R) + 2R(\mathbf{m}_j - \mathbf{p}_j)]$. The diag of a square matrix is the column vector of its diagonal elements; $\mathbf{1}$ is an n-vector of ones, and $\mathbf{m}_j$ is the j-th column of the matrix $M = (D - P)B$, where $D = \text{diag}(\mathbf{d})$ is a diagonal matrix with $D_{ii} = d_i$ and B is an $n \times n$ matrix with elements $b_{ij} = \mathbf{1}_{\{t_i < t_j\}}$. The elements m_{ij} of M can be interpreted as the estimated martingale residual of individual i just before time t_j.

For purposes of interpretation it is often easier to take

$$T_0 = (\mathbf{d} - \hat{\mathbf{u}})' R (\mathbf{d} - \hat{\mathbf{u}})$$

as the unstandardized test statistic. It has $\hat{\mathrm{E}} T_0 = \text{trace}(R\hat{U} - PP')$ and $\widehat{\mathrm{Var}}(T_0) = \widehat{\mathrm{Var}}(T)$, so that it leads to the same standardized test statistic:

$$Q = \frac{T - \hat{\mathrm{E}} T}{\widehat{\mathrm{Var}}(T)} = \frac{T_0 - \hat{\mathrm{E}} T_0}{\widehat{\mathrm{Var}}(T_0)}.$$

Using estimated regression coefficients

In general the regression coefficients $\gamma_1, \ldots, \gamma_p$ of the covariates are not known but must be estimated. Replacing $\gamma_1, \ldots, \gamma_p$ by their maximum likelihood estimates will still give a valid score test for H_0, but with a different distribution of the test statistic. We use the following approximation to this distribution which is derived by Verweij et al. (1998).

The estimated martingale residuals $\mathbf{d} - \tilde{\mathbf{u}}$ based on the estimated $\hat{\gamma}_1, \ldots, \hat{\gamma}_p$ can be approximated in a first order Taylor approximation by

$$\mathbf{d} - \tilde{\mathbf{u}} \approx (I - V)(\mathbf{d} - \hat{\mathbf{u}}) \tag{3.7}$$

with $V = WZ(ZWZ')^{-1}Z'$, $W = \hat{U} - PP'$ and Z the $n \times p$ data matrix of the fixed covariates. Therefore the unstandardized test statistic T_0 can be approximated as

$$T_0 \approx (\mathbf{d} - \hat{\mathbf{u}})'\tilde{R}(\mathbf{d} - \hat{\mathbf{u}})$$

with $\tilde{R} = (I - V)'R(I - V)$. The expectation of T_0 can be estimated using the formulae in section 3.3. They are approximately

$$\hat{\mathrm{E}}T_0 \approx \mathrm{trace}(\tilde{R}W)$$

and

$$\widehat{\mathrm{Var}}(T_0) \approx \sum_{j=1}^{n} \mathbf{p}_j' \,\mathrm{diag}(\tilde{\mathbf{t}}_j\tilde{\mathbf{t}}_j'),$$

with $\tilde{\mathbf{t}}_j = (I - \mathbf{1}\mathbf{p}_j')[\mathrm{diag}(\tilde{R}) + 2\tilde{R}(\mathbf{m}_j - \mathbf{p}_j)]$. To evaluate $\hat{\mathrm{E}}T_0$ and $\widehat{\mathrm{Var}}(T_0)$ we replace the parameter values of $\gamma_1, \ldots, \gamma_p$ by their estimates. Simulations in Verweij et al. (1998) show this approximation to be quite accurate.

The distribution of the test statistic

There are two ways to calculate the p-value of the test: by asymptotic theory and by permutation arguments. We outline both options and their advantages.

In equation (3.3) it will be shown that the centered test statistic $T - \hat{\mathrm{E}}T$ can be written as a linear combination of n martingales. Therefore by the martingale central limit theorem (Andersen et al., 1993) the distribution of the standardized Q converges to a standard normal distribution as $n \to \infty$. This fact motivates the use of a normal approximation to the distribution of Q for calculating the one-sided p-value (see also simulation results by Verweij et al., 1998).

For small samples the asymptotic distribution may not be reliable enough. An alternative is to calculate Q for all, or a random sample of many (10,000), permutations of the martingale residuals of the n samples. This randomly redistributes the vectors of gene expression measurements over the individuals,

while keeping the relationship between the fixed covariates and survival the same. The resulting distribution is another approximation to the null distribution of Q, which can be used to find the p-value. Use of the permutation null distribution requires the assumption that there is no relationship between the gene expressions on the one hand and the covariates and the censoring mechanism on the other hand: permuting destroys these associations. This makes the permutation null distribution less useful when covariates are present.

The main advantage of the permutation-based p-value is that it gives an "exact" p-value, which is guaranteed to keep the alpha level, provided enough permutations are used. This is especially useful for smaller sample sizes, where we may not trust the normality of the distribution of Q. The advantage of the asymptotic theory p-value—aside from being much quicker to calculate—is that it has more power: the permutation based p-value does not use the full null distribution, but the null distribution conditional on the set of observed martingale residuals. With this conditioning the test loses some power, as the set of observed residuals is informative on the parameter τ^2.

Counting process calculations

In this technical section we calculate the mean and variance of the test statistic T under the null hypothesis for known $c_1, \ldots, c_n$ but estimated $H(t)$, as given in (3.5) and (3.6). For this we will use a counting process notation (Andersen et al., 1993; Fleming and Harrington, 1991). The strategy we will use is common in martingale theory: we write our test statistic T as the limit of a process $T(t)$ as $t \to \infty$ and decompose $T(t)$ into a martingale and a compensator. The limit of the compensator is the estimator of the mean of T and the limit of the predictable variation process is the estimate of the variance. For an alternative derivation, see Verweij et al. (1998).

Let $\mathbf{Y}(t) = (Y_1(t), \ldots, Y_n(t))'$ be the vector of at-risk processes of individuals $1, \ldots, n$ and $\mathbf{N}(t) = (N_1(t), \ldots, N_n(t))'$ the vector of their counting processes. Then $\mathbf{N}$ has intensity process $\mathbf{\Lambda} = C\mathbf{Y}(t)H(t)$, where C is a diagonal matrix with $C_{ii} = e^{c_i}$, $i = 1 \ldots, n$. Write $N(t) = \mathbf{1}'\mathbf{N}(t)$, the total counting process.

In the counting process notation, $\mathbf{d} = \mathbf{N}(\infty)$ and $\hat{\mathbf{u}} = \hat{\mathbf{\Lambda}}(\infty)$ with $\hat{\mathbf{\Lambda}}(t) = \int_0^t \mathbf{V}(s)\mathbf{1}' \, \mathrm{d}\mathbf{N}(s)$, where $\mathbf{V} = C\mathbf{Y}(\mathbf{1}'C\mathbf{Y})^{-1}$. Wherever possible we will drop the dependence on time for convenience of notation.

Note that the compensator of $\hat{\mathbf{\Lambda}}$ is $\mathbf{\Lambda}$, which is also the compensator of $\mathbf{N}$. Write $\widehat{\mathbf{M}} = \mathbf{N} - \hat{\mathbf{\Lambda}}$. Then $\mathbf{d} - \hat{\mathbf{u}} = \widehat{\mathbf{M}}(\infty)$ and $\widehat{\mathbf{M}}(t) = \int_0^t (I - \mathbf{V}\mathbf{1}') \, \mathrm{d}\mathbf{N}$ is a martingale vector. Subtracting the intensities and writing $\mathbf{M} = \mathbf{N} - \mathbf{\Lambda}$,

$$\widehat{\mathbf{M}}(t) = \int_0^t (I_n - \mathbf{Y}\mathbf{1}') \, \mathrm{d}\mathbf{M}.$$

The statistic T is $T(\infty)$, with

$$T(t) = \text{trace}[R\widehat{\mathbf{M}}\widehat{\mathbf{M}}' - R\,\text{diag}(\hat{\mathbf{\Lambda}})].$$

From the integration by parts formula (Fleming and Harrington, 1991, theorem A.1.2) it follows that, almost surely,

$$\begin{aligned}\widehat{\mathbf{M}}\widehat{\mathbf{M}}' &= \int_0^t \widehat{\mathbf{M}}^-\,\mathrm{d}\widehat{\mathbf{M}}' + \int_0^t \mathrm{d}\widehat{\mathbf{M}}\,(\widehat{\mathbf{M}}^-)' \\ &\quad + \int_0^t (I - \mathbf{1}\mathbf{V}')\text{diag}(\mathrm{d}\mathbf{N})(I - \mathbf{V}\mathbf{1}') \qquad (3.8)\end{aligned}$$

where $\widehat{\mathbf{M}}^-(s) = \widehat{\mathbf{M}}(s-)$ is a predictable process. Using (3.8) and some linear algebra we can say that, almost surely,

$$T(t) = \int_0^t (\text{diag}(R)' + 2(\widehat{\mathbf{M}}^-)'R - \mathbf{V}'R)(I - \mathbf{V}\mathbf{1}')\,\mathrm{d}\mathbf{N} - \int_0^t \mathbf{V}'R\,\mathrm{d}\mathbf{N}.$$

Because $\int_0^t (I - \mathbf{V}\mathbf{1}')\,\mathrm{d}\mathbf{N}$ is a martingale and $\text{diag}(R)' + 2(\widehat{\mathbf{M}}^-)'R - \mathbf{V}'R$ is predictable, the compensator of the process T is $-\int_0^t \mathbf{V}'R\,\mathrm{d}\mathbf{\Lambda}$, which we can estimate by

$$\hat{\mathrm{E}}T = -\int_0^t \mathbf{V}'R\,\mathrm{d}\hat{\mathbf{\Lambda}} = -\int_0^t \mathbf{V}'R\mathbf{V}\mathbf{1}'\,\mathrm{d}\mathbf{N}$$

The process $S = T - \hat{\mathrm{E}}T$ is a martingale. It can be written in the following way

$$S = \int_0^t (\text{diag}(R)' + 2(\widehat{\mathbf{M}}^- - \mathbf{V})'R)(I - \mathbf{V}\mathbf{1}')\,\mathrm{d}\mathbf{M} \qquad (3.9)$$

as the integral of the predictable process vector

$$\mathbf{K} = (\text{diag}(R)' + 2(\widehat{\mathbf{M}}^- - \mathbf{V})'R)(I - \mathbf{V}\mathbf{1}')$$

over the martingale vector $\mathbf{M}$. The predictable variation process of S is therefore $\langle S\rangle = \int_0^t \text{diag}(\mathbf{K}\mathbf{K}')'\,\mathrm{d}\mathbf{\Lambda}$, which we can estimate by

$$\widehat{\text{Var}}(T) = \int_0^t \text{diag}(\mathbf{K}\mathbf{K}')'\,\mathrm{d}\hat{\mathbf{\Lambda}} = \int_0^t \text{diag}(\mathbf{K}\mathbf{K}')'\mathbf{V}\mathbf{1}'\,\mathrm{d}\mathbf{N}$$

To evaluate $\hat{\mathrm{E}}T$ and $\widehat{\text{Var}}(T)$ we use

$$p_{ij} = \int_0^\infty \frac{e^{c_i}Y_i}{Y}\,\mathrm{d}N_j = \mathbf{1}_{\{t_i \geq t_j\}}\frac{e^{c_i}d_j}{\sum_{t_k \geq t_j} e^{c_k}}.$$

and

$$m_{ij} = \int_0^\infty \widehat{M}_i^-\,\mathrm{d}N_j = \mathbf{1}_{\{t_i < t_j\}}d_i - \sum_{k=1}^n \mathbf{1}_{\{t_k < t_j\}}p_{ik}$$

Writing P for the $n \times n$ matrix with elements p_{ij} and M for the $n \times n$ matrix with elements m_{ij}, the results (3.5) and (3.6) follow.

3.4 Interpretation

When testing a specific pathway for a specific sample of patients, it is usually not satisfactory to only report the resulting p-value. In this section we will discuss some issues related to interpretation of the test result. We show how to calculate and visualize the influence of individual genes on the test result. We also propose an diagnostic which can be used when many genes are associated with survival, to assess whether a gene group is exceptional. We only give the theory here; for an example see section 3.5.

Similarity

The test of this paper is derived from the Cox model in the same way as the Global Test in Goeman et al. (2004) was derived from the generalized linear model. The functional form of the test statistic is therefore quite similar, the martingale residuals taking the place of the residuals from the generalized linear model in that paper. Much of the interpretation of the test statistic is also quite similar.

Central to all interpretation of the test outcome is the matrix $R = XX'$ which figures prominently in the formula for the test statistic. It is an $n \times n$ matrix which can be seen as describing the similarities in expression profile between the samples. The entry R_{ij} is relatively large if samples i and j have a relatively similar expression profile over the pathway of interest.

To show the role of the matrix R, we can rewrite the unstandardized test statistic T_0 as

$$T_0 = \sum_{i=1}^{n} \sum_{j=1}^{n} R_{ij}(d_i - \hat{u}_i)(d_j - \hat{u}_j),$$

which is the sum over the term-by-term product of the entries of R and the entries of the matrix $(\mathbf{d} - \hat{\mathbf{u}})(\mathbf{d} - \hat{\mathbf{u}})'$. The i,j-th entry of the latter matrix is large whenever samples i and j have similar martingale residuals. The test statistic T_0 is, therefore, relatively large whenever the entries of the matrices R and $(\mathbf{d} - \hat{\mathbf{u}})(\mathbf{d} - \hat{\mathbf{u}})'$ are correlated, which happens when similarity in gene expressions tends to coincide with similarity in the martingale residual. Hence, the test statistic is large if individuals who die sooner than expected tend to be relatively similar in their gene expression profile and, similarly, the individuals who live longer than expected also tend to be similar in their gene expression profile.

Gene plot

To investigate the influence of individual genes on the test outcome we can rewrite $R = \sum_{i=1}^{m} \mathbf{x}_i \mathbf{x}_i'$, where $\mathbf{x}_i$ is the i-th column of X ($i = 1, \ldots, m$), contain-

ing the measurements for the i-th gene. The unstandardized test statistic then becomes

$$T_0 = \sum_{i=1}^{m} T_i$$

where $T_i = (\mathbf{d} - \hat{\mathbf{u}})'\mathbf{x}_i\mathbf{x}_i'(\mathbf{d} - \hat{\mathbf{u}})$ is exactly the unstandardized 'global' test statistic for testing whether the 'pathway' containing only gene i is associated with survival. The test statistic of a pathway is therefore a weighted average of the test statistics for the m genes in the pathway.

In a plot we can visualize the influence of the individual genes by showing the values $T_i - \hat{\mathrm{E}}T_i$, with their standard deviation under the null hypothesis (calculated using the methods of section 3.3). An example of such a 'gene plot' is given in figure 3.1. In this plot, large positive values indicate genes with a large (positive or negative) association with survival and hence genes that make the pathway more significant. As $T_i \propto \|\mathbf{x}_i\|^2$, genes with more expression variance tend to carry more weight in the pathway.

Note that the visualized values of the gene influences T_i in the gene plot are essentially univariate: they only depend on the gene i itself. The multivariate nature of the test statistic Q is therefore not visible in the gene plot. It comes in because, although T_0 is the sum of the T_i and $\hat{\mathrm{E}}T_0$ is the sum of the $\hat{\mathrm{E}}T_i$, the variance of T_0 is generally not the sum of the variances of the T_i.

The comparative p

The global test tests the null hypothesis that the pathway is not associated with survival. This null hypothesis only depends on the observed survival and on the genes in the pathway itself: the result is absolute, not relative to the other pathways.

However, there are situations in which one would be more interested in a relative result. If the global test on the set of all genes is very significant, we can usually expect a sizeable proportion of the genes on the array to be associated with survival. In that case we can expect many pathways to show association with survival as well. This will hold especially for the larger pathways, which will often include some of the genes which are associated with survival.

In such situations we propose a diagnostic called "comparative p", which can help interpret the p-value that comes out of the test. The comparative p for a pathway of size m with p-value $\bar{p}$ is defined as the proportion of randomly selected sets of genes of the size m that have an global test p-value smaller than or equal to $\bar{p}$. To calculate this comparative p we draw 1,000 or 10,000 random gene sets from the array without replacement.

The comparative p fulfills a role different from the p-value and should only be used alongside it. It is a diagnostic, not a p-value in the statistical sense.

It tells whether the p-value of a group of genes is much lower than could be expected from a gene group of its size in this data set.

3.5 Application: osteosarcoma data

We applied the above methodology to a data set of 17 osteosarcoma patients from the Leiden University Medical Center.

Data

A genome wide screen of gene expression in osteosarcoma was done using Hu133a gene expression chips (Affymetrix, Santa Clara, CA). This chip contains 22,283 genes. A successful hybridization was obtained for 17 osteosarcoma biopsies. Three of the samples were amplified, labelled and hybridized in duplicate, one sample in triplicate. These technical replicates were averaged after gene expression measures were obtained, which was done using *gcrma* (Wu et al., 2004). No pre-selection of genes was made.

The 17 patients were followed up to 10 years. Median survival time was 40 months. Available covariates included the presence of metastasis at diagnosis, histology and response to neo-adjuvant chemotherapy. However, as treatment was not uniform over all patients, these covariates were not prognostic and we did not consider them.

Pathway information was obtained from the Gene Ontology (GO Ashburner et al., 2000) database, using the BioConductor (Gentleman et al., 2004) GO package (Zhang, 2004). Pathways that were considered of specific interest were cell cycle (GO: 7049), DNA repair (GO: 6281), Angiogenesis (GO: 1525), Skeletal development (GO: 1501) and Apoptosis (GO: 6915).

Analysis

When testing pathways of interest, it is advisable to also test the 'pathway' of all genes on the chip for association with survival. This shows whether the overall gene expression profile is associated with survival. The results for the pathway of all genes and for the five pathways of primary interest are given in table 3.1. We calculated the p-value using both the asymptotic theory method and the permutation method (using 100,000 permutations).

The permutation p-values tend to be somewhat more conservative than the asymptotic p-values, reflecting both the slight loss of power for the permutation test and a deviation from asymptotic normality due to the small number of samples.

In this data set the expression profile over the set of all genes on the chip is significantly associated with survival. Note that this does *not* mean that every

TABLE 3.1: *Global Test results for the Osteosarcoma data and the pathways of primary interest. The p-values were calculated using the permutation and asymptotic method. The final column gives the comparative p (see section 3.4) .*

pathway	genes	Q	perm. p	asym. p	comp. p
All genes	22283	2.446	0.0120	0.0072	—
Cell cycle	1115	2.957	0.0042	0.0016	0.006
DNA rep.	271	3.123	0.0006	0.0009	0.011
Angiogen.	66	0.917	0.1429	0.1795	0.774
Skel. dev.	185	0.002	0.4133	0.4992	0.998
Apoptosis	656	2.533	0.0093	0.0057	0.210

gene on the chip is associated with survival. It means that the patients who die early are relatively similar to each other in terms of their overall expression profile, while patients who live long are likewise relatively similar. It also means that there is some potential for prediction of survival based on gene expression, even before any pre-selection of genes. The cell cycle, DNA repair and apoptosis pathways are clearly associated with survival, while there is no evidence for this association in angiogenesis and skeletal development.

Because the test for all genes was significant, we expect a sizeable proportion of genes to be associated with survival, so that many pathways will be associated with survival. The comparative p gives a measure whether the p-value found for the pathway is unusually low given that it is a pathway of its size from this data set (see section 3.4). For the results in table 3.1 10,000 gene sets were sampled for each pathway. We used the asymptotic p-values for the comparative p calculations.

We conclude that cell cycle and DNA repair are more clearly associated than could be expected from a gene set of its size in this data set: only around 60 out of 10,000 random gene sets of size 1,115 have a lower p-value than the cell cycle pathway. The expression profile of the apoptosis pathway is clearly associated with survival, as can be seen from the p-values; however it is not exceptional in that: more than 20% of random gene sets have a lower p-value than apoptosis. The Skeletal development pathway is interesting in its own way: it is clearly not associated with survival ($p = 0.5$) and this is quite exceptional for a pathway of this size in this data set: only around 20 in 10,000 random gene sets had a higher p-value. The skeletal development pathway seems to include uncommonly few genes which are associated with survival.

It can occur in some data sets that the set of all genes is not significant, while some pathways (eg. DNA repair) are significant. This occurs in table 3.1 for example if we use FDR-adjusted p-values with a threshold of 0.01 (Benjamini

and Yekutieli, 2001). The result for all genes can be seen as a false negative test result. However, another valid interpretation is that prediction of survival without biological pre-selection of genes is uncertain, but if it is known a priori that the genes in the DNA repair pathway are likely to be informative, some prediction of survival is possible.

Mining the GO database

If it is not a priori known which pathways are of specific interest, one can also use a data-mining approach, trying to find those pathways which are most significantly associated with survival.

For the osteosarcoma data we explored the Gene Ontology database. Of all GO terms, 4,032 matched at least one gene on the hu133a chip. We excluded all terms which matched only one gene, because the interesting single genes pathways would already have been found in single gene testing. This left 3,080 pathways, which we all tested for association with survival. We used the asymptotic p-value, because due to the randomness in the permutation p-value it does not give a unique list. Table 3.2 gives the ten GO-terms with the smallest p-values.

To adjust for multiple testing, one can use the Benjamini and Hochberg FDR (Benjamini and Yekutieli, 2001). All 10 pathways in table 3.2 are significant on an FDR of 0.05. The p-values of the pathways tend to have positive correlations because of pathway overlap and pathways being subsets of other pathways. A FDR-controlling procedure that would make use of these dependencies would potentially gain much power in this situation.

TABLE 3.2: *Global Test results for the Osteosarcoma data on 3,080 Gene Ontology pathways, showing the top 10 FDR-adjusted p-values.*

pathway	# genes	Q	FDR-adjusted p
GO:0015630	21	4.306	0.016
GO:0019932	8	4.176	0.016
GO:0045192	2	4.148	0.016
GO:0045595	17	4.060	0.016
GO:0042518	7	4.054	0.017
GO:0000158	8	3.993	0.018
GO:0040008	9	3.944	0.018
GO:0010033	10	3.844	0.023
GO:0006479	13	3.791	0.026
GO:0030111	9	3.766	0.026

The literature confirmed the importance of many of these GO-terms in tu-

morigenesis. For example, both microtubule cytoskeleton (GO:0015630) and phosphorylation of Stat3 protein (GO:0042518) are known to be involved in growth and differentiation signaling, processes which are often disturbed in tumors. Second-messenger mediated signaling (GO:0019932) is a superset of the Stat3 pathway. Protein amino acid methylation (GO:0006479) is involved in protein degradation. Alterations in the stability of proteins is often a hallmark of tumors and may affect the aggressiveness of a tumor and thereby the patient's survival.

A diagnostic plot

To learn more about the outcome of the Global Test than just the p-value one can use the diagnostic plot described in section 3.4. We illustrate the use of this plot on the microtubule cytoskeleton pathway, which emerged on top of table 3.2.

The gene plot for the 21 genes in this pathway is given in figure 3.1. Each bar gives the global test statistic for testing whether the gene set containing only that single gene is associated with survival. The test statistic for the whole pathway is a weighted average of the bars of the genes (see section 3.4). The colour of the bars distinguishes between positive and negative association with survival.

Figure 3.1 shows that only four out of 21 genes in the microtubule cytoskeleton pathway show a significant association with survival on their own. Further, the pathway is a mix of genes which are positively and negatively associated with survival. Looking more closely at the gene plot can be a basis for investigating more deeply into the structure of the pathway, perhaps to formulate hypotheses on interesting subpathways.

3.6 Discussion

It has often been remarked that the key to successful microarray data analysis lies in an intelligent integration of advanced statistical methods with the vast domain of biological knowledge that is already available. The global test for survival presented in this paper is a step forward in this direction, combining known biological pathway information with the statistical sophistication of the Cox proportional hazards model.

Due to its complexity the Cox model has been slow to find its way to microarray methodology. Most methods require survival to be reduced to a two-valued variable, using an arbitrary cut-off, resulting in unnecessary loss of information. By using the Cox model for survival, gene expression analysis can improve performance and also become better connected to traditional medical

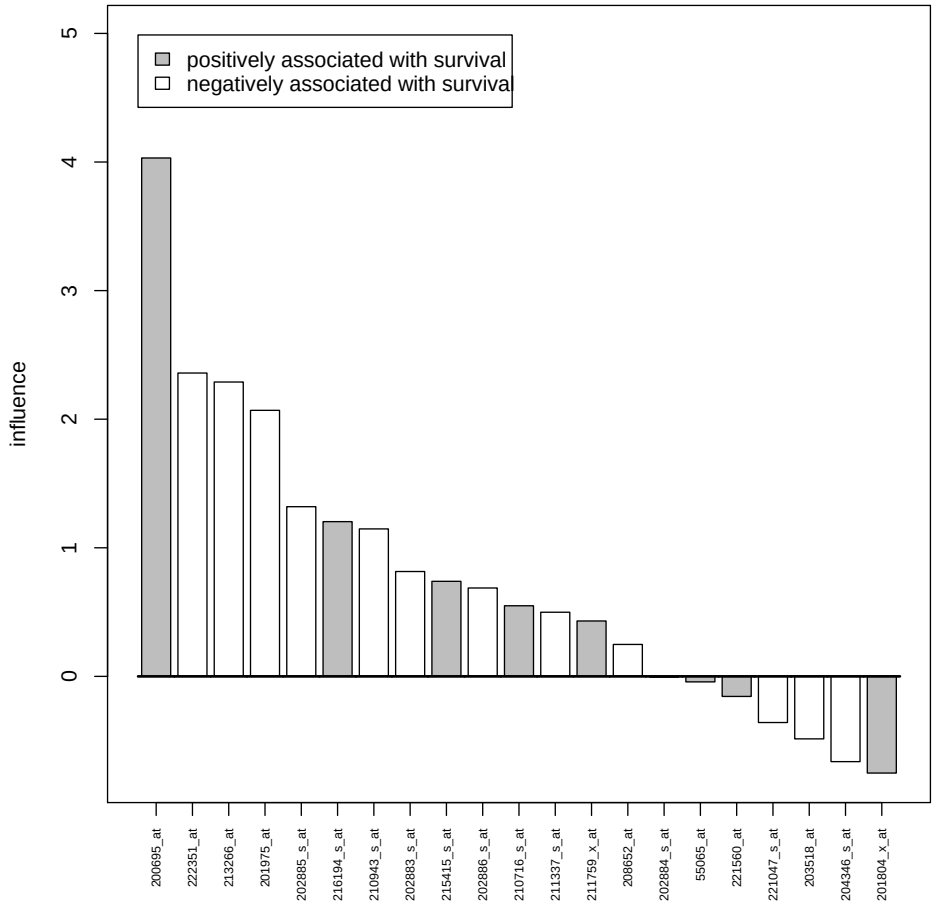

FIGURE 3.1: *Gene plot of microtubule cytoskeleton pathway, showing the sorted global test statistics for testing the 21 single gene pathways which make up the pathway.*

statistics.

Pathway information is available from many databases and is essential for the understanding of the outcomes of a microarray experiment. The Global Test methodology allows researchers to look directly for important pathways, without first having to go through single gene testing. This may lead to a better use of pathway information and more directly interpretable results.

CHAPTER 4

A goodness-of-fit test for multinomial logistic regression

Abstract

This paper presents a score test to check the fit of a logistic regression model with two or more outcome categories. The null hypothesis that the model fits well is tested against the alternative that residuals of samples close to each other in covariate space tend to deviate from the model in the same direction. We propose a test statistic that is a sum of squared smoothed residuals, and show that it can be interpreted as a score test in a random effects model. By specifying the distance metric in covariate space, users can choose the alternative against which the test is directed, making it either an omnibus goodness-of-fit test or a test for lack of fit of specific model variables or outcome categories.

4.1 Introduction

The multinomial logistic regression model is a generalization of logistic regression to outcomes with more than two levels. The model is also known as polytomous or polychotomous logistic regression in the health sciences and as the discrete choice model in econometrics (Hosmer and Lemeshow, 2000). Two variants exist: one for nominal and one for ordinal scale outcomes. This paper considers only the nominal scale version.

When fitting a model it is important to have tools to test for lack of fit. This is especially important for the multinomial logistic model, whose fit is notoriously difficult to visualize. The modelling toolbox should include general tests for the fit of the whole model, but also more specific tests for lack of fit in specific

This chapter will appear as: J. J. Goeman and S. le Cessie (2006). A goodness-of-fit test for multinomial logistic regression. *Biometrics* **62**, in press. The definitive version will be available at `http://www.blackwell-synergy.com`.

covariates or outcome categories. Such tools are remarkably scarce in multinomial logistic regression. Hosmer and Lemeshow (2000) suggested looking at the multinomial model as if it were a set of independent ordinary logistic models of each outcome against the reference outcome, and testing the fit of each of these separately. Lesaffre and Albert (1989) give diagnostics for detecting influential, leverage and outlying samples in multinomial logistic regression, but provided no explicit goodness-of-fit test. The only actual test for the fit of the multinomial logistic regression model is given by Pigeon and Heyse (1999). It is an extension of the test of Hosmer and Lemeshow (2000) for binary regression, which is well known to have low power for detecting quadratic effects (Le Cessie and Van Houwelingen, 1991).

In this paper we present an alternative and flexible goodness-of-fit test for the multinomial logistic regression model. It can be directed against the general alternative that the model does not fit or against more specific alternatives. The test extends the goodness-of-fit test of Le Cessie and Van Houwelingen (1991) for ordinary logistic regression to the multinomial case. The approach is to smooth the regression residuals and to test whether these smoothed residuals have more variance than expected under the null hypothesis, which occurs when residuals which are close together in the covariate space are correlated. This type of test was shown by Le Cessie and van Houwelingen (1995) to be equivalent to a score test in a random effects model, which tests for the presence of a pre-specified correlation structure between the residuals. Their approach to goodness-of-fit testing is quite generally applicable, and has already been extended to generalized linear models (Le Cessie and van Houwelingen, 1995) and to the Cox proportional hazards model (Verweij et al., 1998). This paper extends the methodology to multinomial logistic regression.

The properties of the resulting test are verified using simulated data and illustrated on a liver enzyme data set (Albert and Harris, 1987). Software in *R* for fitting and testing the fit of the model is available on request from the authors.

4.2 The multinomial logistic regression model

Suppose the multinomial outcome variable Y takes values in the unordered set $\{1, \ldots, g\}$. The multinomial logistic regression model assumes that the probability for observation i to have outcome s depends on i's covariates $x_{i1}, \ldots, x_{ip}$ as

$$P(Y_i = s) = \frac{e^{\eta_{is}}}{\sum_{t=1}^{g} e^{\eta_{it}}} \tag{4.1}$$

where $\eta_{is} = \sum_{k=1}^{p} x_{ik}\beta_{ks}$ is a linear predictor. In this formulation of the model we have a regression coefficient β_{ks} for each combination of covariate k and outcome category s, and a separate linear predictor η_{is} for each outcome category (for a more detailed description of the model, see Hosmer and Lemeshow, 2000).

The model as defined in (4.1) is overparametrized. Replacing $(\beta_{k1}, \ldots, \beta_{kg})$ with $(\beta_{k1} + c, \ldots, \beta_{kg} + c)$, for any $c \in \mathbb{R}$ and $k \in \{1, \ldots, p\}$, leads to exactly the same probabilities. The most common way to solve this overparametrization is to designate one outcome category, say outcome 1, as the "reference" category, setting all regression coefficients $\beta_{11}, \ldots, \beta_{p1}$ to zero. A good choice of the reference category will usually facilitate interpretation of the resulting parameter estimates. However, in this paper we are not concerned with estimation but rather with assessment of the fit, which does not depend on the choice of the reference category. We will therefore refrain from choosing a reference category, but instead treat the outcome categories symmetrically, leaving the model overparametrized.

Suppose we have sampled outcomes $Y_1, \ldots, Y_n$ and a corresponding $n \times p$ design matrix X. Then let y_{is} be the indicator of the event $\{Y_i = s\}$, for $i = 1, \ldots, n$, and $s = 1, \ldots, g$, and call the corresponding probabilities $\mu_{is} = \mathrm{P}(Y_i = s)$. Let $\hat{\mu}_{is}$ be the maximum likelihood estimates of μ_{is} for all i and s. The fitted model has $n \times g$ residuals $\hat{r}_{is} = y_{is} - \hat{\mu}_{is}$, one for each individual i and outcome category s. These residuals fulfill $\sum_{s=1}^{g} \hat{r}_{is} = 0$. It will be convenient to gather the residuals together in a long vector $\hat{\mathbf{r}} = \mathbf{y} - \hat{\boldsymbol{\mu}}$, where $\mathbf{y} = (y_{11}, \ldots, y_{n1}, \ldots, y_{1g}, \ldots, y_{ng})'$ and $\boldsymbol{\mu} = (\mu_{11}, \ldots, \mu_{n1}, \ldots, \mu_{1g}, \ldots, \mu_{ng})'$.

4.3 Testing goodness-of-fit by smoothing

A goodness-of-fit test tests a model against the alternative that the model 'does not fit'. This is an extremely broad class of alternatives: lack of fit comes in many different shapes and sizes. A linear model, for example, can display lack of fit when the distribution of the residuals is skewed or heavy-tailed, or when there are non-linear relationships which fit the data better. Typically, there is no single goodness-of-fit test which has good power against all kinds of lack of fit. For better interpretation, a goodness-of-fit test should therefore be specific about the type of lack of fit is directed against.

The goodness-of-fit test of this paper is directed against the alternative that any non-linearities or interaction effects have been missed. Such neglected effects can be detected by looking for patterns in the residuals: observations close to each other in covariate space which deviate from the model in the same direction. One looks for this same kind of behaviour when making a scatterplot

of the residuals against a covariate. The test can also detect different kinds of lack of fit which show up as patterns of correlation in the residuals, such as over-dispersion.

One can formally test for patterns in the residuals by smoothing the residuals: the smoothed residuals are a weighted average of the residual itself and the other residuals which are close to it in covariate space. If residuals close to each other are strongly correlated, the smoothing will not affect the magnitude of the residuals much, while if they are not correlated smoothing will shrink the residuals toward zero. The sum of squares of the smoothed residuals is therefore a good measure of the correlations of residuals close to each other in covariate space (Le Cessie and Van Houwelingen, 1991).

Based on these arguments we propose to reject for large values of the test statistic

$$Q = \sum_{s=1}^{g} \sum_{i=1}^{n} \Big[\sum_{j=1}^{n} u_{ij}(y_{js} - \hat{\mu}_{js}) \Big]^2 \tag{4.2}$$

where $u_{ij} \geq 0$ is the i,j-th entry of a smoothing matrix U, fulfilling $\sum_{j=1}^{n} u_{ij} = 1$ for all i. The statistic Q is a sum of squared smoothed residuals, as each $\tilde{r}_{is} = \sum_{j=1}^{n} u_{ij}(y_{js} - \hat{\mu}_{js})$ is a smoothed version of the residual $\hat{r}_{is}$. Note that smoothing of the residual $\hat{r}_{is}$ only involves residuals of the same outcome category s, as the residuals corresponding to different categories are not expected to be positively correlated.

There are various possibilities for the choice of the smoothing matrix U. This choice has two aspects: the choice of a distance measure and the choice of a smoothing method. Of these two, the choice of distance measure deserves most consideration. To test globally for lack of fit one could take euclidian distance using all covariates. As euclidian distance is sensitive to the scaling of the variables, it is wise to rescale the variables to unit variance to prevent one covariate dominating the distance measure. If, on the other hand, the interest is in testing lack of fit for a specific subset of the covariates, one should only use that subset for constructing the distance measure. The choice of a smoothing method is less of an issue. Let d_{ij} be the chosen distance between observations i and j. Following Le Cessie and van Houwelingen (1995) one could choose a kernel smoother based on this distance. A convenient choice is the uniform kernel which has $K(t) = 1$ if $-1 \leq t \leq 1$, and $K(t) = 0$ otherwise. The resulting smoothing matrix U will have entries

$$u_{ij} = \frac{K(d_{ij}/h)}{\sum_{k=1}^{n} K(d_{ik}/h)}.$$

Here h is the bandwidth, which should be chosen carefully: taking h too large results in oversmoothing, while taking h too small results in undersmoothing.

Both will lead to low power. The choice of h can be related to the distribution of the distances d_{ij}, $i \neq j$: our experience is that taking h as the 25-th percentile of this distribution is a often good choice. Using a kernel smoother, the smoothed residual $\tilde{r}_{is}$ will be the average of all residuals $\hat{r}_{js}$ with $d_{ij} \leq h$.

4.4 Distribution of the test statistic

To be able to use the test statistic Q for testing we must calculate or approximate its distribution function.

Write $\mathbf{U} = I_g \otimes U$, where $\otimes$ denotes the Kronecker product and I_g the $g \times g$ identity matrix, and write $\mathbf{R} = \mathbf{U}\mathbf{U}'$. Then we can write $\tilde{\mathbf{r}} = \mathbf{U}'\hat{\mathbf{r}}$ and

$$Q \;=\; \|\tilde{\mathbf{r}}\|^2 \;=\; (\mathbf{y} - \hat{\boldsymbol{\mu}})'\mathbf{R}(\mathbf{y} - \hat{\boldsymbol{\mu}}),$$

which is a non-negative quadratic form.

There is no exact expression for the null distribution function of Q, but there are various approaches for finding an approximation. The most promising approach follows asymptotic arguments. Assuming that as n grows new observations are added which have the same covariate patterns as those already present, it can be shown that Q converges in distribution to a linear combination of chi-squared variables with one degree of freedom. There is no simple explicit expression for the distribution function of a such a distribution, but it is known that it can be well approximated by a general scaled chi-squared (or gamma) distribution. This is often used as an approximate distribution for quadratic forms (Cox and Hinkley, 1974, p. 462–463), although more accurate approximations exist (Solomon and Stephens, 1978). The gamma approximation was also used for the test of Le Cessie and van Houwelingen (1995) which this paper extends. It should be calibrated to have the same mean and variance as Q as well as to the fact that $Q \geq 0$, resulting in a gamma distribution with parameters $\alpha = (\mathrm{E}Q)^2/\mathrm{Var}(Q)$ and $\lambda = \mathrm{E}Q/\mathrm{Var}(Q)$. The accuracy of this approximation will be checked with a simulation example in section 4.7.

To use this approximation we have to calculate expectation and variance of Q. This involves the distribution of the estimated residuals $\mathbf{y} - \hat{\boldsymbol{\mu}}$, which can be related to the easier distribution of the true residuals $\mathbf{y} - \boldsymbol{\mu}$ through its first order approximation, using standard theory from generalized linear models (McCullagh and Nelder, 1989). If n is not too small,

$$\mathbf{y} - \hat{\boldsymbol{\mu}} \approx (I - \mathbf{H})(\mathbf{y} - \boldsymbol{\mu}) \tag{4.3}$$

where $\mathbf{H}$ is the asymmetric form of the hat matrix for the multinomial logistic regression model. It is defined as $\mathbf{H} = \mathbf{W}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-}\mathbf{X}'$, where $\mathbf{X} = I_g \otimes X$,

superscript minus denotes a generalized inverse, and $\mathbf{W}$ is given by

$$\mathbf{W} = \begin{pmatrix} W^{11} & W^{12} & \dots & W^{1g} \\ W^{21} & W^{22} & & \vdots \\ \vdots & & \ddots & \\ W^{g1} & \dots & & W^{gg} \end{pmatrix}, \tag{4.4}$$

where each W^{ij} is an $n \times n$ diagonal matrix with

$$\text{diag}(W^{st}) = \text{diag}(W^{ts}) = \begin{cases} (-\mu_{1s}\mu_{1t}, \dots, -\mu_{ns}\mu_{nt})' & \text{if } s \neq t \\ (\mu_{1s}(1-\mu_{1s}), \dots, \mu_{ns}(1-\mu_{ns}))' & \text{if } s = t \end{cases}$$

The hat matrix $\mathbf{H}$ also plays an important role in the paper of Lesaffre and Albert (1989), where it is used to detect influential observations. From the approximation (4.3) it follows that if n is not too small, the distribution of Q is approximately the same as the distribution of

$$\tilde{Q} = (\mathbf{y} - \boldsymbol{\mu})'\tilde{\mathbf{R}}(\mathbf{y} - \boldsymbol{\mu}),$$

where $\tilde{\mathbf{R}} = (I - \mathbf{H})'\mathbf{R}(I - \mathbf{H})$.

Under the null hypothesis, $\text{E}[(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})'] = \mathbf{W}$, so that

$$\text{E}\tilde{Q} = \text{trace}(\tilde{\mathbf{R}}\mathbf{W}).$$

The variance under H_0 of Q is calculated in Section 4.10. It is given by

$$\text{Var}(\tilde{Q}) = 2\text{trace}(\tilde{\mathbf{R}}\mathbf{W}\tilde{\mathbf{R}}\mathbf{W}) + \sum_{s=1}^{g}\sum_{t=1}^{g}\sum_{u=1}^{g}\sum_{v=1}^{g}\sum_{i=1}^{n} \tilde{R}_{ii}^{st}\tilde{R}_{ii}^{uv}\kappa_i^{stuv} \tag{4.5}$$

In this expression, $\tilde{R}_{ij}^{st}$ is the i,j-th element of the submatrix $\tilde{R}^{st}$ of $\tilde{\mathbf{R}}$, which is similarly decomposed as $\mathbf{W}$ in (4.4). The value of κ_i^{stuv} does not depend on the order of s, t, u and v: it can be calculated with

$$\begin{aligned} \kappa_i^{ssss} &= \mu_{is} - 7\mu_{is}^2 + 12\mu_{is}^3 - 6\mu_{is}^4 \\ \kappa_i^{ssst} &= -\mu_{it}\mu_{is} + 6\mu_{it}\mu_{is}^2 - 6\mu_{it}\mu_{is}^3 \\ \kappa_i^{sstt} &= -\mu_{is}\mu_{it} + 2\mu_{is}\mu_{it}^2 + 2\mu_{is}^2\mu_{it} - 6\mu_{is}^2\mu_{it}^2 \\ \kappa_i^{sstu} &= 2\mu_{is}\mu_{it}\mu_{iu} - 6\mu_{is}^2\mu_{it}\mu_{iu} \\ \kappa_i^{stuv} &= -6\mu_{is}\mu_{it}\mu_{iu}\mu_{iv}, \end{aligned} \tag{4.6}$$

after recoding s, t, u and v to denote unique outcomes.

The mean and variance of Q involve the unknown vector $\boldsymbol{\mu}$, which should be estimated by its maximum likelihood estimate $\hat{\boldsymbol{\mu}}$ in applications.

4.5 Testing for the presence of a random effect

The test proposed in Section 4.3 was motivated by heuristic arguments. These arguments give a good impression of the type of alternative the test can be expected to have good power against, but the alternative was not yet precisely specified. In this section we present a fully specified alternative model from which the goodness-of-fit test proposed in Section 4.3 can be derived as a score test. This model explicitly lets observations which are close to each other in covariate space have correlated residuals.

We propose to add an extra random effect z_{is} to the linear predictor η_{is} for each combination of observation i and outcome category s. Given the random effect, the distribution of Y becomes

$$\mathrm{P}(Y_i = s \mid \mathbf{z}) = \frac{e^{\eta_{is}+z_{is}}}{\sum_{t=1}^{g} e^{\eta_{it}+z_{it}}} \tag{4.7}$$

where $\mathbf{z} = (z_{11}, \ldots, z_{n1}, \ldots, z_{1g}, \ldots, z_{ng})'$ is the vector of all random effects. We do not specify a distributional form for $\mathbf{z}$, but we specify its first and second moments as $\mathrm{E}(\mathbf{z}) = \mathbf{0}$ and $\mathrm{Var}(\mathbf{z}) = \tau^2\mathbf{R}$, where τ^2 is an unknown parameter. The matrix $\mathbf{R} = \mathbf{U}\mathbf{U}'$ here is the same matrix as defined in section 4.4. It can be written $\mathbf{R} = I_g \otimes R$ where $R = UU'$. Let R_{ij}^{st} be the element of $\mathbf{R}$ corresponding to the covariance of the random effects z_{is} and z_{jt}. If U is a smoothing matrix, R_{ij}^{st} is positive when $s = t$ and the distance d_{ij} is small, and zero otherwise. For example, when using a uniform kernel with bandwidth h, $R_{ij}^{st} > 0$ if $s = t$ and there is a k such that $d_{ki} \leq h$ and $d_{kj} \leq h$; R_{ij}^{st} is zero otherwise. If $\tau^2 > 0$, the presence of the random effect causes extra variation in the regression residuals with a covariance structure similar to $\mathbf{R}$: correlated random effects cause correlated residuals. Therefore, if $\tau^2 > 0$ observations which are close to each other tend to have correlated residuals.

The null hypothesis that the multinomial logistic regression model fits well can be phrased in terms of the above random effects model (4.7) as

$$H_0 : \ \tau^2 = 0,$$

which implies $\mathbf{z} = \mathbf{0}$, against the one-sided alternative

$$H_A : \ \tau^2 > 0.$$

We test H_0 with a score test. An advantage of score testing is that it only requires fitting the model under the null hypothesis, not under the alternative hypothesis. This is an important advantage for our H_A, because the random effects model (4.7) is difficult to fit. Furthermore, the score test is by definition

a one-sided test, so problems due to a null hypothesis on the boundary of the parameter space do not arise.

The score test statistic is the derivative of the loglikelihood $\ell(\tau^2)$ with respect to τ^2 at $\tau^2 = 0$. If nuisance parameters are present, as in this case the model parameters $\boldsymbol{\beta}$, the loglikelihood is replaced by the profile loglikelihood $\hat{\ell}(\tau^2) = \ell(\tau^2, \hat{\boldsymbol{\beta}}(\tau^2))$. We have

$$\frac{\partial \hat{\ell}}{\partial \tau^2} = \frac{\partial \ell}{\partial \tau^2} + \frac{\partial \ell}{\partial \boldsymbol{\beta}} \cdot \frac{\partial \hat{\boldsymbol{\beta}}}{\partial \tau^2}.$$

As $\partial \ell / \partial \boldsymbol{\beta}$ is zero if $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, the score test statistic of the profile likelihood is simply the score test statistic of the ordinary likelihood with maximum likelihood estimates of the nuisance parameters under the null plugged in.

The loglikelihood of the general model (4.7) is given by

$$\ell(\tau^2) = \log \Big[\mathrm{E}_{\mathbf{z}} \Big\{ \exp \Big(\sum_{i=1}^{n} \sum_{s=1}^{g} y_{is} \log\{\nu_{is}(\mathbf{z})\} \Big) \Big\} \Big], \tag{4.8}$$

where $\nu_{is}(\mathbf{z}) = \mathrm{P}(Y_i = s \mid \mathbf{z})$ and $\mathrm{E}_{\mathbf{z}}$ denotes taking the expectation over $\mathbf{z}$. In Section 4.11 we calculate the derivative of this likelihood with respect to τ^2 at $\tau^2 = 0$, in the spirit of Le Cessie and van Houwelingen (1995), using a Taylor approximation of $\nu_{is}(\mathbf{z})$ with respect to $\mathbf{z}$ at $\mathbf{z} = \mathbf{0}$. This results in the score test statistic

$$T = \frac{\partial \hat{\ell}(0)}{\partial \tau^2} = \tfrac{1}{2}(\mathbf{y} - \hat{\boldsymbol{\mu}})' \mathbf{R} (\mathbf{y} - \hat{\boldsymbol{\mu}}) - \tfrac{1}{2} \mathrm{trace}(\mathbf{R}\hat{\mathbf{W}}). \tag{4.9}$$

We see that the score test statistic in this model is equivalent to the test statistic proposed in (4.2), as, ignoring the constants, T is simply Q minus the estimated expectation of Q.

This alternative construction of Q as a score test statistic gives interesting insights in the power properties of the test. A score test is a locally most powerful test (Cox and Hinkley, 1974) in the sense that it optimizes the slope of the power function at the test value of $\tau^2 = 0$. It is therefore the optimal test to use against the alternative model (4.7) when the value of τ^2 is small. These alternatives tend to have small, but non-zero values of the random effect $\mathbf{z}$. The goodness-of-fit test proposed in this paper is therefore the optimal test for detecting a small, but consistent deviation from the model.

The random effects model of this section is interesting in its own right as a general test for the existence of a random effect with a specified covariance structure $\mathbf{R}$, which may be any positive semi-definite matrix. This type of test has many applications outside the context of goodness-of-fit testing, for example in variance components analysis in genetics (Houwing-Duistermaat et al., 1995) and in high-dimensional data analysis in genomics (Goeman et al., 2004).

4.6 Connection to binary logistic regression

Here we show that for $g = 2$, when multinomial logistic regression becomes binary logistic regression, the test in this paper is exactly the same as the goodness-of-fit test of Le Cessie and van Houwelingen (1995), so that it is a generalization of that test.

Take $g = 2$. Call $R = UU'$, $W = W^{11}$, as defined in (4.4), and $H = WX(X'WX)^{-1}X'$. Call $\mathbf{y}_1 = (y_{11}, \ldots, y_{n1})'$, $\boldsymbol{\mu}_1 = (\mu_{11}, \ldots, \mu_{n1})'$, using the notation of Section 4.2. Then the test statistic of Le Cessie and van Houwelingen (1995) is given by

$$Q_1 = (\mathbf{y}_1 - \hat{\boldsymbol{\mu}}_1)'R(\mathbf{y}_1 - \hat{\boldsymbol{\mu}}_1)$$

To show that this test statistic is equivalent to the test statistic in this paper for $g = 2$, remark that $\mathbf{y} - \hat{\boldsymbol{\mu}} = \mathbf{f} \otimes (\mathbf{y}_1 - \hat{\boldsymbol{\mu}}_1)$ where $\mathbf{f} = (1, -1)'$. Combining this with $\mathbf{R} = I_g \otimes R$, it follows that

$$Q = \mathbf{f}'\mathbf{f} \otimes (\mathbf{y}_1 - \hat{\boldsymbol{\mu}}_1)'R(\mathbf{y}_1 - \hat{\boldsymbol{\mu}}_1) = 2Q_1.$$

The test statistics are therefore equivalent.

To show that also the approximations to the distribution of the test statistic are the same, we must show that also $\tilde{Q} = 2\tilde{Q}_1$, where

$$\tilde{Q}_1 = (\mathbf{y}_1 - \hat{\boldsymbol{\mu}}_1)'(I - H)'R(I - H)(\mathbf{y}_1 - \hat{\boldsymbol{\mu}}_1)$$

This can be shown by remarking that $\mathbf{W} = F \otimes W$, where

$$F = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

Writing $\mathbf{X} = I \otimes X$, remarking that F has generalized inverse $F^- = (1/4)F$ and expanding the Kronecker products, we get $\mathbf{H} = (1/2)F \otimes H$, from which $(I - \mathbf{H})(\mathbf{y} - \boldsymbol{\mu}) = \mathbf{f} \otimes (I - H)(\mathbf{y}_1 - \boldsymbol{\mu}_1)$. Finally, combining this with $\mathbf{R} = I \otimes R$, the result $\tilde{Q} = 2\tilde{Q}_1$ follows. Therefore the two test statistics and the approximations to their distribution are completely equivalent in case of binary logistic regression.

4.7 Simulation results

To check the adequacy of the gamma approximation to the distribution of Q in a concrete case and to give an illustration of the power of the test, we conducted a small simulation experiment (compare Le Cessie and van Houwelingen, 1995, for the case $g = 2$).

We constructed a data set of 108 observations and three covariates x_1, x_2 and x_3, each taking values 1, 0 and -1. The 108 observations were taken as

TABLE 4.1: *Fraction rejected for the goodness-of-fit test of this paper, based on 10,000 simulated data sets under the null hypothesis ($t = 0$) and under alternatives with a quadratic effect ($t > 0$).*

alternative	*nominal test size* α				
	0.10	0.05	0.01	0.005	0.001
$t = 0$	0.125	0.061	0.014	0.007	0.002
$t = 1$	0.243	0.148	0.046	0.026	0.009
$t = 2$	0.618	0.487	0.259	0.189	0.088
$t = 3$	0.882	0.800	0.581	0.485	0.300
$t = 4$	0.979	0.954	0.844	0.781	0.606

four replicates from each of the 27 possible combinations of the three covariate values. We modelled the probabilities of three possible outcomes as in (4.1) with

$$\begin{aligned} \eta_1 &= 2x_1 + tx_1^2 \\ \eta_2 &= 2x_2 \\ \eta_3 &= 2x_3 \end{aligned}$$

By varying the value of t we can generate outcomes both from the null hypothesis that a multinomial logistic regression model in x_1, x_2 and x_3 fits well, and various alternative hypotheses.

We generated 10,000 multinomial outcome vectors Y from the model, taking $t = 0$, 1, 2, 3 and 4. For each realisation of Y we fitted a multinomial logistic regression model in x_1, x_2 and x_3 and calculated the goodness-of-fit test statistic Q, estimated its expectation and variance, and calculated the p-value using the gamma approximation. The smoothing matrix U was constructed using a uniform kernel with a bandwidth at the 25-th percentile of the distance distribution, which meant that each smoothed residual was the average of all residuals at most $\sqrt{2}$ distance away. The results are given in table 4.1, rounded to three decimal places.

Judging from this table, it seems that the gamma approximation to the distribution of the test statistic performs quite well, although it is slightly anti-conservative. The rejection rates for $t = 0$ are close to the nominal α level. It can also be concluded that the goodness-of fit test has good power for detecting deviations from the null hypothesis. It would be interesting to look at the effect of different choices of the bandwidth and to study different alternatives, but we lack space in this paper.

TABLE 4.2: *The p-values of the goodness-of-fit test for the liver enzyme data, with different choices of bandwidth (measured as percentiles of the distribution of distances between observations).*

model	*bandwidth (percentile)*						
	10	20	30	40	50	60	70
non-log-transformed	0.004	0.001	0.000	0.000	0.013	0.022	0.091
log-transformed	0.491	0.576	0.341	0.297	0.579	0.580	0.397

4.8 Application: liver enzyme data

We applied the goodness-of-fit test to a dataset of patients with liver disease (Albert and Harris, 1987). This data set has 218 patients in four disease categories: acute viral hepatitis (57 patients), persistent chronic hepatitis (44), aggressive chronic hepatitis (40) and post-necrotic cirrhosis (77). For each patient the concentrations of three liver enzymes was measured: aspartate aminotransferase (AST), alanine aminotransferase (ALT) and glutamate dehydrogenase (GLDH). All these variables had markedly skewed distributions. The data were analyzed with a multinomial logistic regression model by Albert and Harris (1987), but Lesaffre and Albert (1989) argued for a multinomial logistic regression model with log-transformed covariates.

We tested the fit of the model with AST, GLDH and ALT using the goodness-of-fit test of this paper and kernel smoothing using a uniform kernel with bandwidth equal to the 25-th percentile of the distribution of the distances between observations. We found $Q = 8.41$ with $\mathrm{E}Q = 2.78$ and $\mathrm{sd}(Q) = 1.27$. On a scaled chi-squared distribution with 9.52 degrees of freedom (gamma{4.76, 1.71}), this gave a p-value of 0.001, clearly indicating lack of model fit. Log-transforming the covariates before fitting the model gives a clearly non-significant p-value of 0.37.

To investigate the sensitivity of this result to the choice of the smoothing method, we calculated the p-value for different choices of the bandwidth parameters (table 4.2). Bandwidth values are given as percentiles of the distribution of distances between the observations. From table 4.2 it can be seen that the test is quite robust to the choice of bandwidth.

There are various ways of making use of the flexibility of the goodness-of-fit test of this paper for looking more closely into a more significant test result. One is to break down the omnibus test for all variables to see which variables are responsible for the lack of fit. This can be done by using subsets of the original covariates AST, ALT and GLDH for constructing the distance measure for use by the test, testing whether the relationship between that subset of the covari-

TABLE 4.3: *Results of goodness-of-fit test of the liver enzyme data in which the distance measure between observations depends on different subsets of the covariates. The table gives raw p-values and multiplicity adjusted p-values from a closed testing procedure.*

Distance based on	*p-value*	*adjusted p-value*
AST, ALT and GLDH	0.001	0.001
AST and GLDH	0.003	0.003
AST and ALT	0.001	0.001
GLDH and ALT	0.000	0.001
AST	0.000	0.003
GLDH	0.314	0.314
ALT	0.001	0.001

ates and the outcome has been adequately modelled. Taking all $2^3 - 1$ subsets, we can set up a closed testing procedure (Marcus et al., 1976) to control for multiple testing. In this procedure each subset of covariates is only tested when all its supersets are significant (for example the subset {ALT} is only tested when tests based on the subsets {AST, ALT}, {GLDH, ALT} and {GLDH, ALT, AST} are all significant). In that case all tests can be performed at level α, while still keeping the family-wise error rate at α (Marcus et al., 1976). The multiplicity adjusted p-values (Dudoit et al., 2003) for this procedure are the maximum of the p-values of the test itself and all supersets. These multiplicity-adjusted p-values are never smaller than the p-value for the test for global lack of fit. We performed these tests using kernel smoothing with a bandwidth at the 25-th percentile of the distance distribution as above. The raw and multiplicity adjusted p-values are given in table 4.3. The lack of fit is most clear in ALT and AST, while there is no evidence for lack of fit in GLDH. This is in line with the analysis of Lesaffre and Albert (1989), who concluded that there was no real need to log-transform GLDH.

Just as the test result can be split up in its component variables, it can be split into its component outcome categories. The test statistic can be written as

$$Q = \sum_{s=1}^{g} Q_s$$

where Q_s is the sum of the squared smoothed residuals $\tilde{r}_{1s}, \ldots, \tilde{r}_{ns}$, corresponding to outcome category s. We plotted the Q_s, $s = 1, \ldots, 4$ in figure 4.1, standardized to z-scores. From the plot we can see that the lack of fit is clear in the residuals of categories 1, 2 and 3 (acute viral, persistent chronic and aggressive chronic hepatitis), but that there is no clear evidence for lack of fit in the residuals of category 4 (post-necrotic cirrhosis).

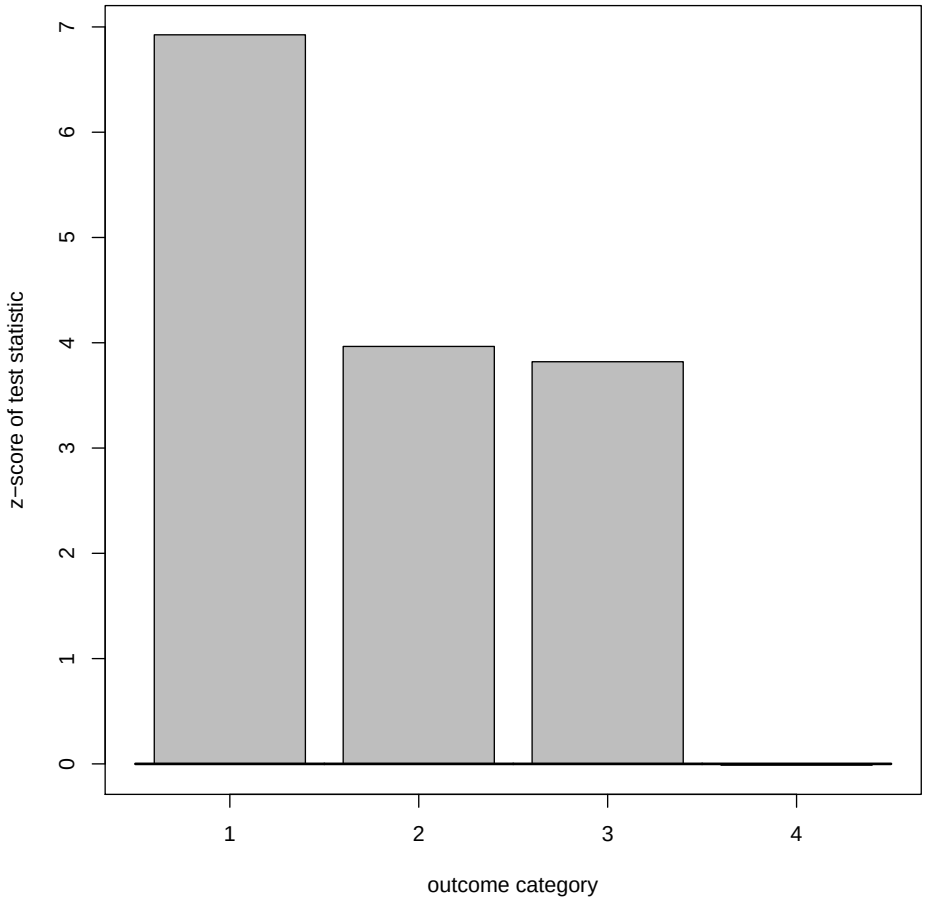

FIGURE 4.1: *Influence of the four outcome categories on the goodness-of-fit test result. Depicted are the sum of squared smoothed residuals Q_s of each outcome category s, standardized to z-scores. The total goodness-of-fit test statistic is the sum of the unstandardized Q_s-scores.*

4.9 Discussion

Formal goodness-of-fit testing is important in model-building of the multinomial logistic regression model, because the fitted model is very difficult to visualize. So far, however, only one goodness-of-fit test was available for this model (Pigeon and Heyse, 1999), which stands in the tradition of the goodness-of-fit test of Hosmer and Lemeshow (2000) for binary logistic regression. In this paper we have presented a very different goodness-of-fit test based on a sum of squared smoothed residuals, extending a test of Le Cessie and Van Houwelin-

gen (1991). It has power against consistent patterns of non-linearity: observations close to each other in covariate space which deviate in the same direction.

To illustrate the power properties of the test, we have constructed a random effects model for which the proposed test is optimal. Such a precise specification of the alternative hypothesis against which the test is optimal clarifies the type of lack of fit the test is directed against and therefore gives some insight into its power properties. Tools were also provided to look more closely into a significant test result.

Just like the test of Le Cessie and van Houwelingen (1995), the test proposed in this paper has potential applications outside the goodness-of-fit testing context, for example in genetics (Houwing-Duistermaat et al., 1995) and in high-dimensional data analysis (Goeman et al., 2004). This paper allows these applications to be generalized to the case of multinomial outcome variables.

4.10 Variance of the test statistic

We calculate the variance of $\tilde{Q}$ as given in (4.5). Write $\tilde{Q} = \sum_{s=1}^{g}\sum_{t=1}^{g} Q_{st}$, where $Q_{st} = \sum_{i=1}^{n}\sum_{j=1}^{n} \tilde{R}_{ij}^{st}(y_{is}-\mu_{is})(y_{jt}-\mu_{jt})$ for $s,t = 1,\ldots g$. We will calculate the g^4 covariances of all Q_{st} terms and sum them to find the variance of $\tilde{Q}$.

Define $S_{ij}^{st} = (y_{is}-\mu_{is})(y_{jt}-\mu_{jt})$. Then $\mathrm{Cov}(S_{ij}^{st}, S_{kl}^{uv}) = 0$ unless $i=k$ and $j=l$ or $i=l$ and $j=k$, due to the independence of the samples under the null hypothesis. Therefore

$$\begin{aligned}\mathrm{Cov}(Q_{st}, Q_{uv}) &= \sum_i \tilde{R}_{ii}^{st}\tilde{R}_{ii}^{uv}\mathrm{Cov}(S_{ii}^{st}, S_{ii}^{uv}) + \sum_i\sum_{j\neq i}\tilde{R}_{ij}^{st}\tilde{R}_{ij}^{uv}\mathrm{Cov}(S_{ij}^{st}, S_{ij}^{uv}) \\ &\quad + \sum_i\sum_{j\neq i}\tilde{R}_{ij}^{st}\tilde{R}_{ji}^{uv}\mathrm{Cov}(S_{ij}^{st}, S_{ji}^{uv}).\end{aligned}$$

If $i \neq j$, $\mathrm{E}(S_{ij}^{st}) = 0$, for all s and t, so that

$$\begin{aligned}\mathrm{Cov}(S_{ij}^{st}, S_{ij}^{uv}) &= \mathrm{E}[(y_{is}-\mu_{is})(y_{iu}-\mu_{iu})]\cdot\mathrm{E}[(y_{jt}-\mu_{jt})(y_{jv}-\mu_{jv})] \\ &= W_{ii}^{su}W_{jj}^{tv},\end{aligned}$$

while if $i=j$,

$$\mathrm{Cov}(S_{ii}^{st}, S_{ii}^{uv}) = \mathrm{E}[S_{ii}^{st}S_{ii}^{uv}] - \mathrm{E}[S_{ii}^{st}]\mathrm{E}[S_{ii}^{uv}] = \mathrm{E}[S_{ii}^{st}S_{ii}^{uv}] - W_{ii}^{st}W_{ii}^{uv}.$$

Using these expressions,

$$\begin{aligned}\mathrm{Cov}(Q_{st}, Q_{uv}) &= \sum_{i=1}^{n}\tilde{R}_{ii}^{st}\tilde{R}_{ii}^{uv}\kappa_i^{stuv} + \sum_{i=1}^{n}\sum_{j=1}^{n}\tilde{R}_{ij}^{st}\tilde{R}_{ji}^{vu}W_{ii}^{su}W_{jj}^{tv} \\ &\quad + \sum_{i=1}^{n}\sum_{j=1}^{n}\tilde{R}_{ij}^{st}\tilde{R}_{ji}^{uv}W_{ii}^{sv}W_{jj}^{tu}\end{aligned}$$

where $\kappa_i^{stuv} = \mathrm{E}(S_{ii}^{st} S_{ii}^{uv}) - W_{ii}^{st} W_{ii}^{uv} - W_{ii}^{su} W_{ii}^{tv} - W_{ii}^{sv} W_{ii}^{tu}$. It is easy to check that the value of κ_i^{stuv} does not depend on the order of s, t, u and v. Calculation of the values of κ_i^{stuv} as given in (4.6) is straightforward but tedious.

Taking all covariances of the Q_{st} terms together, we have

$$\mathrm{Var}(\tilde{Q}) = \sum_{s=1}^{g} \sum_{t=1}^{g} \sum_{u=1}^{g} \sum_{v=1}^{g} \mathrm{Cov}(Q_{st}, Q_{uv}).$$

The result (4.5) follows by rewriting

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \tilde{R}_{ij}^{st} \tilde{R}_{ji}^{vu} W_{ii}^{su} W_{jj}^{tv} = \mathrm{trace}(\tilde{R}^{st} W^{tv} \tilde{R}^{vu} W^{su})$$

and

$$\sum_{s=1}^{g} \sum_{t=1}^{g} \sum_{u=1}^{g} \sum_{v=1}^{g} \mathrm{trace}(\tilde{R}^{st} W^{tu} \tilde{R}^{uv} W^{sv}) = \mathrm{trace}(\tilde{\mathbf{R}}\mathbf{W}\tilde{\mathbf{R}}\mathbf{W}).$$

4.11 Derivation of the test statistic

We derive the expression (4.9) for the score test statistic from the random effects model (4.7). The likelihood $L(\tau^2) = \exp\{\ell(\tau^2)\}$ can be written

$$L(\tau^2) = \mathrm{E}_{\mathbf{z}}\big[\prod_{i=1}^{n} f_i(\mathbf{z})\big],$$

where $f_i(\mathbf{r}) = \exp\{l_i(\mathbf{z})\}$ and

$$l_i(\mathbf{z}) = \sum_{s=1}^{g} y_{is} \log\{\nu_{is}(\mathbf{z})\}.$$

Compare (4.8). Note that $f_i(\mathbf{z})$ only depends on $(z_{i1}, \ldots, z_{ig})$. Therefore, Taylor expanding $L(\tau^2)$ with respect to $\mathbf{z}$ at $\mathbf{z} = \mathbf{0}$ gives

$$\begin{aligned}
L(\tau^2) &= \mathrm{E}_{\mathbf{z}}\Bigg[\prod_{i=1}^{n} f_i(\mathbf{0}) + \sum_{s=1}^{g} \sum_{i=1}^{n} z_{is} \frac{\partial f_i(\mathbf{0})}{\partial z_{is}} \prod_{j \neq i} f_j(\mathbf{0}) \\
&\quad + \frac{1}{2} \sum_{s=1}^{g} \sum_{t=1}^{g} \sum_{i=1}^{n} z_{is} z_{it} \frac{\partial^2 f_i(\mathbf{0})}{\partial z_{is} \partial z_{it}} \prod_{j \neq i} f_j(\mathbf{0}) \\
&\quad + \frac{1}{2} \sum_{s=1}^{g} \sum_{t=1}^{g} \sum_{i=1}^{n} \sum_{j \neq i} z_{is} z_{jt} \frac{\partial f_i(\mathbf{0})}{\partial z_{is}} \frac{\partial f_j(\mathbf{0})}{\partial z_{jt}} \prod_{k \neq i,j} f_k(\mathbf{0}) + o(\mathbf{z}\mathbf{z}')\Bigg] \\
&= \prod_{i=1}^{n} f_i(\mathbf{0}) + \frac{1}{2}\tau^2 \sum_{s=1}^{g} \sum_{t=1}^{g} \sum_{i=1}^{n} R_{ii}^{st} \frac{\partial^2 f_i(\mathbf{0})}{\partial z_{is} \partial z_{it}} \prod_{j \neq i} f_j(\mathbf{0}) \\
&\quad + \frac{1}{2}\tau^2 \sum_{s=1}^{g} \sum_{t=1}^{g} \sum_{i=1}^{n} \sum_{j \neq i} R_{ij}^{st} \frac{\partial f_i(\mathbf{0})}{\partial z_{is}} \frac{\partial f_j(\mathbf{0})}{\partial z_{jt}} \prod_{k \neq i,j} f_k(\mathbf{0}) + o(\tau^2).
\end{aligned}$$

Using $\frac{\partial f_i(\mathbf{z})}{\partial z_{is}} = f_i(\mathbf{z})\frac{\partial l_i(\mathbf{z})}{\partial z_{is}}$ and $\frac{\partial^2 f_i(\mathbf{z})}{\partial z_{is}\partial z_{it}} = f_i(\mathbf{z})\left[\frac{\partial^2 l_i(\mathbf{z})}{\partial z_{is}\partial z_{it}} + \frac{\partial l_i(\mathbf{z})}{\partial z_{is}}\frac{\partial l_i(\mathbf{z})}{\partial z_{it}}\right]$, this expression can be rewritten to

$$\begin{aligned} L(\tau^2) &= \prod_{i=1}^{n} f_i(\mathbf{0})\left[1 + \frac{1}{2}\tau^2 \sum_{s=1}^{g}\sum_{t=1}^{g}\sum_{i=1}^{n} R_{ii}^{st}\frac{\partial l_i(\mathbf{0})}{\partial z_{is}\partial z_{it}}\right. \\ &\quad \left. + \frac{1}{2}\tau^2 \sum_{s=1}^{g}\sum_{t=1}^{g}\sum_{i=1}^{n}\sum_{j=1}^{n} R_{ij}^{st}\frac{\partial l_i(\mathbf{0})}{\partial z_{is}}\frac{\partial l_j(\mathbf{0})}{\partial z_{jt}}\right] + o(\tau^2) \end{aligned}$$

Because $\frac{\partial \ell(0)}{\partial \tau^2} = \frac{1}{L(0)}\frac{\partial L(0)}{\partial \tau^2}$, the score function at $\tau^2 = 0$ is

$$\frac{\partial \ell(0)}{\partial \tau^2} = \frac{1}{2}\left[\sum_{s=1}^{g}\sum_{t=1}^{g}\sum_{i=1}^{n} R_{ii}^{st}\frac{\partial^2 l_i(\mathbf{0})}{\partial z_{is}\partial z_{it}} + \sum_{s=1}^{g}\sum_{t=1}^{g}\sum_{i=1}^{n}\sum_{j=1}^{n} R_{ij}^{st}\frac{\partial l_i(\mathbf{0})}{\partial z_{is}}\frac{\partial l_j(\mathbf{0})}{\partial z_{jt}}\right]$$

The result (4.9) follows from $\frac{\partial l_i(\mathbf{0})}{\partial z_{is}} = y_{is} - \mu_{is}$ and $\frac{\partial^2 l_i(\mathbf{0})}{\partial z_{is} z_{it}} = -W_{it}^{st}$.

CHAPTER 5

Testing against a high-dimensional alternative

Abstract

As the dimensionality of the alternative increases, the power of classical tests tends to diminish quite rapidly. This is especially true for high-dimensional data in which there are more parameters than observations. In this paper we discuss a score test on a hyperparameter in an empirical Bayesian model as an alternative to classical tests. It gives a general test statistic which can be used to test a point null hypothesis against a high-dimensional alternative, even when the number of parameters exceeds the number of samples. This test will be shown to have optimal power on average in a neighbourhood of the null, which makes it a proper generalization of the locally most powerful test to multiple dimensions. To illustrate this new locally most powerful test we investigate the case of testing the global null hypothesis in a linear regression model in more detail. The score test is shown to have significantly more power than the F-test whenever under the alternative the large-variance principal components of the design matrix explain substantially more of the variance of the outcome than the low-variance principal components. The score test is also useful for detecting sparse alternatives in truly high-dimensional data, where its power is comparable to the test based on the maximum absolute t-statistic.

5.1 Introduction

In a linear regression model one traditionally uses the F-test to test the global null hypothesis that all regression coefficients are zero. However, it is well known that the F-test has low power when the number of covariates in the

This chapter will appear as: J. J. Goeman, S. A. van de Geer and J. C. van Houwelingen (2006) Testing against a high-dimensional alternative. *Journal of the Royal Statistical Society, Series B* **68**, in press. The definitive version will be available at `http://www.blackwell-synergy.com`.

model is close to the number of samples. The F-test even breaks down completely when the number of covariates exceeds the number of samples. Similar behaviour is known for the likelihood ratio test in generalized linear models. In general, classical tests tend to perform badly when used against high dimensional alternatives.

This paper explores testing of a simple null hypothesis against a high-dimensional alternative. We shall formulate a simple test which can be used in high-dimensional models regardless of the number of parameters. This test is constructed as a locally most powerful test (score test) on the hyperparameter in an empirical Bayesian model. The same type of test has been introduced for specific models in the context of microarray gene expression data, where it is used to generalize a test for association between a clinical variable and a single gene to a test for association between a clinical variable and a group of genes. Goeman et al. (2004) have applied this methodology in generalized linear models with a canonical link function and Goeman et al. (2005) in the Cox proportional hazards model. For examples of real data applications we refer to these papers.

In the present paper we explore the general power properties of this type of test in more detail, adopting a purely frequentist point of view. The test will be shown to have optimal average power in a neighbourhood of the null hypothesis, a property which follows as a corollary to the Neyman-Pearson lemma. This property makes the test a natural generalization of the locally most powerful test to higher dimensions, and motivates us to refer this high-dimensional version of the locally most powerful test simply as the locally most powerful test.

We shall also look more closely into the relatively simple case of a high-dimensional alternative in a linear model. In this model there are few distracting details and many quantities can be explicitly calculated. We investigate the regions of the parameter space where the empirical Bayes score test has most and least power and situations where we may expect good power.

In the linear model it is also relatively easy to investigate links with other tests, most notably the F-test. It turns out that the F-test can be formulated as an empirical Bayesian score test with a different prior distribution, a fact which gives insight into the power properties of the F-test. We also investigate relationships between our empirical Bayes procedure with principal components tests and with a typical multiple testing procedure from microarray data analysis which uses the maximum of all absolute univariate T-statistics as a test statistic for the global null. All these comparisons will be illustrated with simulations based on real microarray data (taken from Van de Vijver et al., 2002).

5.2 Empirical Bayes testing

Suppose we have observations $\mathbf{y}$ (typically an n-vector), the distribution of which is assumed to depend on a p-vector of parameters $\boldsymbol{\beta}$. In this model we want to test

$$H_0 : \boldsymbol{\beta} = \mathbf{0}$$

against $H_A : \boldsymbol{\beta} \neq \mathbf{0}$. There may also be some nuisance parameters, but we assume them known for the moment.

If the dimension p of the alternative is large, the alternatives can range over a huge space and H_A typically allows many widely different distributions of $\mathbf{y}$. Some of the alternatives may even induce the same distribution of $\mathbf{y}$ as H_0, especially if $p > n$. In a generalized linear model, for example, the distribution of $\mathbf{y}$ depends on $\boldsymbol{\beta}$ only through $X\boldsymbol{\beta}$, where X is an $n \times p$ design matrix. If $p > n$, there are many alternatives which have $\boldsymbol{\beta} \neq \mathbf{0}$ but $X\boldsymbol{\beta} = \mathbf{0}$. These alternatives give rise to the same distribution of $\mathbf{y}$ as the null hypothesis, which means we can never hope to have any power against these alternatives. This is typical for high-dimensional alternatives: a minimax type approach which tries to have power against all alternatives is bound to fail.

Therefore it seems a sensible approach to focus the power of the test on what we choose to be the most interesting alternatives. This can be done in a Bayesian fashion by assigning the vector $\boldsymbol{\beta}$ a distribution. This distribution should give most probability mass to the alternatives which are perceived as more likely (as in a prior distribution) or simply as more 'interesting' to detect.

What this distribution should be depends very much on the model and the purpose of the test. However, a good choice for such a distribution is usually one that is 'unbiased', i.e. it is symmetric around the null hypothesis and therefore has $\mathrm{E}(\boldsymbol{\beta}) = \mathbf{0}$. This is sensible, because we are usually equally interested in detecting the alternative that $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ as in detecting $\boldsymbol{\beta} = -\boldsymbol{\beta}_0$ for every $\boldsymbol{\beta}_0$. The covariance matrix of $\boldsymbol{\beta}$ may then be chosen in general as $\mathrm{E}(\boldsymbol{\beta}\boldsymbol{\beta}') = \tau^2\Sigma$ for some well-chosen positive (semi-)definite $p \times p$ matrix Σ. The choice $\Sigma = I$ deserves special attention, because it follows from an exchangeability assumption: the density of all permutations of the vector $\boldsymbol{\beta}$ is equal (Bernardo and Smith, 1994, p. 180). Under this exchangeable assumption one is not prejudiced as to which elements of $\boldsymbol{\beta}$ are expected to be large or which elements of $\boldsymbol{\beta}$ are expected to be similar. This assumption is useful when there is no structure or ordering in the parameters that can be readily exploited and when the typical range of the parameter values is similar.

One can complete the specification of the distribution of $\boldsymbol{\beta}$ by choosing a value τ_0^2 for τ^2 and a distributional shape. In the generalized linear model setting, taking the maximum likelihood estimate of $\boldsymbol{\beta}$ will then result in one of

many familiar penalized regression methods, depending on the choice of the distribution of $\boldsymbol{\beta}$. Choosing $\boldsymbol{\beta}$ to have i.i.d. normal entries results in a (generalized) ridge regression (Hoerl and Kennard, 1970). Choosing the regression coefficients $\boldsymbol{\beta}$ i.i.d. double exponential results in the LASSO method (Tibshirani, 1996). These methods are frequently used in estimation and prediction problems in high-dimensional regression models.

We can also use the chosen distribution of $\boldsymbol{\beta}$ as a tool to rephrase our testing problem, rewriting it in terms of the marginal distribution of $\mathbf{y}$. Let $f(\boldsymbol{\beta}; \mathbf{y})$ be the likelihood of $\boldsymbol{\beta}$ for given $\mathbf{y}$. The marginal density of $\mathbf{y}$ is

$$\bar{f}(\tau^2; \mathbf{y}) = \mathrm{E}_{\boldsymbol{\beta}|\tau^2}[f(\boldsymbol{\beta}; \mathbf{y})],$$

which can be interpreted as the likelihood of τ^2 in a new marginal model of $\mathbf{y}$. In this new model, rejecting the new null hypothesis $\bar{H}_0 : \tau^2 = 0$ implies rejecting the old $H_0 : \boldsymbol{\beta} = \mathbf{0}$, as the two imply the same distribution of $\mathbf{y}$.

The testing procedure based on testing $\bar{H}_0 : \tau^2 = 0$ against $\bar{H}_A : \tau^2 = \tau_1^2$ can be called "empirical Bayes testing", because we have put a prior on the parameter vector $\boldsymbol{\beta}$ of the model, which depends on an unknown hyperparameter τ^2, and our inference on $\boldsymbol{\beta}$ proceeds through inference on τ^2. On the other hand it can also simply be called "Bayesian testing", because once the shape of the distribution and the value of τ_1^2 are chosen, the model H_A is fully Bayesian.

One important use of testing $\bar{H}_0$ in the marginal model of $\mathbf{y}$ lies in Lemma 1, a corollary to the Neyman-Pearson Lemma. It says that if we take a specific distribution of $\boldsymbol{\beta}$ and construct a likelihood ratio test in the marginal model, the resulting test has optimal power on average over the chosen distribution of alternatives.

Lemma 1 (Empirical Bayes version of Neyman-Pearson) *Let A_1 be the critical region of a likelihood ratio test of $\bar{H}_0 : \tau^2 = 0$ against $\bar{H}_A : \tau^2 = \tau_1^2$ in the marginal model $\bar{f}$, with associated power function $\bar{w}_{\tau_1^2}(\boldsymbol{\beta}) = \mathrm{P}_{\mathbf{y}|\boldsymbol{\beta}}[A_1]$; and let A be the critical region of any test of $H_0 : \boldsymbol{\beta} = \mathbf{0}$, with power function $w(\boldsymbol{\beta}) = \mathrm{P}_{\mathbf{y}|\boldsymbol{\beta}}[A]$. Then*

$$w(\mathbf{0}) \leq w_{\tau_1^2}(\mathbf{0})$$

implies

$$\mathrm{E}_{\boldsymbol{\beta}|\tau_1^2}[w(\boldsymbol{\beta})] \leq \mathrm{E}_{\boldsymbol{\beta}|\tau_1^2}[w_{\tau_1^2}(\boldsymbol{\beta})].$$

This is a well-known result. The proof is immediate from the Neyman-Pearson Lemma when it is observed that $\mathrm{E}_{\boldsymbol{\beta}|\tau_1^2}[w(\boldsymbol{\beta})] = \mathrm{E}_{\boldsymbol{\beta}|\tau_1^2}\{\mathrm{P}_{\mathbf{y}|\boldsymbol{\beta}}[A]\} = \mathrm{P}_{\mathbf{y}|\tau_1^2}[A]$.

The result of Lemma 1 could immediately be used in practice if we were willing to completely specify the distribution of $\boldsymbol{\beta}$, or at least to specify the

shape of the distribution up to a number of parameters which can be estimated. In most cases, however, we should be reluctant to do this, for two reasons. In the first place, the marginal likelihood is a complicated p-dimensional integral, which often makes it difficult to estimate hyperparameters and usually almost impossible to find the distribution of the test statistic, except in very special cases. Secondly, specifying the distributional shape of $\boldsymbol{\beta}$ means specifying whether the interesting alternatives have a $\boldsymbol{\beta}$ with a few large entries or many small ones. This is a kind of judgement which is typically very difficult to make in high-dimensional data. In a high-dimensional regression model, for example, it is usually not known whether there are few large or many small regression coefficients. A wrong choice of the distribution of $\boldsymbol{\beta}$ could mean low power. How can we avoid specifying the distributional shape of $\boldsymbol{\beta}$?

5.3 The locally most powerful test

It turns out that we can design a test for $\bar{H}_0$ in the marginal model which manages to avoid full specification of the distribution of $\boldsymbol{\beta}$ and avoids evaluation of the complicated marginal likelihood as well. This can be done by constructing the test as a score test.

The traditional score test is a one-sided test of $H_0^* : \theta = \theta_0$ against $H_A^* : \theta > \theta_0$ in a one parameter model with likelihood $f^*(\theta; \mathbf{y})$. It rejects when the score test statistic $S^*(\mathbf{y}) = \frac{\mathrm{d}}{\mathrm{d}\theta} \log f^*(\theta_0; \mathbf{y}) \geq k$ for some constant k. If θ_0 is on the edge of the parameter space, $S^*(\mathbf{y})$ should be taken as the right-sided derivative. For typical values of the test size α the critical value k is almost invariably positive, because, by the properties of the score function, $S^*(\mathbf{y})$ has zero expectation under the null hypothesis.

The score test is known as the "locally most powerful test" as a consequence of Lemma 2. This lemma says that the score test has optimal slope of the power function among all tests of at most the same size, so that it has optimal power against local alternatives close to the null.

Lemma 2 (Score test property) *Suppose that the derivative $\frac{\mathrm{d}}{\mathrm{d}\theta} f^*(\theta; \mathbf{y})$ exists a.e. and is bounded in a (right-)neighbourhood of θ_0. Then for any test of H_0^* with critical region A and power function $w(\theta) = \mathrm{P}_{\mathbf{y}|\theta}[A]$, the derivative $\frac{\mathrm{d}}{\mathrm{d}\theta} w(\theta_0)$ exists. Moreover, if $w^*(\theta) = \mathrm{P}_{\mathbf{y}|\theta}[S^* \geq k]$ is the power function of the score test, then either of*

(i) $w(\theta_0) = w^*(\theta_0)$

(ii) $w(\theta_0) \leq w^*(\theta_0)$ *and* $k \geq 0$

implies

$$\frac{\mathrm{d}}{\mathrm{d}\theta} w(\theta_0) \leq \frac{\mathrm{d}}{\mathrm{d}\theta} w^*(\theta_0).$$

The proof of this lemma is given in Section 5.12.

A more extensive treatment of locally most powerful tests in one dimension is given in Cox and Hinkley (1974). They show that the score test can be interpreted as the limit for $\theta_1 \downarrow \theta_0$ of the likelihood ratio test of H_0^* against the point alternative $H_1^* : \theta = \theta_1$. Score tests are typically useful when testing an 'easy' null hypothesis against a 'complicated' alternative, because score testing does not require estimation of θ. Our high-dimensional alternative is a good example of such a complicated alternative.

We shall apply score testing in the empirical Bayesian setting by testing $\bar{H}_0 : \tau^2 = 0$ against $\bar{H}_A : \tau^2 > 0$ in the marginal model using the score test statistic

$$S = \frac{\mathrm{d}}{\mathrm{d}\tau^2} \log \bar{f}(0; \mathbf{y}),$$

which is automatically a right-sided derivative as $\bar{f}$ is only defined for $\tau^2 \geq 0$. This test has two very useful properties, which we have formulated as Lemma 3 and Lemma 4.

The first property is important both for computation and for modelling. Lemma 3 says that the test statistic S can be found with simple matrix operations from the conditional likelihood $f(\boldsymbol{\beta}; \mathbf{y})$ and the covariance matrix of $\boldsymbol{\beta}$. This implies that we do not need numerical integration to find the value of the test statistic and that we do not have to specify the distributional shape of the distribution of $\boldsymbol{\beta}$.

Lemma 3 (Score test statistic) *Suppose $\boldsymbol{\beta} = \tau \mathbf{b}$, where $\mathrm{E}\mathbf{b} = \mathbf{0}$ and $\mathrm{E}(\mathbf{b}\mathbf{b}') = \Sigma$ and the distribution of $\mathbf{b}$ does not depend on τ. Suppose also that loglikelihood $\log f(\boldsymbol{\beta}; \mathbf{y})$ and its first two derivatives exist a.e. and are bounded in a neighbourhood of $\boldsymbol{\beta} = \mathbf{0}$. Then the empirical Bayes score test statistic $S = \frac{\mathrm{d}}{\mathrm{d}\tau^2} \log \bar{f}(0; \mathbf{y})$ exists and is given by*

$$S = \frac{1}{2}\mathbf{s}'\Sigma\mathbf{s} - \frac{1}{2}\mathrm{trace}[\Sigma\mathbf{I}]$$

where $\mathbf{s} = \frac{\partial}{\partial \boldsymbol{\beta}} \log f(\mathbf{0}; \mathbf{y})$ is the score function and $\mathbf{I} = \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \log f(\mathbf{0}; \mathbf{y})$ the observed Fisher information of $\boldsymbol{\beta}$ in H_0.

The proof of this lemma is a simple calculation, which is given in Section 5.12.

The second and most important property of the score test based on S is given in Lemma 4. It is again an optimality property, which effectively combines the statements of Lemmas 1 and 2. Lemma 4 says that the empirical Bayes score test, which has optimal slope of the power function in the marginal model $\bar{f}$, has optimal expected slope of the power function in the conditional model f.

This lemma only holds for the exchangeable version of the test with $\Sigma = I$, although a more general version can also be formulated.

Lemma 4 (Locally Optimal Power) *Suppose the conditions of Lemma 3 hold with $\Sigma = I$. Let $\bar{w}(\boldsymbol{\beta}) = \mathrm{P}_{\mathbf{y}|\boldsymbol{\beta}}[S \geq k]$ be the power function of the exchangeable score test of H_0. Let $w(\theta) = \mathrm{P}_{\mathbf{y}|\boldsymbol{\beta}}[A]$ be the power function of any test of H_0. Then either of*

(i) $w(\mathbf{0}) = \bar{w}(\mathbf{0})$

(ii) $w(\mathbf{0}) \leq \bar{w}(\mathbf{0})$ *and* $k \geq 0$

implies

$$\mathrm{E}_{\boldsymbol{\xi}}[\tfrac{\mathrm{d}}{\mathrm{d}\tau^2} w_{\boldsymbol{\xi}}(0)] \leq \mathrm{E}_{\boldsymbol{\xi}}[\tfrac{\mathrm{d}}{\mathrm{d}\tau^2} \bar{w}_{\boldsymbol{\xi}}(0)]$$

where $w_{\boldsymbol{\xi}}(\tau) = w(\tau\boldsymbol{\xi})$, $\bar{w}_{\boldsymbol{\xi}}(\tau) = \bar{w}(\tau\boldsymbol{\xi})$ and $\boldsymbol{\xi}$ has a uniform distribution on the unit p-ball ($p = \dim(\boldsymbol{\beta})$). The same result holds when $\boldsymbol{\xi}$ has any other distribution on the unit p-ball such that $\mathrm{E}(\boldsymbol{\xi}) = \mathbf{0}$ and $\mathrm{E}(\boldsymbol{\xi}\boldsymbol{\xi}') \propto I$.

The proof of the lemma is given in Section 5.12. In fact, Lemma 4 follows from Lemma 2 in more or less the same way as Lemma 1 follows from the Neyman-Pearson Lemma.

By Lemma 4 we see that the score test in the exchangeable empirical Bayesian model has optimal expected slope of the power function, where the expectation is with respect to taking a random direction in p-space. This is the property that motivates its name of locally most powerful test. It is an interesting sidenote that even if $p = 1$, by Lemma 3 the high-dimensional score test based on S is not the same as the ordinary one-dimensional score test based on S^*, because the test based on S is a two-sided test, whereas the test based on S^* is one-sided. By Lemmas 3 and 4 the test based on S is the proper generalization of the one-dimensional score test from one-sided to two-sided alternatives.

5.4 Nuisance parameters

The presence of nuisance parameters complicates some of the issues described above. When nuisance parameters are present, the null hypothesis is not simple anymore but composite. In that case strict optimality in the sense of Lemma 4 is impossible.

The issue of nuisance parameters is usually tackled by switching to the profile likelihood (Pawitan, 2001). When using a score test, switching to the profile likelihood is very easy: one can simply plug in the maximum likelihood estimate of the nuisance parameter under the null hypothesis. This can be easily

seen in a simple two parameter model with loglikelihood $g(\theta, \eta)$ and profile likelihood $\hat{g}(\theta) = g\{\theta, \hat{\eta}(\theta)\}$. In this situation

$$\frac{\partial \hat{g}}{\partial \theta} = \frac{\partial g}{\partial \theta} + \frac{\partial g}{\partial \eta}\frac{\partial \eta}{\partial \theta}. \tag{5.1}$$

The second term on the right hand side is zero, because $\partial g/\partial \eta$ is always zero in $\hat{\eta}$.

This simple plugging in of the null estimate of the nuisance parameters can also be understood by viewing the score test again as a (profile) likelihood ratio test of $\theta = \theta_0$ versus $\theta = \theta_1$ for $\theta_1 \downarrow \theta_0$. In the limit the maximum likelihood estimate of η is the same under the alternative as under the null.

In the empirical Bayes model of this paper the situation is basically the same. A similar argument to (5.1) can be used to check in the proof of Lemma 3 that plugging in the estimate under the null is equivalent to using the profile likelihood. For this derivation it makes no difference whether one uses the conditional profile likelihood, starting with likelihood f and the maximum likelihood estimate $\hat{\eta}(\boldsymbol{\beta}; \mathbf{y})$ of the nuisance parameter η as a function of $\boldsymbol{\beta}$, or whether one uses the marginal likelihood $\bar{f}$ and the maximum likelihood estimate $\bar{\eta}(\tau^2; \mathbf{y})$ from the marginal model for given τ^2. Both profile likelihoods lead to the same test.

See section 5.6 for an example of a model with nuisance parameters.

5.5 Distribution of the test statistic

The specification of the locally most powerful test in the previous sections is not fully complete, as it only provides us with the test statistic to be used. To be able to use the test in practice, we must also know the distribution of the test statistic under the null, so as to be able to find the cutoff for significance and/or the p-value. There is no general method for finding the null distribution, and this may require some extra work when the concept of the locally most powerful test is to be applied in the context of a specific model. We only give some general comments here. See section 5.6 and Goeman et al. (2004) and Goeman et al. (2005) for concrete examples.

First, we look at the null distribution of S. It should be noted that, aside from having zero expectation under the null, the test statistic S is not yet standardized and, in general, should not be expected to follow any standard textbook distribution. It is usually not easy to directly apply asymptotic results on the distribution of the score statistic, because the marginal likelihood $\bar{f}$, from which the score statistic was derived, is not generally a product of n contributions of the individuals. Asymptotic arguments may be used in specific models (as in Goeman et al., 2005), but we have no general theory yet.

In many cases, however, one can find a reasonably good approximation to the distribution of S because the expression for S, as given in Lemma 3, is relatively easy. The mean of S and its variance can often be explicitly calculated. This allows approximation of the null distribution by moment matching to a tabulated distribution (this strategy was used in Goeman et al., 2004). Other practical options for finding the distribution of S include numerical integration or permutation methods. Exact calculation of the distribution function of S is possible in special cases, such as testing the global null hypothesis in the linear model with normal errors, which is the case we shall turn to now.

5.6 The linear model

The optimality property implied in Lemma 4 is very appealing, but it has its limitations. Good power is guaranteed, but only locally near the null and on average over many possible alternatives. To investigate more closely what Lemma 4 is worth for specific alternatives, we shall examine the simplest case of the linear model in detail.

Assume that $\mathbf{y} \sim \mathcal{N}(X\boldsymbol{\beta}, \sigma^2 I)$, where X is an $n \times p$ design matrix of full rank $\min(n, p)$. For simplicity we ignore the intercept parameter α which would normally be included (See Goeman et al., 2004, on how to deal with the nuisance parameter α). The score vector for this model is $\mathbf{s} = \sigma^{-2} X'\mathbf{y}$ and the observed Fisher information is $\mathbf{I} = \sigma^{-2} X'X$, so the general empirical Bayes score test statistic is

$$\tilde{S}_\Sigma = \tfrac{1}{2\sigma^4}\mathbf{y}'X\Sigma X'\mathbf{y} - \tfrac{1}{2\sigma^2}\text{trace}(X\Sigma X').$$

It is more convenient to work with the equivalent test statistic $\sigma^{-2}\mathbf{y}'X\Sigma X'\mathbf{y}$, whose distribution does not depend on σ^2. Because σ^2 is not known, we plug in its maximum likelihood estimate $\hat{\sigma}_0^2 \propto \mathbf{y}'\mathbf{y}$ under the null hypothesis. The resulting test statistic is

$$S_\Sigma = \frac{\mathbf{y}'X\Sigma X'\mathbf{y}}{\mathbf{y}'\mathbf{y}}, \tag{5.2}$$

whose distribution also does not depend on the nuisance parameter σ^2. We study the exchangeable case $\Sigma = I$, as 'the' locally most powerful test statistic

$$S = \frac{\mathbf{y}'XX'\mathbf{y}}{\mathbf{y}'\mathbf{y}}.$$

To find the distribution function of S, we can use the following identity (Azzalini and Bowman, 1993):

$$\mathrm{P}\{S > t\} = \mathrm{P}\{\mathbf{y}'(XX' - tI)\mathbf{y} > 0\}.$$

The distribution function of the quotient S can therefore be found through the distribution function of a quadratic form in normal variables. We use numeric methods developed by Imhof (1961) to calculate the latter distribution function. Reasonably good approximations to the 5% and 1% cutoff values can also be found by equating the moments of S to those of a gamma distribution, a strategy which was used in Goeman et al. (2004).

It is interesting to note a connection between the test statistic S and the method of partial least squares (PLS), which is often used for high-dimensional data in chemometrics (Brown, 1993). The first component of a partial least squares regression is $XX'\mathbf{y}$, so the test statistic S can be viewed as a test for correlation between the first PLS component and $\mathbf{y}$.

5.7 Power of the score test

We want to gain insight in the power of the locally most powerful test in practice. It has already been said that when the alternatives are high-dimensional, it is impossible to have power against all alternatives. To see which are the alternatives that our score test cannot detect, we check which alternatives have an expected test statistic that is smaller than expected under the null. These alternatives have power below the size α of the test.

Under the null hypothesis, the test statistic S has expectation

$$\mathrm{E}_{\mathbf{y}|\mathbf{0}}[S] = \frac{1}{n}\mathrm{trace}(XX').$$

Under the alternative the expectation of S can be well approximated by taking the expectations of the numerator and the denominator separately

$$\mathrm{E}_{\mathbf{y}|\boldsymbol{\beta}}[S] \approx \frac{\boldsymbol{\beta}'X'XX'X\boldsymbol{\beta} + \sigma^2\mathrm{trace}(XX')}{\boldsymbol{\beta}'X'X\boldsymbol{\beta} + n\sigma^2}.$$

This approximation is not only asymptotically exact, but also for small sample size if $\mathbf{y}$ is either dominated by $X\boldsymbol{\beta}$ or by σ^2 (i.e. in any of the limits $n \to \infty$, $\sigma^2 \to 0$, $\sigma^2 \to \infty$ or $\boldsymbol{\beta} \to \mathbf{0}$).

The difference between the expectations is

$$\mathrm{E}_{\mathbf{y}|\boldsymbol{\beta}}[S] - \mathrm{E}_{\mathbf{y}|\mathbf{0}}[S] \approx \frac{\boldsymbol{\beta}'X'XX'X\boldsymbol{\beta} - \frac{1}{n}\boldsymbol{\beta}'X'X\boldsymbol{\beta}\cdot\mathrm{trace}(XX')}{\boldsymbol{\beta}'X'X\boldsymbol{\beta} + n\sigma^2}.$$

To interpret this expression we must look at the principal components of X and the amount of variance of $\mathbf{y}$ that each principal component explains. Call

$$r^2 = \frac{\boldsymbol{\beta}'X'X\boldsymbol{\beta}}{\boldsymbol{\beta}'X'X\boldsymbol{\beta} + n\sigma^2},$$

the fraction of the variance of $\mathbf{y}$ explained by the alternative. We use the spectral decomposition. Write $X'X = \sum_{i=1}^{n} \lambda_i Q_i$, where $\lambda_1 \geq \ldots \geq \lambda_n \geq 0$ are eigenvalues of $X'X$ and Q_i is the $p \times p$ projection matrix that projects onto the eigenvector of $X'X$ corresponding to the eigenvalue λ_i. Note that we can stop the decomposition at the n-th component because the rank of $X'X$ is $\min(n, p) \leq n$. Use of the spectral decomposition gives $r^2 = \sum_{i=1}^{n} r_i^2$, with $r_i^2 = \lambda_i \boldsymbol{\beta}' Q_i \boldsymbol{\beta} / (\boldsymbol{\beta}' X'X \boldsymbol{\beta} + n\sigma^2)$, and

$$\mathrm{E}_{\mathbf{y}|\boldsymbol{\beta}}[S] - \mathrm{E}_{\mathbf{y}|\mathbf{0}}[S] = \sum_{i=1}^{n} \lambda_i r_i^2 - \frac{1}{n} \sum_{i=1}^{n} \lambda_i \sum_{j=1}^{n} r_j^2.$$

This can be recognized as proportional to the covariance of the vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)'$ of variances of the principal components of X and the vector $\mathbf{r} = (r_1^2, \ldots, r_n^2)'$, which gives the fraction of the variance of $\mathbf{y}$ explained by these components.

This small exercise has a few interesting conclusions. In the first place there are many alternatives, especially in the $p \geq n$ case, for which the locally most powerful test has negligible power. These are the alternatives for which the low-variance principal components of X explain most of the variance of $\mathbf{y}$. These undetectable alternatives may have any value of r^2, even $r^2 = 1$: an alternative with $\mathrm{E}_{\mathbf{y}|\boldsymbol{\beta}}[S] \leq \mathrm{E}_{\mathbf{y}|\mathbf{0}}[S]$ and $r^2 = 1$ will even have zero power.

Fortunately for the score test, a negative covariance of $\boldsymbol{\lambda}$ and $\mathbf{r}$ occurs only seldomly in real data, because the measurements in X are often noisy or inaccurate. The uninformative noise tends be dominant in the small-variance principal components of X.

How can a test be most powerful on average if it has such low power against many alternatives? The reason for this lies in the assumption of exchangeability that underlies the test. By Lemma 4 the power is optimal on a small p-ball with $\boldsymbol{\beta}'\boldsymbol{\beta} = c$. The alternatives on this ball have very diverse values of r^2: alternatives which have $\boldsymbol{\beta}$ in directions corresponding to the eigenvectors of the large eigenvalues of $X'X$ have large r^2, others have small r^2. It is very difficult to have much power against alternatives with small r^2. Even an 'oracle' which knows the direction of $\boldsymbol{\beta}$ and only tests whether $\|\boldsymbol{\beta}\| = 0$ will have low power if the true $\boldsymbol{\beta}$ has low r^2. Average power will increase, therefore, if some power on the low-potential alternatives is sacrificed in exchange for a gain in power for the high-potential alternatives. This is the advantageous trade-off that the exchangeable empirical Bayes score test makes.

If negative covariance of $\boldsymbol{\lambda}$ and $\mathbf{r}$ leads to $\mathrm{E}_{\mathbf{y}|\boldsymbol{\beta}}[S] < \mathrm{E}_{\mathbf{y}|\mathbf{0}}[S]$, conversely a positive covariance of the same $\boldsymbol{\lambda}$ and $\mathbf{r}$ leads to $\mathrm{E}_{\mathbf{y}|\boldsymbol{\beta}}[S] > \mathrm{E}_{\mathbf{y}|\mathbf{0}}[S]$ and potentially good power. Against some of these alternatives the score test must even have very good power, as the test is locally most powerful on average by Lemma 4.

We come back to this in Sections 5.8 and 5.10, where we compare the locally most powerful test with the F-test.

It has to be remarked that the problems of lower expectation of the test statistic S under the alternative than under the null typically disappear when n is large. If we let n grow to kn by observing k samples from each covariate pattern, $\mathrm{E}_{\mathbf{y}|\boldsymbol{\beta}}[S]$ will eventually become larger than $\mathrm{E}_{\mathbf{y}|\mathbf{0}}[S]$, because letting n grow in this setup means augmenting both $\boldsymbol{\lambda}$ and $\mathbf{r}$ with zeros, so that the correlation between the two increases. Similarly, if we have $p < n$ to begin with, there are at least $n - p$ zero elements of $\boldsymbol{\lambda}$ with corresponding zero elements of $\mathbf{r}$, so that the smallest elements of $\boldsymbol{\lambda}$ and $\mathbf{r}$ automatically coincide and there are few alternatives with $\mathrm{E}_{\mathbf{y}|\boldsymbol{\beta}}[S] \leq \mathrm{E}_{\mathbf{y}|\mathbf{0}}[S]$.

5.8 A new look at the F-test

In the $p < n$ situation it is possible to apply both the locally most powerful test and the F-test, which makes it interesting to compare the two. The F-test statistic in our linear model is a constant times

$$\tilde{F} = \frac{\mathbf{y}'X(X'X)^{-1}X'\mathbf{y}}{\mathbf{y}'(I - X(X'X)^{-1}X')\mathbf{y}}.$$

We find it convenient to transform $\tilde{F}$ by the strictly increasing function $g(x) = (x^{-1} + 1)^{-1}$ to the equivalent test statistic $F = g(\tilde{F})$, which is given by

$$F = \frac{\mathbf{y}'X(X'X)^{-1}X'\mathbf{y}}{\mathbf{y}'\mathbf{y}}.$$

Under the null the transformed F has a beta distribution with parameters $\frac{1}{2}p$ and $\frac{1}{2}(n-p)$.

It is now easy to compare F with the locally most powerful test statistic $S = (\mathbf{y}'XX'\mathbf{y})/(\mathbf{y}'\mathbf{y})$. We can immediately notice that, if the design is orthogonal (i.e. $X'X \propto I$) both tests are equivalent. Note that the design is always orthogonal if $p = 1$, so the locally most powerful test for $p = 1$ is equivalent to the F-test and hence to the two-sided t-test.

More fundamental insights follow when comparing F with the general expression for the empirical Bayesian score test statistic given in (5.2): as $S_\Sigma = \mathbf{y}'X\Sigma X'\mathbf{y}/(\mathbf{y}'\mathbf{y})$, we have $F = S_{(X'X)^{-1}}$. It follows that we can look at the F-test as the empirical Bayes score test based on the prior covariance $\mathrm{E}(\boldsymbol{\beta}\boldsymbol{\beta}') = \tau^2(X'X)^{-1}$ for τ^2 very small. By Lemma 1, the F-test therefore optimizes the power on average over this distribution of $\boldsymbol{\beta}$. The F-test is therefore especially directed against alternatives in directions where the variance of the distribution of $\boldsymbol{\beta}$ is large. These directions are the directions of the eigenvectors

of small eigenvalues of $X'X$. These are also the directions where a large r^2 requires a very large $\|\boldsymbol{\beta}\|$. Vice versa, the directions of the eigenvectors of large eigenvalues of $X'X$ get a small prior variance of $\boldsymbol{\beta}$. These are, therefore, of small importance to the F-test: $\boldsymbol{\beta}$ is a priori not expected to lie in these directions. The directions of the eigenvectors of large eigenvalues of $X'X$ are the directions in which a small investment of $\|\boldsymbol{\beta}\|$ results in a large r^2.

We get a similar look at the power properties of the F-test if we orthogonalize the design by taking $\tilde{\boldsymbol{\beta}} = (X'X)^{1/2}\boldsymbol{\beta}$ and $\tilde{X} = X(X'X)^{-1/2}$. This results in $X\boldsymbol{\beta} = \tilde{X}\tilde{\boldsymbol{\beta}}$ for all $\boldsymbol{\beta}$ so the distribution of $\mathbf{y}$ is unchanged. Unlike the F-test, the locally most powerful test is not invariant under change of parametrization: under the assumption of exchangeability $\mathrm{E}[\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}'] = \tau^2 I$ on $\tilde{\boldsymbol{\beta}}$ we now get the F-test as the locally most powerful test for the new parametrization. Applying the reasoning of Section 5.6 to the new parametrization, we see that the F-test optimizes power not over small balls with $\boldsymbol{\beta}'\boldsymbol{\beta} = c$, but on small ellipsoids with $\tilde{\boldsymbol{\beta}}'\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}'X'X\boldsymbol{\beta} = c$, which are ellipsoids of alternatives that have the same r^2. All alternatives with the same r^2 have the same potential power, so there is no trade-off and all alternatives in the ellipsoid are given equal power. The expected test statistic under the alternative minus the expected test statistic under the null for the F-test is

$$\mathrm{E}_{\mathbf{y}|\boldsymbol{\beta}}[F] - \mathrm{E}_{\mathbf{y}|\mathbf{0}}[F] = r^2(1 - \frac{p}{n}),$$

which only depends on $\boldsymbol{\beta}$ through r^2. It is positive whenever $r^2 > 0$ and $p < n$.

The main difference between the empirical Bayes score test and the F-test is therefore that, while for the F-test all alternatives with the same r^2 are as credible and interesting to detect, the score test is explicitly directed at finding parsimonious alternatives, which can explain $\mathbf{y}$ with minimal expenditure of $\|\boldsymbol{\beta}\|$.

There is no easy analytic expression which shows for which alternatives in the $p \leq n$ situation the F-test has more power than the score test and vice versa. However, it can be convincingly argued that for those alternatives in which the large variance principal components of X explain most of the variance of $\mathbf{y}$, the score test has more power, while for the alternatives in which the small-variance principal components explain most of the variance of $\mathbf{y}$, the F-test is more powerful. This can be seen by writing XX' in a spectral decomposition as $XX' = \sum_{i=1}^{n} \lambda_i P_i$, where P_i is the $n \times n$ projection matrix for projection on the i-th principal component. Then

$$S = \sum_{i=1}^{n} \lambda_i \frac{\mathbf{y}'P_i\mathbf{y}}{\mathbf{y}'\mathbf{y}},$$

so the test statistic S is a weighted sum of the test statistics $\mathbf{y}'P_i\mathbf{y}$ which test whether the i-th principal component is associated with $\mathbf{y}$. The weights are proportional to the variance of the principal components. In the same way

$$F = \sum_{i=1}^{n} \frac{\mathbf{y}'P_i\mathbf{y}}{\mathbf{y}'\mathbf{y}},$$

the statistic F is the unweighted sum of the same test statistics. Comparing the two composite tests, one can argue that one has more power than the other if it puts heavier weights on the terms with most power. We'll illustrate this point with simulations in Section 5.10.

An interesting type of alternative against which the locally most powerful test can be expected to have more power than the F-test is a factor-analysis type setup, in which a limited number of latent variables linearly determines both the covariates X and the outcome variable $\mathbf{y}$, but both are measured with error (Bartholomew and Knott, 1999). In this case the latent variables tend to show up in the large-variance principal components of X, while the uninformative noise tends to dominate the small-variance principal components. This setup is not unrealistic for many practical problems, especially in high-dimensional data, as the covariates can often be seen as noisy measurements of more or less the same underlying mechanisms. In this kind of alternative one would normally apply principal components testing: reducing the matrix X to its first few principal components and then applying the F-test. An important advantage of the locally most powerful test over principal components testing is that there is no need to choose the number of principal components. We come back to principal components testing in Section 5.10.

5.9 Sparse alternatives

In the previous sections we have established that the locally most powerful test is especially directed against parsimonious alternatives with small $\|\boldsymbol{\beta}\|$. A different type of parsimonious alternative is the sparse alternative, in which only a few entries of $\boldsymbol{\beta}$ are non-zero. This type of alternative is especially of interest in regression modelling.

A test which specifically aims to detect this type of sparse alternative in a regression model is a multiple testing procedure. This type of testing procedure is often used in microarray data analysis. There are many variants, but the most basic form is the following: for $i = 1, \ldots, p$ a t-test statistic t_i is calculated to test for association of each covariate with the outcome $\mathbf{y}$. The test statistic $\tilde{T}_{\max} = \max(|t_1|, \ldots, |t_p|)$ is used to test whether there is an association between any covariate and $\mathbf{y}$. The critical value of $T_{\max}$ can be found either conservatively using the Bonferroni adjustment, or using numerical methods.

Different though this test may seem from the locally most powerful test, there is still a connection. First, we can transform each $|t_i|$ to $g(t_i^2)$, using the function $g(x) = (x^{-1}+1)^{-1}$ also used in section 5.8. This results in test statistics with a beta distribution with parameters $\frac{1}{2}$ and $\frac{1}{2}(n-1)$. As $g(x^2)$ is increasing in $|x|$, the test statistic

$$T_{\max} = \max\{g(t_1^2), \ldots, g(t_p^2)\}$$

is equivalent to $\tilde{T}_{\max}$. Next, we write $\mathbf{x}_i$ for the i-th column of X, then $g(t_i^2) = \mathbf{y}'\mathbf{x}_i\mathbf{x}_i'\mathbf{y}/(\mathbf{y}'\mathbf{y} \cdot \mathbf{x}_i'\mathbf{x}_i)$. However, as we can write $XX' = \sum_{i=1}^p \mathbf{x}_i\mathbf{x}_i'$, we can say that

$$S = \sum_{i=1}^{p} \mathbf{x}_i'\mathbf{x}_i g(t_i^2),$$

so the locally most powerful test statistic is a weighted sum of the same (transformed) t-test statistics over which $T_{\max}$ is the maximum. The weights are proportional to the variance of $\mathbf{x}_i$.

Perhaps surprisingly, by Lemma 4 the score test is more powerful than the test based on $T_{\max}$ in the situation where p is large and r^2 is very small, even when only a single regression coefficient is non-zero. Suppose $\boldsymbol{\beta}$ is given a single non-zero entry at random, of fixed size, but with random sign. This distribution of $\boldsymbol{\beta}$ has $\mathrm{E}(\boldsymbol{\beta}) = \mathbf{0}$ and, if p is large $\mathrm{E}(\boldsymbol{\beta}\boldsymbol{\beta}') \approx \tau^2 I$ for some τ^2. By Lemma 4, the score test has optimal power on average to detect these alternatives if τ^2 is small.

This optimality can again be understood in terms of the principal components. If there are few principal components with large variance, it is probable that the $\mathbf{x}_i$ with the positive regression coefficient also has a major part of its variance in the direction of these large-variance principal components. If $\mathbf{y}$ is correlated with $\mathbf{x}_i$, it is therefore automatically correlated with these principal components, and therefore with many other covariates $\mathbf{x}_j$, which also tend to have a large part of their variance in the direction of the large-variance principal components. A single regression coefficient may therefore lead to many significant t-statistics. In this situation there may be more information in the sum of the t-statistics than in the maximum.

Simulations in section 5.10 illustrate these points.

5.10 Simulations

Many of the points raised in the previous sections require some illustration. We'll do this using simulations in the linear model. The simulations are based on real data in the sense that the design matrix X is a taken as a real biological

data set: a microarray data set of gene expression measurements of $p = 4911$ genes, measured for $n = 294$ breast cancer patients (obtained from Van de Vijver et al., 2002, after removing some genes and patients due to missing values). The matrix X was normalized to have both row and column means zero. After this normalization X has rank $n - 1$ and a ratio of the largest to the smallest non-zero singular value of 26.6. Using this design matrix X, values of $\mathbf{y}$ are simulated based on the models chosen below.

First we compare the locally most powerful test with the F-test, to illustrate the statements from Section 5.8 that the score test has more power when the large variance principal components of X explain most of the variance of $\mathbf{y}$. As we cannot use the F-test when $p > n$, we reduce the matrix X to X^* by selecting as covariates only the $p^* = 52$ genes belonging to the apoptosis pathway.

The simulation setup is as follows. We write X^* in a singular value decomposition as $X^* = U\Lambda^{1/2}V'$, with U an $n \times p^*$ semi-orthogonal matrix, V a $p^* \times p^*$ orthogonal matrix and Λ a $p^* \times p^*$ diagonal matrix with diagonal elements $\boldsymbol{\lambda}^* = (\lambda_1, \ldots, \lambda_{p^*})'$, where each λ_i is the variance of the i-th principal component. To vary the amount of variance explained by the principal components, we choose the regression coefficients as $\boldsymbol{\beta} = V\Lambda^{(s/2-1)}\boldsymbol{\lambda}$ for various values of s. In this setup the i-th principal component has regression coefficient $\lambda_i^{s/2}$ and explains a fraction r_i^2 of the variance of $\mathbf{y}$ proportional to λ_1^{s+1}. Hence, if $s > 0$, the large variance principal components have larger regression coefficients and therefore explain more of the variance of $\mathbf{y}$; if $-1 < s < 0$, the large variance principal components have smaller regression coefficients, but still explain more of the variance than the small-variance principal components, while if $s < -1$, the small-variance principal components dominate $\mathbf{y}$. By varying σ^2 as a function of s we can obtain all values of r^2 for every s.

To estimate the power for these alternatives, we generated 10000 $\mathbf{y}$ vectors each from alternatives with various values of s and r^2. The cutoff at level α for the S statistic was found using the methods of Imhof (1961). The results are given in table 5.1. They show that the power of the score test and the F-test is comparable for $s = -1/2$, although the F-test still has a slight advantage here. The score test is substantially more powerful for larger values of s, the F-test is more powerful for smaller values. This is in line with the theoretical discussion in section 5.8.

It is also interesting to compare the locally most powerful test with the test P_1, which is the F-test that tests whether the first principal component of X^* is correlated with $\mathbf{y}$. The results are also given in table 5.1. We can see that the locally most powerful test is comparable in power to the test P_1 for high values of s, but it is consistently better for all the alternatives considered.

In a second simulation experiment we look at sparse alternatives in high-

TABLE 5.1: *Monte Carlo power comparison between the locally most powerful test, the F-test and the test P_1, which uses only the first principal component for testing. The tests use $\alpha = 0.05$. The various alternatives are given by their r^2 and a coefficient s: $s > 0$ means that large-variance principal components get larger regression coefficients, $s < 0$ vice versa.*

	$r^2 = 0.02$			$r^2 = 0.05$		
alternative	F	S	P_1	F	S	P_1
$s = 1.5$	0.14	0.52	0.52	0.35	0.92	0.90
$s = 1$	0.14	0.46	0.44	0.35	0.88	0.82
$s = 0.5$	0.14	0.36	0.31	0.34	0.79	0.66
$s = 0$	0.13	0.24	0.19	0.34	0.58	0.39
$s = -0.5$	0.14	0.13	0.10	0.35	0.32	0.18
$s = -1$	0.14	0.08	0.06	0.34	0.14	0.08
$s = -1.5$	0.14	0.06	0.05	0.35	0.07	0.05

	$r^2 = 0.10$			$r^2 = 0.15$		
alternative	F	S	P_1	F	S	P_1
$s = 1.5$	0.76	1.00	1.00	0.96	1.00	1.00
$s = 1$	0.76	1.00	0.99	0.96	1.00	1.00
$s = 0.5$	0.76	0.99	0.92	0.96	1.00	1.00
$s = 0$	0.75	0.92	0.67	0.96	0.99	0.86
$s = -0.5$	0.76	0.65	0.31	0.96	0.89	0.43
$s = -1$	0.76	0.27	0.10	0.96	0.44	0.13
$s = -1.5$	0.75	0.10	0.05	0.96	0.13	0.05

dimensional data. We compare the power of the locally most powerful test with the power of the test based on $T_{\max}$, the maximum absolute t-statistic, as discussed in Section 5.9.

For this we reverted back to the original high-dimensional data set with $p = 4911$ genes. We generated alternatives $\boldsymbol{\beta}_{m,j}$ for $j = 1, \ldots, p$ and $m = 1, 3, 10, 30$, such that each alternative $\boldsymbol{\beta}_{m,j}$ has the m regression coefficients $\beta_j, \ldots, \beta_{j+m-1}$ equal to 1 and all others equal to zero (taking $\beta_i = \beta_{i-p}$ if $i > p$). Table 5.2 shows the power of the tests based on S and $T_{\max}$ on average against the alternatives $\boldsymbol{\beta}_{m,1}, \ldots, \boldsymbol{\beta}_{m,p}$ with m non-zero regression coefficients. In the simulations the value of σ^2 was taken to be equal for all alternatives $\boldsymbol{\beta}_{m,1}, \ldots, \boldsymbol{\beta}_{m,p}$ and was chosen to get a certain average r^2 over these alternatives. We generated 2 replicates for each of the alternatives, so that each power calculation is based on $2p \approx 10000$ Monte Carlo samples of $\mathbf{y}$.

A complicating factor in this simulation is the lack of a simple and accurate method to find the distribution function of the statistic $T_{\max}$, because of the dependence of the t-statistics. We used simulation to find the α cutoff of

$T_{\max}$ for the design matrix X. The 0.05-cutoff was found at 0.062, using 20000 simulations of $\mathbf{y}$ under the null. Note that this is only slightly below the crude Bonferroni corrected cutoff for p beta$(\frac{1}{2}, \frac{1}{2}(n-1))$ variables, which is at 0.064.

TABLE 5.2: *Monte Carlo power comparison between the locally most powerful test and the test based the maximum of p absolute t-statistics using $\alpha = 0.05$. The reported power values are on average over p different sparse alternatives with m non-zero regression coefficients.*

alter-	$r^2 = 0.01$		$r^2 = 0.02$		$r^2 = 0.05$		$r^2 = 0.10$		$r^2 = 0.20$	
native	S	$T_{\max}$	S	$T_{\max}$	S	$T_{\max}$	S	$T_{\max}$	S	$T_{\max}$
$m = 1$	0.12	0.10	0.17	0.16	0.33	0.40	0.54	0.74	0.76	0.97
$m = 3$	0.11	0.09	0.17	0.14	0.34	0.32	0.55	0.61	0.80	0.90
$m = 10$	0.11	0.09	0.17	0.14	0.35	0.29	0.58	0.54	0.83	0.84
$m = 30$	0.11	0.09	0.17	0.13	0.34	0.28	0.55	0.51	0.80	0.79

The table confirms the theoretical result of Section 5.9 that for sparse alternatives close to the null the score test is slightly superior to the test based on $T_{\max}$. This superiority disappears quite quickly, however, when the single covariate explains a large portion of the variance of $\mathbf{y}$. Looking at decreasingly sparse alternatives, the $T_{\max}$ statistic loses power, as can be expected, but the score test remains more or less stable. What is perhaps most surprising about table 5.2, is that even though the tests are constructed in a very dissimilar way, the average power is still quite similar. The $T_{\max}$ still has good power against not-so-sparse alternatives, while the locally most powerful test has good power against sparse alternatives far from the null.

5.11 Discussion

For testing against a multi-dimensional alternative there are no uniformly most powerful tests. Tests may only be optimal locally for some alternatives, or optimal on average over a region of alternatives. When choosing a test against multi-dimensional alternatives, it is therefore important to consider against which alternatives the chosen test has good power. When constructing such a test, one can use empirical Bayes modelling to design a test which has optimal power on average against a chosen region of alternatives. Thinking about these issues is especially relevant when the data are high-dimensional, because the power of often-used classical tests tends to diminish rapidly when the dimensionality increases.

A drawback of empirical Bayes design of hypothesis tests, is that the construction of the test requires integration over complicated distributions in possibly high-dimensional space. In this paper we have shown in general how to

avoid this problem by using a score test. This test has the property that it is locally most powerful: it has optimal average power in a well-defined neighbourhood of the null hypothesis.

In the linear model, we have shown that this test has good power for many important alternatives, even in the classical low-dimensional situation. The empirical Bayes score test often has better power than the F-test in the situations where there are errors in variables in the design matrix X, when a small set of latent variables influences both the covariates in X and the outcome variable $\mathbf{y}$, or more generally when the large-variance principal components of X explain more of the variance of $\mathbf{y}$ than the small-variance ones. We have also shown that the empirical Bayes score test has good power in truly high-dimensional situations, even against sparse alternatives. If the fraction of variance of $\mathbf{y}$ explained by the covariates is low, the test even outperforms the test based on the maximum absolute t-statistics of all covariates, a test which is designed to find sparse alternatives.

As high-dimensional data become more and more common, so will the need for testing against high-dimensional alternatives. This paper has given a general theoretical outline and presented a concrete example of a model in which the test has good power. But locally most powerful testing in high dimensions has many more potential applications, both in generalized linear models and more generally.

5.12 Proofs of the lemmas

Proof of Lemma 2: To prove Lemma 2, we have to adopt a slightly more formal notation. Shorthand f_θ for the density of $\mathbf{y}$ and let μ be a dominating measure, so that we can write $\mathrm{P}_{\mathbf{y}|\theta}(\mathbf{y} \in A) = \int_A f_\theta \,\mathrm{d}\mu$. Also, let $\mathbf{1}_{\{\cdot\}}$ denote an indicator function.

To prove the existence, we write $w(\theta) = \int_A f_\theta \,\mathrm{d}\mu$, so by the dominated convergence theorem $\frac{\mathrm{d}}{\mathrm{d}\theta} w(\theta_0) = \int_A \frac{\mathrm{d}}{\mathrm{d}\theta} f_{\theta_0} \mathrm{d}\mu < \infty$.

Furthermore, noting that $\frac{\mathrm{d}}{\mathrm{d}\theta} f_{\theta_0} = S^* f_{\theta_0}$, and using $\mathbf{1}_A - \mathbf{1}_B = \mathbf{1}_{A\setminus B} - \mathbf{1}_{B\setminus A}$ twice, we can calculate

$$\begin{aligned}
\tfrac{\mathrm{d}}{\mathrm{d}\theta} w(\theta_0) - \tfrac{\mathrm{d}}{\mathrm{d}\theta} w^*(\theta_0) &= \int_A \tfrac{\mathrm{d}}{\mathrm{d}\theta} f_{\theta_0} \,\mathrm{d}\mu - \int_{S^* \geq k} \tfrac{\mathrm{d}}{\mathrm{d}\theta} f_{\theta_0} \,\mathrm{d}\mu \\
&= \int_{A, S^* < k} S^* f_{\theta_0} \,\mathrm{d}\mu - \int_{A^c, S^* \geq k} S^* f_{\theta_0} \,\mathrm{d}\mu \\
&\leq k \int_{A, S^* < k} f_{\theta_0} \,\mathrm{d}\mu - k \int_{A^c, S^* \geq k} f_{\theta_0} \,\mathrm{d}\mu \\
&= k \int_A f_{\theta_0} \,\mathrm{d}\mu - k \int_{S^* \geq k} f_{\theta_0} \,\mathrm{d}\mu \\
&= k\{w(\theta_0) - w^*(\theta_0)\}.
\end{aligned}$$

The latter term is at most zero whenever condition (i) or (ii) holds. □

Proof of Lemma 3: The assumptions of bounded derivatives combined with the assumption that the distribution of $\mathbf{b}$ is free of τ allows us to interchange limits and integrals in the following calculations. For simplicity we suppress the dependence on $\mathbf{y}$ in the notation.

$$S = \lim_{\tau^2 \downarrow 0} \frac{\frac{\mathrm{d}}{\mathrm{d}\tau^2} \mathrm{E}_{\mathbf{b}}[f(\tau\mathbf{b})]}{\mathrm{E}_{\mathbf{b}}[f(\tau\mathbf{b})]} = \lim_{\tau^2 \downarrow 0} \frac{\mathrm{E}_{\mathbf{b}}[\frac{\mathrm{d}}{\mathrm{d}\tau^2} f(\tau\mathbf{b})]}{\mathrm{E}_{\mathbf{b}}[f(\tau\mathbf{b})]} = \lim_{\tau^2 \downarrow 0} \frac{\mathrm{E}_{\mathbf{b}}[(\frac{\mathrm{d}f(\tau\mathbf{b})}{\mathrm{d}\boldsymbol{\beta}})'\mathbf{b}]}{2\tau \mathrm{E}_{\mathbf{b}}[f(\tau\mathbf{b})]}.$$

The limit evaluates to $0/0$, so we use l'Hôpital's rule to get

$$S = \lim_{\tau^2 \downarrow 0} \frac{\mathrm{E}_{\mathbf{b}}[\mathbf{b}' \frac{\partial^2 f(\tau\mathbf{b})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\mathbf{b}]}{2\mathrm{E}_{\mathbf{b}}[f(\tau\mathbf{b})] + 2\tau \frac{\partial}{\partial\tau^2}\mathrm{E}_{\mathbf{b}}[f(\tau\mathbf{b})]} = \frac{\mathrm{E}_{\mathbf{b}}[\mathbf{b}' \frac{\partial^2 f(\mathbf{0})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\mathbf{b}]}{2\mathrm{E}_{\mathbf{b}}[f(\mathbf{0})]}.$$

Now it only remains to rewrite $\frac{\partial^2 f(\mathbf{0})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} = f(\mathbf{0}) \cdot \{\mathbf{s}\mathbf{s}' - \mathbf{I}\}$. □

Proof of Lemma 4: Assume $w(\mathbf{0}) = \bar{w}(\mathbf{0})$. By Lemma 3 every distribution of $\boldsymbol{\beta}$ with $\mathrm{E}(\boldsymbol{\beta}) = 0$ and $\mathrm{E}(\boldsymbol{\beta}'\boldsymbol{\beta}) \propto \tau^2 I$ leads to the same test statistic and therefore to the same power function. Without loss of generality we can therefore assume that $\bar{w}$ is the power function of the score test in the empirical Bayesian model in which $\boldsymbol{\beta}$ is distributed as $\tau\boldsymbol{\xi}$. By Lemma 2 we have $\frac{\mathrm{d}}{\mathrm{d}\tau^2}\mathrm{E}_{\boldsymbol{\beta}|\tau^2}[w(\boldsymbol{\beta})] \leq \frac{\mathrm{d}}{\mathrm{d}\tau^2}\mathrm{E}_{\boldsymbol{\beta}|\tau^2}[\bar{w}(\boldsymbol{\beta})]$ in $\tau^2 = 0$. The boundedness assumptions of Lemma 3 allow interchanging differentiation and integration, so we get

$$\frac{\mathrm{d}}{\mathrm{d}\tau^2}\mathrm{E}_{\boldsymbol{\beta}|\tau^2}[w(\boldsymbol{\beta})] = \frac{\mathrm{d}}{\mathrm{d}\tau^2}\mathrm{E}_{\boldsymbol{\xi}}[w(\tau\boldsymbol{\xi})] = \mathrm{E}_{\boldsymbol{\xi}}\Big[\frac{\mathrm{d}}{\mathrm{d}\tau^2} w(\tau\boldsymbol{\xi})\Big],$$

both for w and for $\bar{w}$, from which the result follows. □

Chapter 6

Model-based dimension reduction for high-dimensional regression

Abstract

This paper considers the problem of predicting an outcome variable using high-dimensional data. To control the overfit arising from the high-dimensionality one can use dimension reduction methods, which try reduce the set of predictors to a small set of orthogonal linear combinations of the predictor variables, which are subsequently used to predict the outcome. Examples are principal components regression and partial least squares. These methods are usually not motivated by a model and have strictly separated dimension reduction and prediction steps.

This paper looks at dimension reduction for high-dimensional regression from a modelling point of view. We propose a very general model for the joint distribution of the outcome and the predictor variable. This model is based only on the assumption that a set of latent variables exists such that outcome and the predictor variables are conditionally independent given the latent variables. We do not assume that the number of latent variables is known and we allow a very general error structure in the predictor variables. This model allows us to study the dimension reduction and prediction steps jointly.

In this model, we study parameter estimation and prediction in the situation where the number of predictor variables goes to infinity, while the number of samples remains fixed. Based on this analysis, we argue for a doing principal components regression with a relatively small number of components and using only a subset of the predictor variables, selected for their correlation with the outcome variable. This is a variant of the supervised principal components method proposed by Bair et al. on the basis of a much more restrictive model.

This chapter has been submitted as: J. J. Goeman and J. C. van Houwelingen. Model-based dimension reduction for high-dimensional regression.

6.1 Introduction

In recent years high-dimensional data have become increasingly common in many fields of science. This has attracted the attention of the statistical community, resulting in a surge of novel and interesting methodology.

In this paper we consider the basic high-dimensional prediction problem of predicting an outcome y from a vector $\mathbf{x} = (x_1, \ldots, x_p)'$ of predictors. The goal is to predict a new observation y_{new} from an observed data vector $\mathbf{x}_{\text{new}}$. The prediction rule for predicting y_{new} from $\mathbf{x}_{\text{new}}$ is to be constructed using a training sample of size n from the joint distribution of $x_1, \ldots, x_p$ and y. The training data are gathered in an $n \times p$ data matrix X and an $n \times 1$ vector $\mathbf{y}$. The prediction problem is high-dimensional when the number of predictors p is very large, typically larger than the size n of the training sample. The overfit arising from the high number of predictors makes most classical statistical methods unusable.

Many different strategies have been proposed to counter the problem of overfit due to high dimensionality (Hastie et al., 2001). For example, variable selection methods reduce the dimensionality directly by selecting a subset of the predictors to be used for prediction. Shrinkage methods restrain the parameter estimates to prevent overfit (Hoerl and Kennard, 1970; Tibshirani, 1996; Van Houwelingen, 2001). Dimension reduction methods reduce the dimensionality of the prediction problem by using only a small number of orthogonal linear combinations of the original predictor variables (Jolliffe, 2002; Wold et al., 1984). All methods that control overfit in high-dimensional prediction share the property that they reduce the variance of the prediction while introducing bias. The methods differ mainly in the kind of bias introduced.

There is not one strategy or method that is known to be overall superior to all the others. As each method introduces bias, it will tend to perform well especially when the bias introduced is bias 'towards the truth'. For example, a variable selection method can be expected to work best when most of the true regression coefficients are virtually zero. A ridge regression (Hoerl and Kennard, 1970) would most likely work well when most true regression coefficients are in reality small, but not zero. The choice of the method should therefore depend on knowledge or ideas about the underlying 'truth'. Notions about the true relationships between the predictor variables and the outcome can help determine which method is best for which type of data. For a major part, the choice of method should be a modelling issue.

In this paper we investigate high-dimensional regression from a modelling perspective. We formulate a very general model for the joint distribution of the predictors $x_1, \ldots, x_p$ and the outcome y. This model can support the reasoning

behind most methods of the dimension reduction type, such as principal components regression (Jolliffe, 2002), partial least squares (Wold et al., 1984) and more recent methods by Burnham et al. (1999a,b) and Bair et al. (2004).

The joint model we propose is a generalization of factor analysis, a latent variable model often used in psychometry (Bartholomew and Knott, 1999). In this model a set of unobserved latent variables $f_1, \ldots, f_m$ linearly determines both the predictors $\mathbf{x}$ and the outcome y, although both are also subject to error. We assume that the error in $\mathbf{x}$ is independent from the error in y, so that y is conditionally independent of $\mathbf{x}$ given $\mathbf{f} = (f_1, \ldots, f_m)'$. Graphically:

$$\begin{array}{ccc} x_1, \ldots, x_p & & y \\ & \nwarrow \quad \nearrow & \\ & f_1, \ldots, f_m & \end{array} \tag{6.1}$$

We explicitly do not assume that the dimension m of the latent space is known, but only that it is smaller than the sample size n. Furthermore, our model is more general than the factor analysis model in the sense that we do not assume that the error in $x_1, \ldots, x_p$ are independent or identically distributed.

In this model we show how to estimate parameters and how to construct a prediction rule for predicting y from $\mathbf{x}$. We also calculate the mean squared error of prediction for the resulting prediction rule. Based on these calculations, we argue for a prediction rule that first weights the predictors $x_1, \ldots, x_p$ based on their correlation with y and then applies a principal components analysis based on this weighting, using only few components. Essentially, this is a generalization of the method proposed by Bair et al. (2004), which was derived on the basis of a very different and much more restrictive model.

To keep the technicalities limited, we have chosen to limit the discussion in this paper to the linear model, in which the outcome y is continuous and the desired prediction rule is linear. However, the approach is extendable to generalized linear models and we keep the possibility of this extension always in mind.

Before going into the details of the model in section 6.3, in the next section we first investigate some general issues involved in methods of the dimension reduction type and discuss some familiar and less well known methods.

6.2 Bias and variance

Dimension reduction methods combat overfit by replacing the original set of predictor variables $x_1, \ldots, x_p$ with a small set of orthogonal linear combinations of these variables. These linear combinations, sometimes confusingly called 'latent variables', are subsequently used to predict the outcome y.

The most basic dimension reduction method is Principal Components Regression (PCR), which uses the first few principal components of the matrix X for prediction (Jolliffe, 2002, Chapter 8). Other important examples of dimension reduction methods include partial least squares (Wold et al., 1984), various types of continuum regression (Abraham and Merola, 2005; Burnham et al., 1996; Stone and Brooks, 1990) and, more recently, methods proposed by Burnham et al. (1999a,b) and by Bair et al. (2004). There are many more examples. All these methods differ from PCR mainly because they also use the training outcomes $\mathbf{y}$ to determine the principal components, and they differ from each other in the way they use $\mathbf{y}$.

There is a general motivation for all methods of the dimension reduction type. This motivation is best explained through the example of PCR, the mechanics of which are very well known. PCR reduces the matrix X to only its large-variance principal components, ignoring the small-variance principal components for the prediction of y. This will always reduce the variance of the prediction, because the regression coefficients of the small-variance principal components are difficult to estimate accurately. However, it may introduce bias, because the small-variance principal components of X might be important predictors of y. This is the well-known trade-off between bias and variance (Hwang and Nettleton, 2003).

It follows that PCR has best predictive performance when the bias introduced is small, which happens when the small-variance principal components have little or no predictive value for y. A main concern when using PCR is therefore the choice of the number of components: using too many components does not reduce the variance enough, while using too few components may result in missing out those principal components that are important for prediction (Jolliffe, 2002, pp. 173–177). This dilemma is usually solved using methods like cross-validation or AIC, which use the $\mathbf{y}$ data to judge which number of components has the best predictive performance. This 'estimation' of the number of components reintroduces some variance from $\mathbf{y}$ in order to reduce the potential bias.

Other dimension reduction type methods typically introduce the dependence of the choice of the latent variables on $\mathbf{y}$ at an earlier stage. They let the latent variables themselves directly depend on the outcome vector $\mathbf{y}$, so that the important principal components are more likely to be among the first few latent variables selected. Partial Least Squares (Wold et al., 1984) chooses latent variables that have maximum covariance with $\mathbf{y}$ instead of maximum variance. Burnham et al. (1999a,b) choose the latent variables as the eigenvectors corresponding to the largest eigenvalues of the matrix $XX' + \lambda\mathbf{y}\mathbf{y}'$ for some value of λ, instead of those of XX' (see also Tan et al., 2005, for an interesting applica-

tion). Bair et al. (2004) propose a pre-selection of the predictor variables based on their correlation with $\mathbf{y}$, prior to doing principal components. By using $\mathbf{y}$ to choose the latent variables, all these methods essentially reintroduce some variance in order to control the bias.

Ideally, it should follow from a model which method best controls the bias with least reintroduction of variance. Unfortunately, there is little theoretical guidance as to which dimension reduction method is optimal in which situation. The most popular methods of PCR and Partial Least Squares are not based on any model, while the more recent methods of Burnham et al. and Bair et al. are based on relatively restrictive ones (Bair et al., 2004; Burnham et al., 1999a,b).

It is to aid the theoretical discussion on the question which dimension reduction method to use that we formulate our model in section 6.3. This model is the most general model that can motivate a dimension reduction approach. Basically, the only thing it assumes is that a set of latent variables truly exists.

6.3 A basic joint model

The basis of our joint model is the graphical model (6.1), which states that a set of latent variables $f_1, \ldots, f_m$ exists, such that y is conditionally independent of $x_1, \ldots, x_p$ given $f_1, \ldots, f_m$. This assumption implies that prediction of y from $\mathbf{x} = (x_1, \ldots, x_p)'$ should proceed via 'prediction' of the vector $\mathbf{f} = (f_1, \ldots, f_m)'$ of latent variables. This property makes it the basic model underlying methods of the dimension reduction type.

To keep the model simple, we only make a few simple extra assumptions. Firstly, like most dimension reduction methods we assume linearity.

$$\begin{aligned} y &= \mu_y + \boldsymbol{\beta}'\mathbf{f} + \varepsilon \\ \mathbf{x} &= \boldsymbol{\mu} + A'\mathbf{f} + \mathbf{e} \end{aligned} \tag{6.2}$$

Here, μ_y and $\boldsymbol{\mu}$ (a p-vector) are the marginal means of y and $\mathbf{x}$, respectively. The parameters $\boldsymbol{\beta}$ and A are an m-vector and an $m \times p$ matrix of loadings, which determine the relationship between the observed and latent variables. For the error terms $\mathbf{e}$ (a p-vector) and ε we assume zero mean and variance $\mathrm{E}(\mathbf{e}\mathbf{e}') = \Psi$ and $\mathrm{E}(\varepsilon^2) = \sigma^2$. By (6.1), $\mathbf{e}$ and ε are uncorrelated. Further, for deriving maximum likelihood results we shall assume that the distribution of the errors $\mathbf{e}$ and ε is normal. This normality assumption is convenient, but not strictly necessary. The results of this paper can also be rephrased in terms of best linear unbiased prediction (BLUP) under mild conditions. As to the distribution of $\mathbf{f}$, we only assume that it has finite mean and covariance matrix, which can then be taken as $\mathbf{0}$ and I without loss of generality.

It should be remarked that the model, as presented here, is slightly overparametrized if $m > 1$. If W is an $m \times m$ orthogonal matrix, replacing A with WA and $\boldsymbol{\beta}$ with $W\boldsymbol{\beta}$ results in the same joint distribution of $\mathbf{x}$ and y. This means that single parameter values or estimates of A and $\boldsymbol{\beta}$ cannot immediately be interpreted, but only functions of A and $\boldsymbol{\beta}$ that are invariant to multiplication by W. As our prime purpose is prediction, there is no real need to resolve this overparametrization (see Bartholomew and Knott, 1999, for various methods to choose a rotation W).

The terms μ_y and $\boldsymbol{\mu}$ can easily be extended to linear regression functions to incorporate predictors that do not fit in the latent variable structure. The estimation of μ_y and $\boldsymbol{\mu}$ (or their regression extensions) and their use in prediction is straightforward, however, and only complicates notation. For simplicity, we shall therefore assume in the rest of the paper that both μ_y and $\boldsymbol{\mu}$ are known. They can then be taken as zero without loss of generality.

The model presented here is very similar to the factor analysis model frequently used in psychometry. The main difference in the model formulation is that we do not require the matrix Ψ to be diagonal. In psychometry the model is exclusively used in the $p < n$ situation (Bartholomew and Knott, 1999; Magnus and Neudecker, 1999).

We have formulated the above model in terms of the joint distribution of the predictors $\mathbf{x}$ and the outcome y, because this is much more flexible than a model in terms of the conditional distribution of y given $\mathbf{x}$, such as a regression model. One aspect of this flexibility is that the fitted joint model can also be used for prediction when there are missing data in $\mathbf{x}_{\text{new}}$. Another flexible aspect of joint modelling is that it is easier to incorporate theoretical knowledge about relationships between variables into a joint model than in a conditional model. This can already be seen in the assumption (6.1) about the existence of latent variables, which can be directly translated into statements about the joint distribution, but not in statements about the conditional distribution. One may also, for example, have knowledge that certain predictors are uncorrelated with the outcome y. This can be immediately incorporated in the joint model as the statement that the corresponding entries of $A'\boldsymbol{\beta}$ are zero. Such a statement cannot be similarly incorporated into a regression model, because the regression coefficient of a predictor that is uncorrelated with y does not have to be zero. In general, extending the simple joint model with detailed knowledge about dependency relationships of predictors can be easily done using theory on graphical models. In many cases, however, there is only the vague and implicit, but nevertheless important, assumption that the model has a 'simple' structure, which can be translated as the statement that the matrix $A'A$ and the vector $A'\boldsymbol{\beta}$ have many zero elements. We come back to this vague model

structure assumption in Section 6.8.

We shall derive most of the results in this paper using "reverse asymptotics", in which the number of predictors p goes to infinity, while the number of samples n remains fixed or goes to infinity at a much slower rate. We need a few additional assumptions to make these reverse asymptotics well-defined. Essentially, we shall let p grow by simply adding new predictor variables to the vector $\mathbf{x}$. This means that as the parameter space grows, the dimensions of the matrices A and Ψ grow. We impose the following two restrictions:

1. There are constants $0 < k \leq K < \infty$ such that all eigenvalues of Ψ are between k and K for all p.

2. The limit $\lim_{p\to\infty} \frac{1}{p} AA'$ exists and is of full rank m.

First note that by assumption 2 the number of latent variables m does not grow with p. The value of m is therefore assumed to be small relative to p. For simplicity of notation, we also assume that $m < n$ throughout this paper, but this is not a very critical assumption.

The usefulness of the two assumptions 1 and 2 is that they neatly separate the covariance matrix $A'A + \Psi$ of $\mathbf{x}$ into structural covariance ($A'A$) and local covariance (Ψ). The structural covariance is caused by a limited number of latent variables, but each affects a number of the predictors that grows with p. The local covariance is caused by a vector of errors that grows with p, but each independent error term affects only a limited number of predictors.

6.4 Regression

From the model equations (6.2) we can easily derive the joint distribution of the observable variables $\mathbf{x}$ and y. From this joint distribution we can derive any conditional distribution we like. The conditional distribution that is most interesting for prediction is the conditional distribution of y given the whole vector $\mathbf{x}$.

The joint vector $\mathbf{z} = (y, \mathbf{x}')'$ has mean zero and covariance matrix

$$\Sigma_z = \begin{pmatrix} \boldsymbol{\beta}'\boldsymbol{\beta} + \sigma^2 & \boldsymbol{\beta}'A \\ A'\boldsymbol{\beta} & A'A + \Psi \end{pmatrix}.$$

The distribution of $\mathbf{z}$ is normal if the distributions of $\mathbf{x}$ and y are. Therefore, under normality assumptions it follows that y given $\mathbf{x}$ is again normal with a mean that is linear in $\mathbf{x}$:

$$\mathrm{E}[y \mid \mathbf{x}] = \boldsymbol{\gamma}'\mathbf{x} \tag{6.3}$$

where $\gamma = (A'A + \Psi)^{-1}A'\boldsymbol{\beta}$ is a vector of regression coefficients. If normality is not assumed, the equation (6.3) gives the best linear unbiased prediction (BLUP) of y given $\mathbf{x}$.

Using a singular value decomposition on $A\Psi^{-1/2}$ we can also write the regression coefficients γ as

$$\gamma = \Psi^{-1}A'(A\Psi^{-1}A' + I)^{-1}\boldsymbol{\beta}. \tag{6.4}$$

This expression will turn out to be more useful. It is computationally easier, as it does not involve inversion of the complicated $p \times p$ matrix $A'A + \Psi$.

6.5 Easy prediction

We solve the prediction problem in stages. In the previous section we have seen that in the trivial situation that all parameters are known we should simply use equation (6.3) to predict y_{new} from $\mathbf{x}_{\text{new}}$. In this section we study the still relatively easy situation in which the structural parameters A and $\boldsymbol{\beta}$ are known (and hence m is known), but only the error covariance Ψ is not known.

The great difficulty in this situation is that it is almost impossible to estimate Ψ accurately enough from the training data. This can already be inferred from the fact that estimating Ψ means estimating p^2 parameters, while only $n(p+1)$ degrees of freedom are available in the training set. All commonly used estimates of a covariance matrix therefore result in extremely ill-conditioned estimates of Ψ. This ill-conditionedness causes great problems because prediction involves the inverse of the matrix Ψ.

Ledoit and Wolf (2004) studied the general problem of estimating covariance matrices with high-dimensional data. They proposed an estimate that is a linear combination of the naive maximum likelihood estimate and a chosen matrix Θ (typically the identity matrix). This biases the estimated covariance matrix towards the chosen matrix Θ. It also shrinks the eigenvalues of the estimate towards each other and forces them all to be positive, so that the resulting estimate is always invertible. However, as the dimension p of the covariance grows relative to the sample size n, the bias becomes dominant in the estimate of Ledoit and Wolf. In the limit $p \to \infty$, for fixed n, the optimal estimate is all bias and no variance. This essentially means that it is hopeless to try to estimate the covariance matrix Ψ in the $p \to \infty$, n fixed situation that we are interested in.

It does not mean, however, that prediction of y_{new} from $\mathbf{x}_{\text{new}}$ is hopeless. We can simply use the limiting estimate of Ledoit and Wolf, which is all bias and no variance and take any fixed matrix Θ as an 'estimate' for Ψ^{-1}. Such a

Θ is not truly an estimate so we shall refer to it as a *surrogate*. It should have similar properties for $p \to \infty$ as Ψ^{-1}. The properties we need are

1. There are constants $0 \leq l \leq L < \infty$ such that all eigenvalues of Θ are between l and L for all p.

2. The limit
$$G = \lim_{p\to\infty} \frac{1}{p} A\Theta A'$$
exists and $\left(\frac{1}{p}A\Theta A' - G\right)^2 = O(p^{-1})$.

3. The limit
$$\tau^2 = \lim_{p\to\infty} \frac{1}{p} \text{trace}(\Psi\Theta)$$
exist and $\left(\frac{1}{p}\text{trace}(\Psi\Theta) - \tau^2\right)^2 = O(p^{-1})$.

Note that we allow $l = 0$ in property 1, which is importantly different from assumption 1: unlike Ψ^{-1}, the matrix Θ is allowed to be singular. Properties 2 and 3 only serve to rule out some very exotic or degenerate choices of Θ.

Plugging in the surrogate Θ for Ψ^{-1} in the prediction rule (6.3) results in a prediction rule with regression coefficients $\Theta A'(A\Theta A' + I)^{-1}\boldsymbol{\beta}$ instead of $\boldsymbol{\gamma}$. As the identity matrix I is negligible next to $A\Theta A'$ if p is large, we can replace this by the simpler expression

$$\boldsymbol{\gamma}_\Theta = \Theta A'(A\Theta A')^{-1}\boldsymbol{\beta}. \tag{6.5}$$

This vector of regression coefficients $\boldsymbol{\gamma}_\Theta$ will usually be drastically different from the vector $\boldsymbol{\gamma}$ of 'true' regression coefficients. But, surprisingly, the resulting predictions will be precisely the same if p is large enough. This is stated in Theorem 1.

Theorem 1 *If the matrix G has full rank m, we have*

$$\text{E}[(\boldsymbol{\gamma}_\Theta'\mathbf{x} - \boldsymbol{\beta}'\mathbf{f})^2 \mid \mathbf{f}] = O(p^{-1}).$$

for every $\mathbf{f}$.

The theorem says that in the limit $p \to \infty$ using any surrogate Θ for Ψ^{-1} will give perfect predictive performance, as $\boldsymbol{\beta}'\mathbf{f}$ is the expectation of $\mathbf{y}$ given $\mathbf{f}$. We can say that the prediction $\boldsymbol{\gamma}_\Theta'\mathbf{x}$ gives a p-consistent estimate of the mean of $\mathbf{y}$, where p-consistency is defined completely analogously to ordinary consistency. By the property of the best linear unbiased prediction, using the true Ψ^{-1} for Θ would still give a prediction with least variance for finite p, but the extra

variance caused by using a 'wrong' Θ disappears when p grows to infinity. The difference will therefore be negligible relative to many other sources of estimation and prediction error that we will encounter later.

The easiest way to understand the role of the surrogate Θ is view it as a weighting matrix. In the setup (6.1) almost every predictor variable in $\mathbf{x}$ carries information on $\mathbf{f}$ (and through $\mathbf{f}$ on y). Most of this information is redundant, however, if p is much larger than m: a good choice of m predictor variables with small error would be enough to summarize almost all information in $\mathbf{x}$ on y. If we do use all predictor variables, we are free to choose our own weighting to aggregate their information on $\mathbf{f}$. The optimal weighting for finding $\mathbf{f}$ from $\mathbf{x}$ is Ψ^{-1}, by (6.3), which weights each predictor variable inversely to its error variance. However, by Theorem 1 any other weighting which spreads the weight over many predictor variables will do equally well. The intuitive reason for this is the abundance of information on $\mathbf{f}$ in $\mathbf{x}$ if p is large.

The interpretation of Θ as a weighting will be useful throughout this paper. Consequently, all estimates and predictors involving Θ proposed in this paper will be invariant to multiplication of Θ by a constant. Hence only the relative magnitudes of the entries of Θ are important, as is appropriate in a weighting matrix.

The fact that different methods may have very different regression coefficients and still result in very similar predictions has been noted before (Burnham et al., 2001). It should be seen as a warning against using regression coefficients as a basis for modelling or interpretation, and another argument for modelling in terms of the joint distribution.

6.6 Estimation

There is a big difference between estimation of Ψ and estimation of A, the other large matrix of parameters. Estimation of Ψ is difficult, even when A is known, but estimation of A is relatively easy, even if Ψ is unknown. We show this in this section.

Estimation and finding the prediction rule will be based on a training sample: an $n \times p$ matrix X of predictor variables and a corresponding n-vector $\mathbf{y}$ of outcome variables. Call F the $n \times m$ matrix of the realizations of the unobserved latent variables $\mathbf{f}$ for the individuals in the training sample.

We first look at the situation in which both Ψ and m are known. In that situation we can use a standard theorem from factor analysis, formulated by Magnus and Neudecker (1999, Chapter 17, Section 12), which we rephrase here as Theorem 2.

Theorem 2 (Magnus and Neudecker) *Let $\tilde{\Lambda}$ be the $m \times m$ diagonal matrix of the m largest eigenvalues of the matrix $\tilde{S} = \frac{1}{p+1}\{X\Psi^{-1}X' + \sigma^{-2}\mathbf{y}'\mathbf{y}\}$ and let $\tilde{U}$ be the $n \times m$ orthogonal matrix with the corresponding eigenvectors.*
If Ψ and m are known and the distribution of $\mathbf{f}$ is normal, maximum likelihood estimates of A and $\boldsymbol{\beta}$ are given by

$$\begin{aligned} \tilde{A} &= n^{-1/2}\tilde{\Upsilon}^{1/2}\tilde{U}'X \\ \tilde{\boldsymbol{\beta}} &= n^{-1/2}\tilde{\Upsilon}^{1/2}\tilde{U}'\mathbf{y} \end{aligned}$$

where $\tilde{\Upsilon}$ is the $m \times m$ diagonal matrix with $\tilde{\Upsilon}_{ii} = \max(0, 1 - \frac{n}{p}\tilde{\Lambda}_{ii}^{-1})$.

Note that the maximum likelihood estimate is not unique if $m > 1$, due to the overparametrization metioned in Section 6.3. If W is an $m \times m$ orthogonal matrix, then $W\tilde{A}$ and $W\tilde{\boldsymbol{\beta}}$ are also maximum likelihood estimates.

The proof of Theorem 2 is given in Magnus and Neudecker (1999, Chapter 17, Section 12). Their Theorem and its proof are phrased in the context of traditional factor analysis. Consequently, they use the additional implicit assumptions that $p < n$ and that Ψ is diagonal. However, it is a simple exercise to show that their proof also holds for $p \geq n$ and general positive definite Ψ.

The maximum likelihood estimates of Theorem 2 are not immediately useable as they involve the matrix Ψ^{-1}, which cannot be estimated in high-dimensional data. The standard techniques used in factor analysis to estimate A, $\boldsymbol{\beta}$ and Ψ simultaneously (Magnus and Neudecker, 1999, Chapter 17, Sections 12–14) cannot therefore be used. However, we can use the same trick that was used in Section 6.5, simply replacing Ψ^{-1} by a well-conditioned surrogate Θ.

The estimates $\tilde{A}$ and $\tilde{\boldsymbol{\beta}}$ also involve the unknown σ^2. The influence of σ^2 on the estimate disappears very quickly, however, when p becomes large. In the $p \to \infty$, n fixed situation, there is no loss when we simply replace σ^{-2} by zero, just as we replace Ψ^{-1} by Θ. This neglect of y in the estimation of A stands in sharp contrast the method of Burnham et al. (1999b), which makes σ^{-2} an important tuning parameter. Burnham's method is very sensible in the applications it was designed for, where both $\mathbf{x}$ and y are high-dimensional. It is also a sensible strategy when p is small (Wall and Li, 2003). However, in applications with high-dimensional $\mathbf{x}$ but univariate y, the influence of y on the estimate $\tilde{A}$ should always disappear when $p \to \infty$ and $\sigma^2 > 0$.

Aside from the unknown Ψ, there is also the unknown value of m. Just like estimation of Ψ, accurate estimation of the true value of m is very difficult, if not impossible. The solution we propose to the problem of the unknown m is similar to the problem of the unknown Ψ. We simply replace the unknown m with a chosen $q \geq 0$. This q is not trying to estimate the true number of latent

variables; it is the number of latent variables we use for prediction. We shall assume that $q \leq m$, but this is only to keep notation and proofs simple.

We propose to estimate A and $\boldsymbol{\beta}$ as

$$\begin{aligned} \hat{A} &= n^{-1/2}\hat{U}'X \\ \hat{\boldsymbol{\beta}} &= n^{-1/2}\hat{U}'\mathbf{y} \end{aligned} \tag{6.6}$$

Where $\hat{U}$ and is defined as the $n \times q$ orthogonal matrix with the eigenvectors corresponding to the q largest eigenvalues of the matrix

$$S = \frac{1}{p}X\Theta X'.$$

The motivation for these estimates of A and $\boldsymbol{\beta}$ will come from asymptotic arguments very similar to the arguments used in Section 6.5. Theorem 3 shows that when $p \to \infty$ it makes no difference whether we use the matrix $\tilde{S}$ involving the true Ψ and σ^2 or the matrix S involving the surrogate Θ.

Because the matrix S does not involve y, the estimates $\hat{A}$ and $\hat{\boldsymbol{\beta}}$ can easily be adapted to the situation where y does not have a normal distribution, but depends on $\mathbf{f}$ through a generalized linear model. In that case $\boldsymbol{\beta}$ is estimated as the regression coefficients of the Generalized Linear Model with outcome y and $n^{1/2}\hat{U}$ as the design matrix (see also Bair et al., 2004).

By $\|\cdot\|$ we denote the Frobenius norm: $\|C\| = \text{trace}(C'C)^{1/2}$.

Theorem 3 *If $q \leq \text{rank}(G)$, there is an $m \times q$ semi-orthogonal matrix V, depending on $\hat{A}$ and $\tilde{A}$, such that, almost surely in F,*

$$\begin{aligned} \mathrm{E}\big[\, \|\hat{\boldsymbol{\beta}} - V'\tilde{\boldsymbol{\beta}}\|^2 \mid F\big] &= O(p^{-1}) \\ p^{-1}\mathrm{E}\big[\, \|\hat{A} - V'\tilde{A}\|^2 \mid F\big] &= O(p^{-1}) \end{aligned}$$

If $q = m$ or if the q-th and $q+1$-th eigenvalues of $G = \lim_{p\to\infty} p^{-1}A\Theta A'$ are distinct, the result holds uniformly in n.

Theorem 3 states that if p is large enough and $q = m$, the estimates $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ and $\hat{A}$ and $\tilde{A}$ differ only by a rotation. As a rotation of $\tilde{A}$ is also a maximum likelihood estimate of A, this means that if $q = m$, the estimates (6.6) are asymptotically equivalent (for $p \to \infty$) to the maximum likelihood estimates of A and $\boldsymbol{\beta}$. The formulae of (6.6) thus give a good complementary estimation procedure to the traditional iterative estimation procedure for A and $\boldsymbol{\beta}$ used in factor analysis models (Magnus and Neudecker, 1999). The procedure is complementary because it works precisely in the high-dimensional situation in which the traditional procedure does not.

When interpreting Theorem 3 one must keep in mind that estimates which differ only by a $q \times q$ rotation matrix can be considered as equivalent because they lead to the same estimated joint distribution. If $q = m$, therefore, estimates $\hat{A}$ and $\hat{\boldsymbol{\beta}}$ for different surrogates Θ are all asymptotically equivalent because in the limit $p \to \infty$ they only differ by a $q \times q$ rotation matrix V. However, if $q < m$, estimates $\hat{A}$ and $\hat{\boldsymbol{\beta}}$ for different surrogates are different even in the limit $p \to \infty$, because then the $m \times q$ matrices V differ from each other by more than a $q \times q$ rotation.

The effect of the surrogate Θ on the estimate is best understood in terms of weighted principal components. The matrix $\hat{U}$ is the standardized matrix of the first q principal components of the weighted data matrix $X\Theta^{1/2}$. If $p \to \infty$ the principal components of $X\Theta^{1/2}$ will be the same as those of $FA\Theta^{1/2}$ (see the proof of Theorem 3). The combined span of the first m principal components of $FA\Theta^{1/2}$ does not depend on Θ, as it is simply the column span of F. However, the principal component variances do depend on Θ. Therefore the span of the q principal components with largest variance does depend on Θ. For finite p, the minimum variance estimate still remains the estimate with $\Theta = \Psi^{-1}$ (Wentzell et al., 1997).

We have so far only proved that we can define alternative estimates of A and $\boldsymbol{\beta}$ which are non-iterative and are as good as the maximum likelihood estimates for known Ψ if $p \to \infty$. Of course, this only shows that the proposed estimate is good if the maximum likelihood estimate itself is good, which may not be the case if m is close to n. However, in the next section we show that the estimate $\hat{A}$ has good properties when used for prediction.

6.7 Prediction

We want to predict the outcome y from the predictors $\mathbf{x}$ in the model (6.2) in which all parameters are unknown. We propose to combine the results from the previous sections by plugging the estimates of Section 6.6 into the prediction rule of Section 6.5.

This would lead to the prediction rule predicting y_{new} with $\tilde{\boldsymbol{\gamma}}'\mathbf{x}_{\text{new}}$ where the regression coefficients are

$$\tilde{\boldsymbol{\gamma}} = \Theta\hat{A}'(\hat{A}\Theta\hat{A}')^{-1}\hat{\boldsymbol{\beta}}.$$

Compare (6.5). Note that if we take $\Theta = I$ the regression coefficients $\tilde{\boldsymbol{\gamma}}$ are exactly the regression coefficients from a principal components regression. This can be seen by writing $\tilde{\boldsymbol{\gamma}}'\mathbf{x}_{\text{new}} = \mathbf{y}'\hat{U}\hat{U}'(XX')^{-1}\hat{U}\hat{U}'X\mathbf{x}_{\text{new}}$. If we take $\Theta \neq I$, the vector $\tilde{\boldsymbol{\gamma}}$ is the vector of regression coefficients of a weighted principal components regression.

This prediction rule needs an adjustment for the large p, small n situation that we are interested in. It turns out that when n is small, the prediction $\tilde{\gamma}'\mathbf{x}_{\text{new}}$ tends to induce too much shrinkage. This excessive shrinkage is caused by overfit of $\hat{A}$ to the noise in X. This overfit causes $\hat{A}\Theta\hat{A}'$ to be systematically larger than $A\Theta A'$: we have

$$\lim_{p\to\infty} \frac{1}{p}\hat{A}\Theta\hat{A}' = \frac{1}{p}A\Theta A' + \frac{1}{n}\tau^2 I,$$

where $\tau^2 = \lim_{p\to\infty} p^{-1}\text{trace}(\Theta\Psi)$. We propose to remedy this overfit in the small n situation by subtracting an estimate of $\tau^2 I$. This leads to the following prediction rule:

$$\hat{y}_{\text{new}} = \hat{\gamma}'\mathbf{x}_{\text{new}} \tag{6.7}$$

with regression coefficients

$$\hat{\gamma} = \Theta\hat{A}'(\hat{A}\Theta\hat{A}' - \frac{p}{n}\hat{\tau}^2 I)^{-1}\hat{\beta}. \tag{6.8}$$

The difference between the prediction based on $\hat{\gamma}$ and on $\tilde{\gamma}$ disappears very quickly when n becomes large, but can be important in the small n situation. It is essential for the asymptotic $p \to \infty$ result in Theorem 4.

The adjusted prediction rule (6.7) involves a new quantity $\hat{\tau}^2$, which is an estimate of τ^2. We define $\hat{\tau}^2$ as

$$\hat{\tau}^2 = \frac{1}{n-r}\text{trace}(S\hat{Q})$$

where $\hat{Q}$ is the rank $n-r$ projection matrix for projection on the eigenvectors corresponding to the $n-r$ smallest eigenvalues of S. It is shown in the appendix that $\hat{\tau}^2 \to \tau^2$ as $p \to \infty$ uniformly in n whenever $r > m$, so that $\hat{\tau}^2$ is a good estimate of τ^2 when p is large, even when n is small. It is easily checked that the matrix $\hat{A}\Theta\hat{A}' - \frac{p}{n}\hat{\tau}^2 I$ in (6.8) is always positive (semi-)definite.

The most important property of the prediction $\hat{y}_{\text{new}}$ is Theorem 4.

Theorem 4 *If $p \to \infty$,*

$$\text{E}[\,(\hat{y}_{\text{new}} - \tilde{y}_{\text{new}})^2 \mid F] = O(p^{-1})$$

almost surely, where

$$\tilde{y}_{\text{new}} = \mathbf{y}'F(F'F)^{-1/2}VV'(F'F)^{-1/2}\mathbf{f}_{\text{new}}$$

and V is the $m \times q$ semi-orthogonal matrix of the eigenvectors of the q largest eigenvalues of the matrix $(F'F)^{1/2}G(F'F)^{1/2}$. If $q = m$ or if the q-th and $q+1$-th eigenvalues of $G = \lim_{p\to\infty} p^{-1}A\Theta A'$ are distinct, the result holds uniformly in n.

Note that if $q = m$, the $\tilde{y}_{\text{new}}$ in Theorem 4 is simply the least squares prediction of $\mathbf{y}_{\text{new}}$ in the situation where F and $\mathbf{f}_{\text{new}}$ are observed variables instead of latent variables. If $q < m$, the projection matrix VV' introduces bias into the prediction, because not all latent variables are used for prediction.

The prediction based on $\tilde{y}_{\text{new}}$ is not perfect: it has bias if $q < m$, and it also has a prediction variance. The bias and the prediction variance of $\tilde{y}_{\text{new}}$ do not vanish when $p \to \infty$. By Theorem 4, therefore, the prediction error of $\hat{y}_{\text{new}}$ will be dominated by the prediction error of $\tilde{y}_{\text{new}}$ if p is large and n is small, while the difference between $\hat{y}_{\text{new}}$ and $\tilde{y}_{\text{new}}$ will be negligible. We must therefore study the prediction error of our prediction $\hat{y}_{\text{new}}$ through the prediction error of $\tilde{y}_{\text{new}}$

The variance and bias of the prediction $\tilde{y}_{\text{new}}$ are easy to calculate conditional on F and $\mathbf{f}_{\text{new}}$. The variance $v^2 = \text{Var}[\tilde{y}_{\text{new}} \mid F, \mathbf{f}_{\text{new}}]$ is

$$v^2 = \sigma_y^2 \|V'(F'F)^{-1/2} f_{\text{new}}\|^2 \tag{6.9}$$

and the bias $b = \text{E}[\tilde{y}_{\text{new}} - \boldsymbol{\beta}'\mathbf{f}_{\text{new}} \mid F, \mathbf{f}_{\text{new}}]$ is

$$b = \boldsymbol{\beta}'(F'F)^{1/2}(I - VV')(F'F)^{-1/2}\mathbf{f}_{\text{new}}. \tag{6.10}$$

These results would be easier to interpret if we could take the expectation of the variance and of the squared bias over F and $\mathbf{f}_{\text{new}}$. However, there is no analytical solution for these expectations for finite n, mainly because of the complicated dependence between V and F. However, the small n behaviour of the bias and variance of $\tilde{y}_{\text{new}}$ is very similar to the large n behaviour. As $n \to \infty$,

$$\begin{aligned} n\text{E}[v^2] &\to q\sigma^2 \\ \text{E}[b^2] &\to \boldsymbol{\beta}'(I - VV')\boldsymbol{\beta} \end{aligned} \tag{6.11}$$

where V is now a matrix of eigenvectors of G.

A trade-off between v^2 and b determines the performance for different q and Θ of the prediction rule (6.7) proposed above. Some interesting conclusions can be drawn.

In the first place it is not always optimal to take $q = m$, even if m is known. Reducing q below m will usually increase the squared bias b^2, but it will always decrease the prediction variance v^2. If the decrease in variance is larger than the increase in bias, the prediction rule with smaller q will be the better one. This shows that it is often not worthwhile to try to estimate m, as knowledge of m does not necessarily lead to improved prediction accuracy.

There is a value of q somewhere between 0 and m where the trade-off between bias and variance leads to optimal prediction error. The location of this optimum depends on n: larger n means smaller prediction variance v^2, but not

a smaller bias b. Therefore the optimal trade-off will be different. Typically, a larger n will lead to a larger optimal q. The location of the optimal q also depends on Θ, as the choice of Θ also affects the size of the bias.

Unlike the choice of q, the choice of Θ does not involve a trade-off between bias and variance. The reason for this is that, as remarked above, the choice of Θ only influences the distribution of the bias b of the prediction, but not of the variance v^2. Therefore, we can choose a Θ that makes the bias small, without automatically incurring a large prediction variance. Furthermore, if we can find a Θ that produces small bias even for small values of q, we can choose q small, thereby indirectly reducing the variance v^2.

The prediction rule (6.7) therefore has a predictive performance that is as good, if $p \to \infty$, as the prediction $\tilde{y}_{\text{new}}$ of an 'oracle' that observes the unobservable latent variables. We get this predictive performance for any Θ when we choose $q = m$. A smart choice of Θ and q may even result in a better predictive performance than $\tilde{y}_{\text{new}}$. However, whereas finding the optimal q for fixed Θ is doable, finding an optimal Θ is a daunting task even for fixed q, due to the enormously large search space. We discuss one promising strategy in the next section.

6.8 Supervised Principal Components

The model of this paper seemed to be especially designed for supporting Principal Components Regression. In the previous section, however, we have already shown that even when m is known, Principal Components Regression with m components is not necessarily the optimal prediction rule. In this section we show that there are good arguments in favour of a data-driven way to choose Θ, which leads to a variant of the Supervised Principal Components method recently proposed by Bair et al. (2004). Due to the increased complexity of a having a random Θ, exact statements are difficult to prove and this section will be more informal than the previous sections.

Which choice of Θ induces the smallest bias? This is easiest to see in the large n situation of equation (6.11), but it holds similarly for the more complex situation of equation (6.10). The bias is small if $\boldsymbol{\beta}'(I - VV')\boldsymbol{\beta}$ is small, which happens if $\boldsymbol{\beta}$ is in the span of the eigenvectors of the q largest eigenvalues of $G = \lim_{p\to\infty} p^{-1} A\Theta A'$. If Θ is diagonal, with i-th diagonal element θ_i, we have

$$G = \frac{1}{p} \sum_{i=1}^{p} \theta_i \boldsymbol{\alpha}_i \boldsymbol{\alpha}_i'$$

where $\boldsymbol{\alpha}_i$ is the i-th column of A (an m-vector). To push the eigenvectors of the large eigenvalues of G in the direction of $\boldsymbol{\beta}$, we should give a large weight θ_i to

predictor variables which have a large correlation between $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}$, and vice versa give a small weight θ_i to variables which have a low correlation.

As A and $\boldsymbol{\beta}$ are not known, a suitable proxy to the correlation between $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}$ is the correlation between $\mathbf{x}_i$ and $\mathbf{y}$, where $\mathbf{x}_i$ is the i-th column of X. Predictor variables which have a large correlation between $\mathbf{x}_i$ and $\mathbf{y}$ tend to have a large correlation between $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}$, combined with a small error variance.

This suggests a very simple method for choosing Θ, which can be expected to lead to a small bias b even when q is small. This method finds a limited number of predictors which have the largest correlation with the outcome. It gives these predictors equal weight ($\theta_i = 1$), and all other predictors zero weight ($\theta_i = 0$). Using this Θ in combination with a small value of q in the prediction rule (6.7) gives precisely the Supervised Principal Components method proposed by Bair et al. (2004), except for the improvement based on $\hat{\tau}^2$ that we proposed in equation (6.7). Variants of Supervised Principal Components can easily be thought of, for example taking θ_i equal to the squared correlation between $\mathbf{x}_i$ and $\mathbf{y}$. But these variants are not essentially different, so we study the original proposal by Bair et al. (2004).

For such a "supervised" data-driven choice of a surrogate the discussion on variance and bias in Section 6.7 is not strictly valid, because due to the data-dependent construction of Θ the matrix V is not independent of the error in the outcomes $\mathbf{y}$. Therefore the distinction between bias and variance is not the same as it is for fixed Θ. The data-dependent choice of Θ will, therefore, usually not just reduce the bias but also increase the variance of the prediction. There is no explicit expression for bias and variance in this case.

Similarly, Theorem 4 does not hold for a Θ that depends on X. However, we can conclude from Theorem 4 that the prediction result is extremely robust against the choice of Θ, because that theorem holds *for every* fixed Θ. If Θ spreads the weight over many predictor variables, the first few principal components are principal components of a large subset of the columns of the matrix X. By Theorem 4, prediction based on the principal components of any such large subset is indistinguishable from prediction based on the true latent variables $\mathbf{f}$. Because this result is result holds for all fixed Θ, we can also expect it to hold for a data-dependent Θ, as long as Θ is not 'too data-dependent', but still selects a large subset of the predictors. We can therefore expect good prediction results if we combine a data-dependent Θ as in Supervised Principal Components regression, provided Θ does not put all its weight on a small subset of the predictors.

Using a data-dependent weighting Θ becomes even more attractive if n is large, because the variance of Θ will disappear as n grows. For large n the approximations (6.11) of bias and variance are asymptotically still valid. For

large n, the Supervised Principal Components choice of Θ combined with a small value of q will lead to a small prediction variance and a small bias.

Supervised principal components uses the information in $\mathbf{y}$ to choose a weighting of the predictor variables to be used to calculate principal components, but does not let $\mathbf{y}$ influence these components directly. Therefore it has the desirable tendency to choose those principal components that have a relationship with y. However it is relatively robust against pushing the principal components into the directions of the error of $\mathbf{y}$, because only a small number of predictors will have their errors correlated with the error of $\mathbf{y}$. These will typically be too few in number to have a large influence on the first few principal components.

In many applications, for example in microarray data, there is the implicit assumption that many of the predictor variables are not correlated with the outcome y. This can be translated as the assumption that many entries of $A'\boldsymbol{\beta}$ are exactly zero. Predictors with $\boldsymbol{\alpha}_i'\boldsymbol{\beta}$ zero should ideally be given zero weight when doing the principal components calculations, because they only add weight to unimportant principal component directions. A good way to filter out the predictors with $\boldsymbol{\alpha}_i'\boldsymbol{\beta} = 0$ is to remove the predictors with low correlation with the outcome. This assumption that the model is structured in the sense that many predictor variables are uncorrelated with the outcome therefore also gives an argument in favour of using Supervised Principal Components.

6.9 Application

In this section we present a simulation study to illustrate the findings above. To make the simulations realistic, they are based on a microarray gene expression data set of breast cancer patients by Van de Vijver et al. (2002).

Our simulations complement the extensive simulation and data analysis presented by Bair et al. (2004). In their analysis they compare Supervised Principal Components to several other commonly used methods, showing that Supervised Principal Components has good performance relative to other methods. In these simulations they always chose $q = 1$, but chose the number of selected genes using cross-validation.

Our simulation setup is as follows. We used the data by Van de Vijver et al. (2002) in order to obtain realistic values for the parameters A and Ψ. In the original data there were 295 patients and 24,481 predictors, of which 293 patients and 23,862 predictors remained after removing some of the patients and predictors due to missing values. We took 10 as the true value of m and estimated A using the procedures described in Section 6.6. To obtain more highly structured relationships between the latent variables $\mathbf{f}$ and the observed $\mathbf{x}$, as

is biologically expected for microarray data, we applied hard thresholding to A, setting all but the 10% largest absolute values of $\hat{A}$ to zero. We took Ψ as a diagonal matrix, the standard method-of-moments estimate of a diagonal Ψ given A:

$$\hat{\Psi} = \frac{1}{n-m}\text{diag}(X'X - n\hat{A}'\hat{A}).$$

These values of A and Ψ were subsequently used as the true values in the simulation experiments that follow.

In our simulation experiment we compared the performance of our version of the Supervised Principal Components method for all values of q and for various choices of the number s of selected predictors. We investigated what values of q and s tend to do well in prediction. We also investigated how this answer depends on $\boldsymbol{\beta}$. We generated datasets based on different choices of $\boldsymbol{\beta}$, each letting y depend on a different latent variable. In the simulation k, we define $\boldsymbol{\beta}_k = e_k$, the k-th standard basis vector. Depending on the choice of k, between 5,742 ($k = 2$) and 989 ($k = 10$) predictors were correlated with the outcome y. We always chose $\sigma^2 = 1$, so that at most 50% of the variance of y can be predicted from $\mathbf{x}$.

Based on the A and Ψ values chosen above, we generated a training data set of X and $\mathbf{y}$ of $n = 20$ patients and $p = 23{,}862$ predictors using a matrix F of latent variables and based on the model equations (6.2). The performance of the prediction rules created on this data-set was evaluated using an independently generated test set of $n = 100$ patients. The performance of the method depends on the realized values of the latent variables F in the training and test set. Therefore we repeated the whole procedure of generating training and test sets 100 times to be able to average out the effects of F. The construction of Θ for the pre-selection of genes was done for each data set separately.

In the tables we give the results of the simulations for the five different choices of $\boldsymbol{\beta}$. It has to be remarked that due to the thresholding the rows of A are not orthogonal anymore and that the ordering of the norms of the rows may have changed. However, the value of k still gives a good indication how much of the variance of $\mathbf{x}$ is explained by the latent variable $\mathbf{f}'\boldsymbol{\beta}$: the larger k, the more variance $\mathbf{f}'\boldsymbol{\beta}$ explains.

The results of the simulations are given in Tables 1–5. From these simulations we see clearly that it is highly worthwhile to make a pre-selection of predictors prior to doing Principal Components Regression. The lowest prediction error of the methods that select fewer than 23,862 predictors is in all cases lower than the prediction error of the 'plain' Principal Components Regression. Furthermore, this optimum is attained at much lower values of q. However, taking a too small subset to do the Principal Components Regression leads to

TABLE 6.1: *Mean squared prediction error for Supervised Principal Components for various choices of the number of components used and of the number op predictors selected. Simulated data in which the first latent variable is related to y and to 4,201 predictor variables.*

# pred.	# components q									
	1	2	3	4	5	6	7	8	9	10
23,862	0.44	0.28	0.24	0.23	0.22	0.22	0.22	0.22	0.22	0.22
10,000	0.18	0.17	0.18	0.19	0.21	0.22	0.23	0.24	0.24	0.25
5,000	0.14	0.16	0.18	0.20	0.22	0.23	0.23	0.24	0.25	0.25
1,000	**0.12**	0.14	0.16	0.18	0.20	0.20	0.21	0.21	0.22	0.22
200	0.13	0.14	0.16	0.17	0.18	0.20	0.21	0.22	0.22	0.24
50	0.18	0.19	0.20	0.22	0.23	0.25	0.27	0.28	0.30	0.32

TABLE 6.2: *Mean squared prediction error for Supervised Principal Components for various choices of the number of components used and of the number op predictors selected. Simulated data in which the second latent variable is related to y and to 5,742 predictor variables.*

# pred.	# components q									
	1	2	3	4	5	6	7	8	9	10
23,862	0.80	0.43	0.35	0.33	0.31	0.31	0.31	0.31	0.31	0.31
10,000	0.31	0.25	0.26	0.28	0.29	0.30	0.31	0.32	0.33	0.33
5,000	0.22	0.23	0.26	0.28	0.29	0.30	0.31	0.32	0.32	0.32
1,000	0.14	0.22	0.24	0.26	0.26	0.27	0.27	0.28	0.28	0.29
200	**0.14**	0.19	0.21	0.23	0.24	0.25	0.26	0.27	0.28	0.28
50	0.19	0.22	0.23	0.24	0.25	0.27	0.28	0.30	0.31	0.33

TABLE 6.3: *Mean squared prediction error for Supervised Principal Components for various choices of the number of components used and of the number op predictors selected. Simulated data in which the third latent variable is related to y and to 2,718 predictor variables.*

# pred.	# components q									
	1	2	3	4	5	6	7	8	9	10
23,862	0.98	0.86	0.61	0.51	0.47	0.46	0.45	0.44	0.44	0.43
10,000	0.60	0.46	0.38	0.39	0.40	0.41	0.42	0.43	0.43	0.44
5,000	0.45	0.39	0.36	0.38	0.40	0.41	0.42	0.43	0.43	0.43
1,000	0.33	0.32	0.33	0.35	0.38	0.38	0.39	0.40	0.40	0.41
200	0.31	**0.31**	0.33	0.34	0.35	0.36	0.37	0.38	0.39	0.40
50	0.37	0.35	0.36	0.37	0.38	0.40	0.40	0.43	0.44	0.46

inflated prediction error, because the selected data set cannot be called high-dimensional anymore. This is all precisely as expected from the discussion in

TABLE 6.4: *Mean squared prediction error for Supervised Principal Components for various choices of the number of components used and of the number op predictors selected. Simulated data in which the sixth latent variable is related to y and to 1,731 predictor variables.*

# pred.	# components q									
	1	2	3	4	5	6	7	8	9	10
23,862	1.04	1.05	1.00	0.93	0.87	0.80	0.75	0.72	0.70	0.69
10,000	0.96	0.86	0.75	0.70	0.67	0.66	0.65	0.65	0.66	0.66
5,000	0.82	0.72	0.66	0.64	0.63	0.63	0.63	0.64	0.64	0.64
1,000	0.63	0.59	0.58	0.58	0.58	0.59	0.59	0.60	0.60	0.60
200	0.58	0.57	0.56	**0.56**	0.57	0.58	0.58	0.58	0.59	0.59
50	0.62	0.60	0.60	0.61	0.62	0.63	0.63	0.65	0.65	0.66

TABLE 6.5: *Mean squared prediction error for Supervised Principal Components for various choices of the number of components used and of the number op predictors selected. Simulated data in which the tenth latent variable is related to y and to 989 predictor variables.*

# pred.	# components q									
	1	2	3	4	5	6	7	8	9	10
23,862	1.05	1.06	1.07	1.06	1.04	1.00	0.94	0.90	0.87	0.85
10,000	1.11	1.05	0.98	0.92	0.88	0.85	0.83	0.82	0.82	0.82
5,000	1.03	0.96	0.92	0.87	0.84	0.82	0.81	0.81	0.81	0.81
1,000	0.92	0.88	0.85	0.82	0.82	0.81	0.81	0.80	0.80	**0.80**
200	0.90	0.88	0.87	0.86	0.86	0.85	0.85	0.85	0.85	0.84
50	0.96	0.94	0.94	0.94	0.94	0.94	0.96	0.96	0.97	0.97

Section 6.8. It is interesting to note, however, that even when a pre-selection of genes gives better results, it is not always at very low values of q. This can be seen in table 6.5. The differences between methods are small here, however.

A first recommendation from these simulations is that a supervised version of Principal Components Regression is always preferable to a non-supervised one, because the optimal prediction error was never attained without pre-selection of genes. It is already common practice in microarray research to do a pre-selection of genes based on the variance of genes. A pre-selection that also takes correlation with the outcome into account may increase predictive accuracy. A second recommendation is that, when using Supervised Principal Components, it is not advisable to only look at the case $q = 1$ as Bair et al. do. Often a better prediction error can be found at slightly larger values of q, so it can be worthwhile to also search for an optimum among these values. Cross-validation can be a good strategy for that.

6.10 Discussion

We have constructed a basic joint model of the predictor variables $\mathbf{x}$ and the outcome y, in which both $\mathbf{x}$ and y depend on a set of m latent variables. This model is very general because it assumes a general error structure for $\mathbf{x}$ and it does not assume that number of latent variables is known.

We have shown that assuming this model and constructing a prediction rule which has good properties in the $p \to \infty$, n fixed, situation leads to weighted principal components regression in a natural way. The ideal weighting is one that puts most weight on those predictor variables that are correlated with the outcome y. This gives good arguments for using a variant of the method of Supervised Principal Components (Bair et al., 2004), which puts all weight on a subset of the predictor variables that is correlated with the outcome.

This result may be considered surprising, because the method of Supervised Principal Components was originally motivated using a very different and much more restrictive model and its good properties were proved in this model using traditional $n \to \infty$ asymptotics. Furthermore, the model of this paper is actually very similar to the method presented by Burnham et al. (1999a,b) to motivate their method.

The essential assumption for the construction of the method in this paper are that the number of predictors p is very large, specifically much larger than the sample size n, and that the unpredictable part of y is not negligible. These assumptions are very realistic in many statistical applications in modern science, where extremely high-dimensional data are become the rule rather than the exception.

The model presented in this paper also makes a few other assumptions which are not strictly essential, but merely serve to keep the subject matter from becoming too technical. Examples of such assumptions are the normality of the errors, the assumptions that $m < n$ and that $q \leq m$. We expect that these assumptions can be dropped without leading to important difficulties. Of greater practical relevance is to investigate what happens when the surrogate Θ used for estimation of A and $\boldsymbol{\beta}$ is different from the surrogate that is used for prediction. This allows more flexibility in the prediction process, for example when there are missing data in $\mathbf{x}_{\text{new}}$. Another important extension of the model would allow y to depend on $\mathbf{f}$ through a generalized linear model. As Supervised Principal Components is also advocated for these situations, it is interesting to check whether the arguments presented in this paper still apply. At what cost all the above assumptions can be dropped may be an interesting subject for further research.

6.11 Proofs of the theorems

Theorem 1

Proof: Recall that $\mathbf{x} = A'\mathbf{f} + \mathbf{e}$, so

$$\begin{aligned} \gamma'_\Theta \mathbf{x} &= \beta'(A\Theta A')^{-1} A\Theta A'\mathbf{f} + \gamma'_\Theta \mathbf{e} \\ &= \beta'\mathbf{f} + \gamma'_\Theta \mathbf{e}, \end{aligned}$$

so the mean of $\gamma'_\Theta \mathbf{x}$ is $\beta'\mathbf{f}$. The variance is

$$\mathrm{Var}(\gamma'_\Theta \mathbf{e}) = \beta'(A\Theta A' + I)^{-1} A\Theta\Psi\Theta A'(A\Theta A' + I)^{-1}\beta.$$

This latter expression is $O(p^{-1})$ by Assumptions 1 and 2 and the properties of the surrogate Θ. □

Theorem 3

The proofs of Theorem 3 requires the following three Lemmas, which relate the matrix S to its limiting expectation

$$\Sigma = FG_\Theta F' + \tau^2 I$$

where $G_\Theta = \lim_{p\to\infty} p^{-1} A\Theta A'$ and $\tau^2 = \lim_{p\to\infty} p^{-1}\mathrm{trace}(\Theta\Psi)$. The matrix Σ exists and is finite by the properties of Θ.

Lemma 1 *As $p \to \infty$,*

$$\frac{1}{n^2}\mathrm{E}\big[\, \|S - \Sigma\|^2 \mid F\big] = O(p^{-1}).$$

The statement holds almost surely in F uniformly in n.

This lemma is modification of Lemma 1 in Van Houwelingen and Schipper (1981).

Proof: We first prove this lemma under slightly different assumptions. We shall assume that $\Theta = I$ and Ψ is diagonal, but not necessarily invertible. Let $\psi_i^2 \geq 0$ be the i-th diagonal element of Ψ, $\mathbf{x}_i$ the i-th column of X (an n-vector) and $\boldsymbol{\alpha}_i$ the i-th column of A (an m-vector).

Call $R = S - \mathrm{E}(S \mid F)$. Then we have

$$R = \frac{1}{p}\sum_{i=1}^{p} \big(\mathbf{x}_i\mathbf{x}_i' - F\boldsymbol{\alpha}_i\boldsymbol{\alpha}_i'F' - \sigma_i^2 I\big).$$

Given F, R is an average of independent zero mean random variables, so by the law of large numbers and the fact that $\mathrm{E}(S \mid F) = \Sigma + O(p^{-1/2})$, by Assumption 2 on page 95, the statement of the lemma follows for fixed n.

To prove uniformity in n we will show that

$$\frac{1}{n^2}\mathrm{E}\big[\|R\|^2 \mid F\big] = \frac{1}{n^2 p^2}\sum_{i=1}^{p}\mathrm{trace}\{\mathrm{E}[R_i^2 \mid F]\}$$

is bounded in n, where $R_i = \mathbf{x}_i\mathbf{x}_i' - F\boldsymbol{\alpha}_i\boldsymbol{\alpha}_i'F' - \sigma_i^2 I$. For each i we can write $\mathbf{x}_i = F\boldsymbol{\alpha} + \mathbf{e}_i$ with $\mathbf{e}_i$ an n-vector of independent normal errors with fourth moment $\kappa_i < \infty$. Then

$$\begin{aligned}\frac{1}{n^2}\mathrm{trace}\{\mathrm{E}[R_i^2 \mid F]\} &= \frac{1}{n^2}\mathrm{trace}\{(F\boldsymbol{\alpha}_i\mathbf{e}_i + \mathbf{e}_i\boldsymbol{\alpha}_i'F' + \mathbf{e}_i\mathbf{e}_i' - \sigma_i^2 I)^2\}\\ &= \frac{4}{n}\sigma_i^2\boldsymbol{\alpha}_i'F'F\boldsymbol{\alpha}_i + 2\sigma_i^4.\end{aligned}$$

For every i this expression converges a.s. to a limit as $n \to \infty$, because by the strong law of large numbers $n^{-1}F'F \to I$ (a.s.). This proves the lemma for $\Theta = I$ and Ψ diagonal.

For general Θ and Ψ, we write $\Psi^{1/2}\Theta^{1/2}$ in a singular value decomposition as $\Psi^{1/2}\Theta^{1/2} = QD^{1/2}T'$, where Q and T are $p \times p$ orthogonal matrices and D is diagonal. Transform the data matrix X to $\tilde{X} = X\Theta^{1/2}T$. Then $\tilde{X}$ and $\mathbf{y}$ conform to the general model (6.2) with parameters $\tilde{A} = A\Theta^{1/2}T$ and $\tilde{\Psi} = D$. For this model we can take $\tilde{\Theta} = I$ and apply this lemma on $\tilde{S} = p^{-1}\tilde{X}\tilde{\Theta}\tilde{X}'$ and its corresponding $\tilde{\Sigma}$. This immediately proves the statement of the lemma for X, as $\tilde{S} = S$ and $\tilde{\Sigma} = \Sigma$. □

For the next lemma, define $\hat{P} = \hat{U}\hat{U}'$ and, analogously, define P as the projection matrix for projection on the space spanned by the eigenvectors of the q largest eigenvalues of Σ. Let U be an $n \times q$ semi-orthogonal matrix such that $P == UU'$.

Lemma 2 *If $q \le \mathrm{rank}(G)$, the matrix P exists a.s. and is given by*

$$P = F(F'F)^{-1/2}VV'(F'F)^{-1/2}F',$$

where V is the $m \times q$ semi-orthogonal matrix of the eigenvectors of q largest eigenvalues of $(F'F)^{1/2}G(F'F)^{1/2}$.

Proof: Note that $T = F(F'F)^{-1/2}$ is a.s. an $n \times m$ semi-orthogonal matrix and that Σ a.s. has distinct q-th and $q+1$-th eigenvalues. Decompose

$$\Sigma = T(F'F)^{1/2}G(F'F)^{1/2}T' + \tau^2 I.$$

Diagonalize $(F'F)^{1/2}G(F'F)^{1/2} = VDV'$. Then the diagonalization of Σ is

$$\Sigma = TVDV'T' + \tau^2 I.$$

The eigenvectors of the largest eigenvalues of Σ are the same as those of $TVDV'T'$, and these are therefore given by $U = TV$. Hence $P = TVV'T'$. □

Lemma 3 *If $q \leq \text{rank}(G)$, as $p \to \infty$,*

$$\mathrm{E}\big[\, \|\hat{P} - P\|^2 \mid F \big] = O(p^{-1}).$$

almost surely. Furthermore, if $q = m$ or if the q-th and $q+1$-th eigenvalues of G are distinct, the statement holds uniformly in n.

This lemma is a generalization of Lemma 2 in Van Houwelingen (1984).

Proof: By definition $\hat{P}$ maximizes $\text{trace}(S\hat{P})$ among all projection matrices of rank q. Call $R = S - \Sigma$. We have

$$\begin{aligned} \text{trace}\{\Sigma(P - \hat{P})\} &= \text{trace}\{S(P - \hat{P})\} + \text{trace}\{R(\hat{P} - P)\} \\ &\leq \text{trace}\{R(\hat{P} - P)\} \\ &\leq \|R\| \cdot \|\hat{P} - P\|, \end{aligned} \tag{6.12}$$

the last statement being an application of the Schwartz inequality. Let λ_q and λ_{q+1} be the q-th and $q+1$-th largest eigenvalues of Σ. Then

$$\begin{aligned} \text{trace}\{\Sigma(P - \hat{P})\} &= \text{trace}\{P\Sigma P(P - \hat{P})\} \\ &\quad + \text{trace}\{(I - P)\Sigma(I - P)(P - \hat{P})\} \\ &= \text{trace}\{(I - \hat{P})P\Sigma P(I - \hat{P})\} \\ &\quad - \text{trace}\{\hat{P}(I - P)\Sigma(I - P)\hat{P}\} \\ &\geq \lambda_q \text{trace}(P - P\hat{P}) - \lambda_{q+1}\text{trace}(\hat{P} - P\hat{P}) \\ &= \frac{1}{2}(\lambda_q - \lambda_{q+1})\|P - \hat{P}\|^2. \end{aligned} \tag{6.13}$$

The final equation uses $\text{trace}\{(P - \hat{P})^2\} = \text{trace}(P) - 2\text{trace}(P\hat{P}) + \text{trace}(\hat{P})$ and $\text{trace}(P) = \text{trace}(\hat{P})$. Combining (6.12) and (6.13) yields

$$\|P - \hat{P}\| \leq \frac{2}{\lambda_q - \lambda_{q+1}}\|R\|.$$

By the randomness of F and the assumption that $q \leq \text{rank}(G)$, the first $q+1$ eigenvalues of Σ are almost surely distinct, so $\lambda_q - \lambda_{q+1} > 0$ and the first

statement of this lemma for fixed n follows directly from Lemma 1 by squaring and taking expectations.

To prove the uniformity we again remark that $n^{-1}F'F \to I$ (a.s.) as $n \to \infty$. Hence $n^{-1}(F'F)^{1/2}G(F'F)^{1/2} \to G$ and consequently $n^{-1}(\lambda_q - \lambda_{q+1})$ tends to a non-zero limit almost surely, by the assumption on the eigenvalues of G. Therefore the upper bound

$$\frac{2n}{\lambda_q - \lambda_{q+1}} \|n^{-1}R\|.$$

remains bounded in n almost surely. □

Lemma 4 *If $q \leq \text{rank}(G)$, there is a $q \times q$ rotation matrix W, depending on U and $\hat{U}$, such that as $p \to \infty$,*

$$\text{E}\big[\, \|\hat{U} - UW\|^2 \mid F\big] = O(p^{-1})$$

almost surely. Furthermore, if $q = m$ or if the q-th and $q+1$-th eigenvalue of G are distinct, the statement holds uniformly in n.

Proof: First remark that $P\hat{P}$ almost surely has rank q, so that $U'\hat{U}$ also has rank q and is invertible. Choose $W = U'\hat{U}(\hat{U}'UU'\hat{U})^{-1/2}$, which is a $q \times q$ orthogonal matrix. Then

$$\begin{aligned} \|\hat{U} - UW\|^2 &= \text{trace}(\hat{U}\hat{U}') - 2\text{trace}(UW\hat{U}') + \text{trace}(UU') \\ &= \text{trace}(P) - 2\text{trace}\{(\hat{U}'UU'\hat{U})^{1/2}\} + \text{trace}(\hat{P}) \end{aligned}$$

The eigenvalues of $(\hat{U}'UU'\hat{U})^{1/2}$ are the singular values of the matrix $P\hat{P}$ and therefore the square roots of the eigenvalues of the matrix $P\hat{P}P$. The eigenvalues of the latter matrix are between zero and one, so the square roots of these eigenvalues are larger than the eigenvalues themselves. Hence

$$\text{trace}\{(\hat{U}'UU'\hat{U})^{1/2}\} \geq \text{trace}(\hat{U}'UU'\hat{U}) = \text{trace}(P\hat{P}),$$

so that

$$\begin{aligned} \|\hat{U} - UW\|^2 &\leq \text{trace}(P) - 2\text{trace}(P\hat{P}) + \text{trace}(\hat{P}) \\ &= \|P - \hat{P}\|^2 \end{aligned}$$

The statements of the lemma now follow immediately from their counterparts in Lemma 3. □

Proof of Theorem 3: First we apply Lemma 4 both to the matrix $\tilde{U}$. There is an $m \times m$ rotation matrix $\tilde{W}$ such that $\mathrm{E}\big[\, \|\tilde{U} - U_m\tilde{W}\|^2 \mid F\big] = O(p^{-1})$, where U_m is (a rotation of) the $n \times m$ semi-orthogonal matrix of the eigenvectors of the m largest eigenvalues of $\tilde{\Sigma} = F\tilde{G}F' + \tilde{\tau}^2 I$. As $\tilde{G}$ is full rank, we can take U_m as $U_m = F(F'F)^{-1/2}$.

Next we apply Lemma 4 to the matrix $\hat{U}$. There is a $q \times q$ matrix W such that $\mathrm{E}\big[\, \|\hat{U} - U_q W\|^2 \mid F\big] = O(p^{-1})$, where U_q is (a rotation of) the $n \times q$ semi-orthogonal matrix of the eigenvalues of the q largest eigenvalues of $\tilde{\Sigma} = FGF' + \tau^2 I$. As the matrix of the eigenvectors of the m largest eigenvalues is a rotation of U_m, the matrix of the eigenvectors of the q largest eigenvalues is a $U_m V_0$ for some $m \times q$ semi-orthogonal matrix V_0.

Remark that $\tilde{\Lambda} = \tilde{U}'\tilde{S}\tilde{U} < \infty$ as $p \to \infty$, so it is easily checked that $\tilde{\Upsilon} = I + O_P(p^{-1})$ uniformly in n.

Define $V = \tilde{W}V_0W$. Then

$$\begin{aligned}\frac{1}{p}\mathrm{E}\big[\, \|\hat{A} - V'\tilde{A}\|^2 \mid F\big] &= \frac{1}{np}\mathrm{E}\big[\, \|(\hat{U}' - V'\tilde{\Upsilon}^{1/2}\tilde{U}')X\|^2 \mid F\big] \\ &\leq \frac{1}{np}\mathrm{E}\big[\, \|X\|^2 \mid F\big] \cdot \mathrm{E}\big[\, \|\hat{U}' - V'\tilde{\Upsilon}^{1/2}\tilde{U}'\|^2 \mid F\big]\end{aligned}$$

by the Schwartz inequality. The first factor on the right hand side of the inequality converges to a finite limit as $p \to \infty$, a.s. uniformly in n. The second factor can be bounded further by

$$\begin{aligned}\mathrm{E}\big[\, \|\hat{U}' - V'\tilde{\Upsilon}^{1/2}\tilde{U}'\|^2 \mid F\big] \leq\ & \mathrm{E}\big[\, \|\hat{U}' - WV_0'U_m'\|^2 \mid F\big] \\ & + \mathrm{E}\big[\, \|V'\tilde{\Upsilon}^{1/2}\tilde{U}' - WV_0'U_m'\|^2 \mid F\big]\end{aligned}$$

Both terms are $O(p^{-1})$ (a.s.) uniformly in n by Lemma 4, which proves the statement about $\hat{A}$ and $\tilde{A}$. The proof of the statement about $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ is completely analogous. □

Theorem 4

We first formulate and prove a lemma on $\hat{\tau}^2$

Lemma 5 *If $r > m$, as $p \to \infty$,*

$$\mathrm{E}\big[(\hat{\tau}^2 - \tau^2)^2 \mid F\big] = O(p^{-1})$$

almost surely and uniformly in n.

Proof: We have

$$\begin{aligned}\hat{\tau}^2 - \tau^2 &= \tfrac{1}{n-r}\text{trace}(S\hat{Q}) - \tfrac{1}{n-r}\text{trace}(\tau^2\hat{Q}) \\ &= \tfrac{1}{n-r}\text{trace}\{\hat{Q}(S-\Sigma)\} + \tfrac{1}{n-r}\text{trace}\{\hat{Q}(\Sigma-\tau^2 I)\}.\end{aligned}$$

Construct $\hat{P}$ and P from S and Σ for $q = \text{rank}(G)$. Then $P(\Sigma - \tau^2 I) = \Sigma - \tau^2 I$ and $\text{trace}(\hat{Q}T) \leq \text{trace}\{(I-\hat{P})T\}$ for any positive semi-definite T. We have

$$\begin{aligned}|\hat{\tau}^2 - \tau^2| &\leq \tfrac{1}{n-r}|\text{trace}(S-\Sigma)| + \tfrac{1}{n-r}\text{trace}\{(I-\hat{P})P(\Sigma-\tau^2 I)\} \\ &\leq \tfrac{1}{n-r}\|S-\Sigma\| + \tfrac{1}{n-r}\|(I-\hat{P})P\| \cdot \|\Sigma - \tau^2 I\|.\end{aligned}$$

The result follows when we remark that $\|P - P\hat{P}\|^2 = \frac{1}{2}\|P - \hat{P}\|^2$ and apply Lemmas 1 and 3. □

Proof of Theorem 4: Write $\mathbf{x}_{\text{new}} = A\mathbf{f}_{\text{new}} + \mathbf{e}_{\text{new}}$ and $X = FA + \mathcal{E}$. We have

$$\begin{aligned}\hat{\gamma}'\mathbf{x}_{\text{new}} &= \mathbf{y}'\hat{U}(\hat{U}'X\Theta X'\hat{U} + p\hat{\tau}^2 I)^{-1}\hat{U}'X\Theta A\mathbf{f}_{\text{new}} + \hat{\gamma}\mathbf{e}_{\text{new}} \\ &= \frac{1}{p}\mathbf{y}'\hat{P}(S-\hat{\tau}^2 I)^{-1}\hat{P}FA\Theta A\mathbf{f}_{\text{new}} \qquad (6.14) \\ &\quad + \frac{1}{p}\mathbf{y}'\hat{P}(S-\hat{\tau}^2 I)^{-1}\hat{P}\mathcal{E}\Theta A\mathbf{f}_{\text{new}} \qquad (6.15) \\ &\quad + \frac{1}{p}\mathbf{y}'\hat{P}(S-\hat{\tau}^2 I)^{-1}\hat{P}X\Theta\mathbf{e}_{\text{new}}, \qquad (6.16)\end{aligned}$$

which is a sum of three complicated-looking terms, which we shorthand t_1 (6.14), t_2 (6.15) and t_3 (6.16) in the order they appear above. We study the behaviour of the three terms when $p \to \infty$. The calculations are tedious but straightforward. They mainly involve repeated application of Lemmas 1 and 3.

We have

$$\begin{aligned}\text{E}[(t_1 - \tilde{y}_{\text{new}})^2 \mid F] &= O(p^{-1}) \\ \text{E}[t_2^2 \mid F] &= O(p^{-1}) \\ \text{E}[t_3^2 \mid F] &= O(p^{-1}),\end{aligned}$$

all uniformly in n under the conditions given. This proves the theorem. □

CHAPTER 7

Enhancing Scatterplots with Smoothed Densities

Abstract

Scatterplots of microarray data generally contain a very large number of dots, making it difficult to get a good impression of their distribution in dense areas. We present a fast an simple algorithm for two-dimensional histogram smoothing to visually enhance scatterplots. Functions for Matlab and R are available from the authors.

7.1 Introduction

The scatterplot is a simple but effective tool in microarray analysis. It is one of the best ways to visualize expressions of two arrays (or of two dye colours on one array). Still the scatterplot leaves much to be desired. Because of the large number of dots, up to ten thousand or more, large parts of the picture can become completely black. Then it is hard to get a good impression of the distribution of the spots. Figure 7.1 shows an example. When the plotting symbols are large, as in the left panel, the center of the graph gets completely filled with ink. As the right panel shows, it helps to use very small symbols, but then isolated dots can easily be missed.

A solution is to move from plotting of the individual dots to a presentation of their empirical distribution. An obvious choice is the two-dimensional histogram. Unfortunately, either one has to use rather wide bins, or to accept a rather choppy histogram. Figure 7.2 shows examples.

This is a pre-copy-editing, author-produced version of an article accepted for publication in *Bioinformatics* following peer review. The definitive publisher-authenticated version: P. H. C. Eilers and J. J. Goeman (2004). Enhancing scatterplots with smoothed densities. *Bioinformatics 20*(5), 623–628 is available online at: `http://dx.doi.org/10.1093/bioinformatics/btg454`.

We can achieve large improvements if we use a histogram with narrow bins and additional smoothing, as is shown in Figure 7.3, which is based on histograms with 200 bins in both directions. In this paper we present an algorithm for fast and effective smoothing of two-dimensional histograms. Speed is important, because in everyday work many scatterplots are made on a computer screen to help in exploratory data analysis.

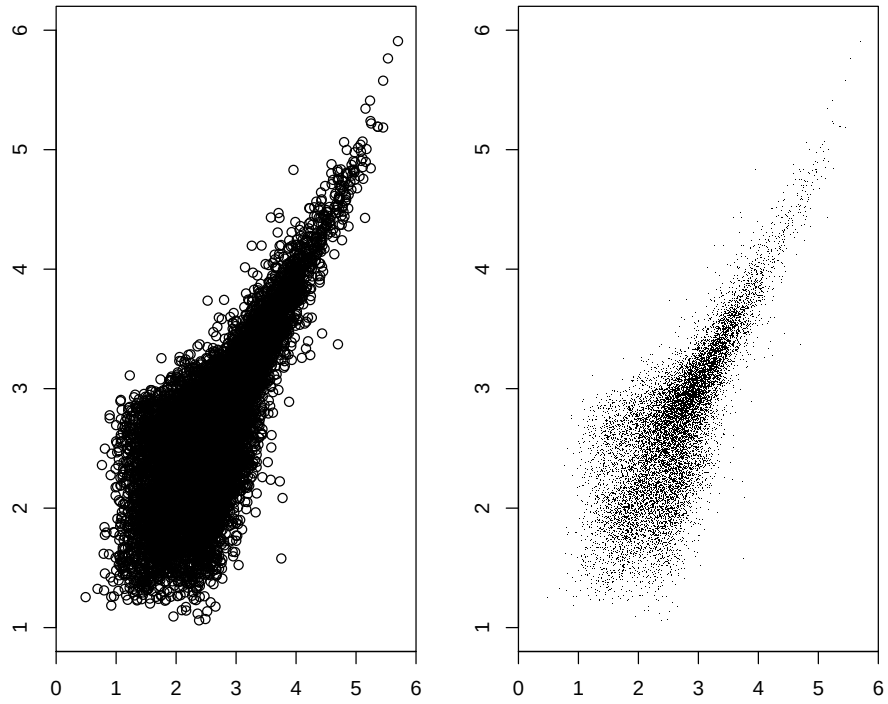

FIGURE 7.1: *Two scatterplots of log-expressions of a pair of microarrays. Left: large symbols, right: small symbols.*

7.2 Algorithm

The two-dimensional histogram is a natural generalization of the well-known histogram. The x-y domain is cut into rectangles and the number of observations in each rectangle is counted. As Figure 7.2 shows, a graphical display of this raw histogram is not a great succes. We can make it more informative (and

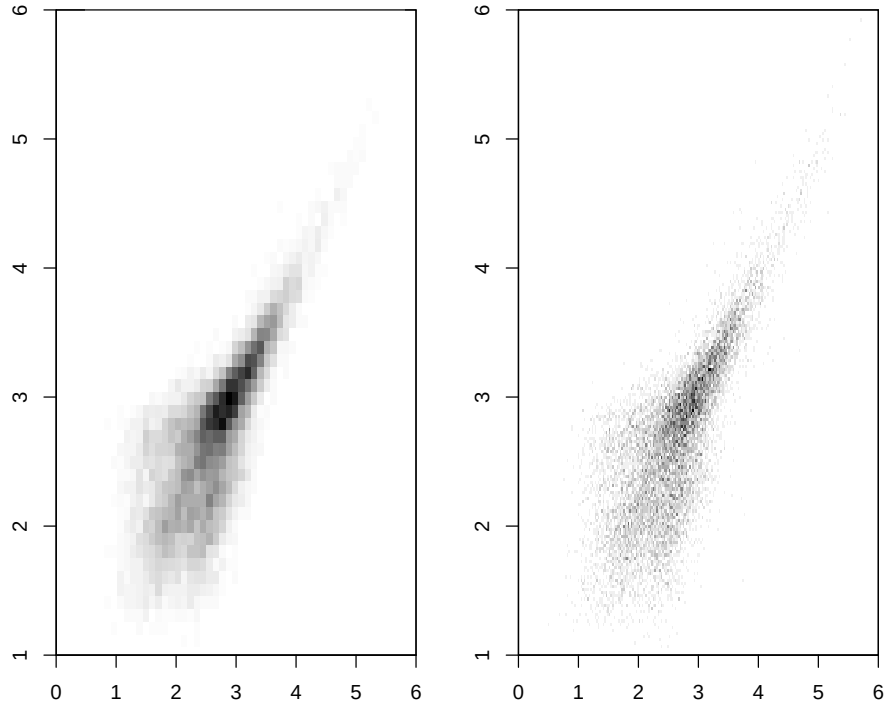

FIGURE 7.2: *Two-dimensional histograms, derived from the scatterplots in Figure 1. Left: 50 by 50 bins, right: 200 by 200 bins.*

attractive) with a simple smoothing algorithm.

Let H be the matrix of counts resulting from a two-dimensional histogram. Consider smoothing of one column of H, we will call it the vector y, to get a vector z. The distance from z to y can be measured as the sum of squares of the residuals $y - z$:

$$S = \sum_i (y_i - z_i)^2 = |y - z|^2.$$

The roughness of z can be measured by first computing differences,

$$\Delta z_i = z_i - z_{i-1},$$

and then summing their squares:

$$R = \sum_i (\Delta z_i)^2 = |D_1 z|^2.$$

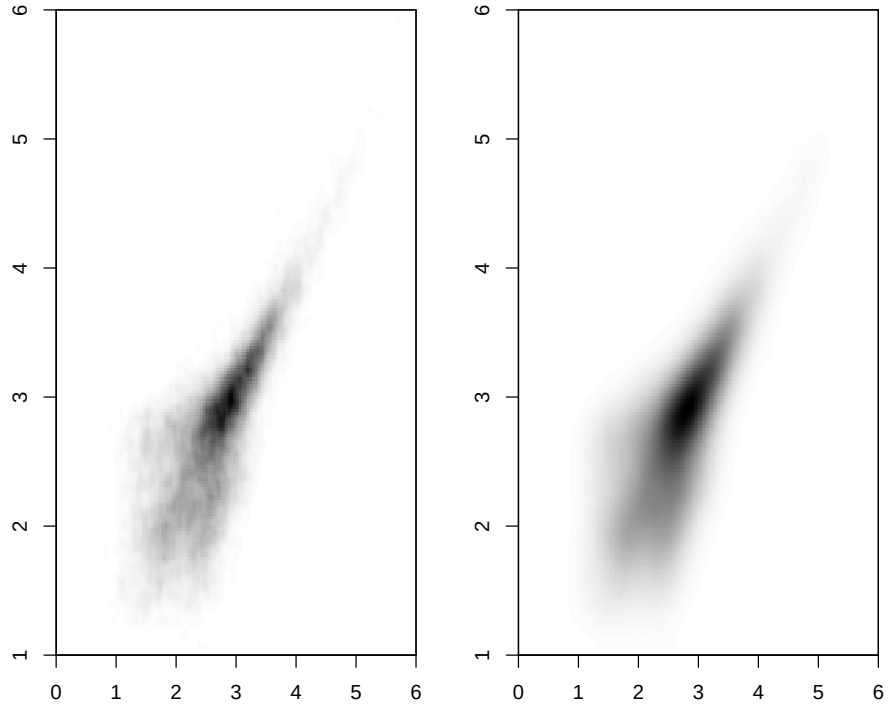

FIGURE 7.3: *Smoothed two-dimensional histograms, derived from the scatterplots in Figure 1, using 200 by 200 bins. Left: $\lambda = 1$, right: $\lambda = 10$.*

Here D_1 is a first-order difference matrix such that $D_1 z = \Delta z$:

$$D_1 = \begin{pmatrix} -1 & 1 & & \emptyset \\ & \ddots & \ddots & \\ \emptyset & & -1 & 1 \end{pmatrix}.$$

We combine S and R in one penalized least squares function Q:

$$Q = S + \lambda R = |y - z|^2 + \lambda |D_1 z|^2, \tag{7.1}$$

and compute the vector $\hat{z}$ that minimizes Q. By changing λ we can balance our preference between fit to the data y (the first term) and roughness of z (the second term). The higher λ, the more the roughness of z will be penalized, leading to a smoother result, at the cost of the fit to y getting worse. The minimizer of

Q is the solution to the following linear system of equations:

$$(I + \lambda D_1' D_1)\hat{z} = y, \tag{7.2}$$

where I is the identity matrix. For moderate lengths of y, say 200 or less, it can be solved very quickly on modern computers.

A refinement uses second order differences:

$$\Delta^2 z_i = \Delta(\Delta z_i) = (z_i - z_{i-1}) - (z_{i-1} - z_{i-2}) = z_i - 2z_{i-1} + z_{i-2}.$$

The only change to the system (7.2) is that D_1 is changed to a second order difference matrix

$$D_2 = \begin{pmatrix} 1 & -2 & 1 & & \varnothing \\ & \ddots & \ddots & \ddots & \\ \varnothing & & 1 & -2 & 1 \end{pmatrix}.$$

and λ changed to λ^2.

Figure 7.4 shows one column of H and the effect of smoothing with different values of λ, using first or second order differences. The latter choice gives a somewhat smoother result and follows the peaks better. However, there is a slight problem: we can get negative values if we use second order differences, especially with strong smoothing. The explanation is shown in Figure 7.5, where the impulse response of the smoothers is displayed. Imagine a degenerate histogram with zeroes in all cells but one, which contains a one. Smoothing this impulse shows what happens to one count. Any histogram can be interpreted as a sum of many of these impulses, with different positions of the single count. Because the smoother is linear, the smoothed histogram is the sum of the corresponding smoothed impulses. With first-order differences the impulse response has the shape of decaying exponentials in both directions and it cannot become negative. With second-order differences, each branch of the impulse response consists of two exponentials, in a combination that leads to a negative minimum.

For visual display, negative values of the smoothed histogram are not really a problem. But it is inelegant and it can be harmful when results are used for further computations that expect non-negative probabilities. A solution is to use both a first and second-order penalty (Eilers, 1994). We use the penalty $\lambda^2|D_2 z|^2 + \alpha\lambda|D_1 z|^2$ and search for (a "pleasant" number) α that keeps the impulse response from becoming non-positive; $\alpha = 2$ is a round number that works well. The penalized least squares function becomes.

$$Q = |y - z|^2 + \lambda^2|D_2 z|^2 + 2\lambda|D_1 z|^2. \tag{7.3}$$

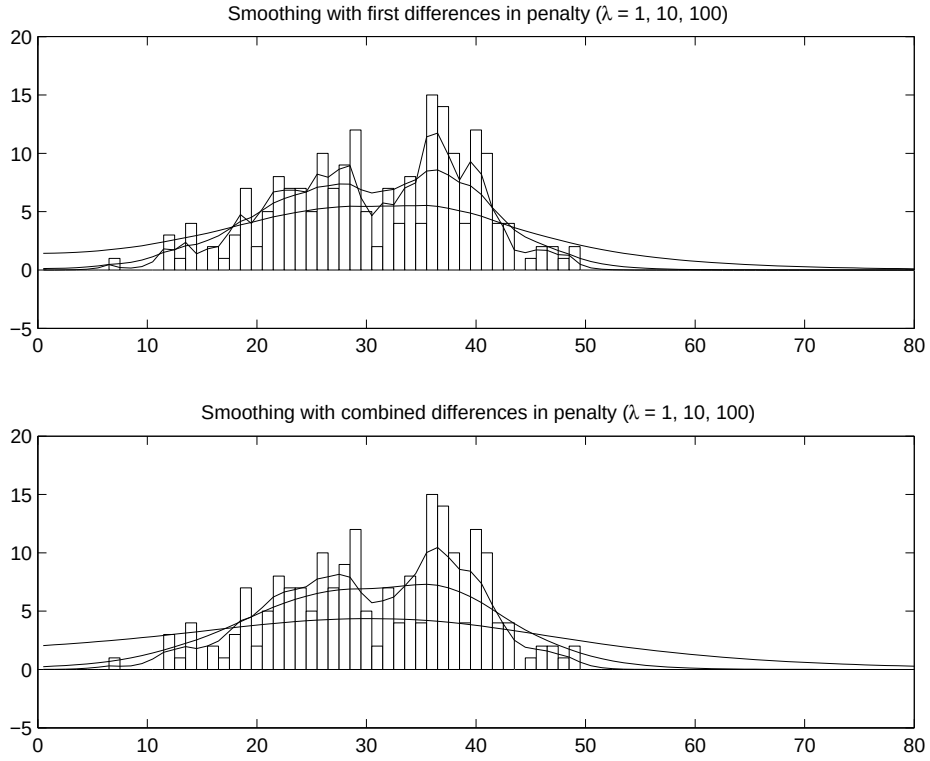

FIGURE 7.4: *Smoothing of a one-dimensional histogram. Top: using first-order differences, bottom: using second-order differences.*

The impulse response of this smoother is also shown in Figure 7.5. The peak is rounded like that of the second-order smoother and the tails are like that of the first-order smoother.

The idea of using differences in a penalty goes back at least to Whittaker (1923). Extensions and fast algorithms for one-dimensional smoothing have been presented elsewhere (Eilers, 1994, 2003). An attractive property of this smoother is that it respects boundaries. This is unlike a kernel density smoother, which computes a weighted local mean and implicitly assumes zero counts past the boundaries of a histogram. This can do little harm on densities with tails that gradually slope down. But it is when a density has its peak at, or near, zero. The peak will be rounded too much by a kernel smoother. This type of density frequently occurs with when one studies squares of absolute values of data, to get an impression of variance or standard deviation.

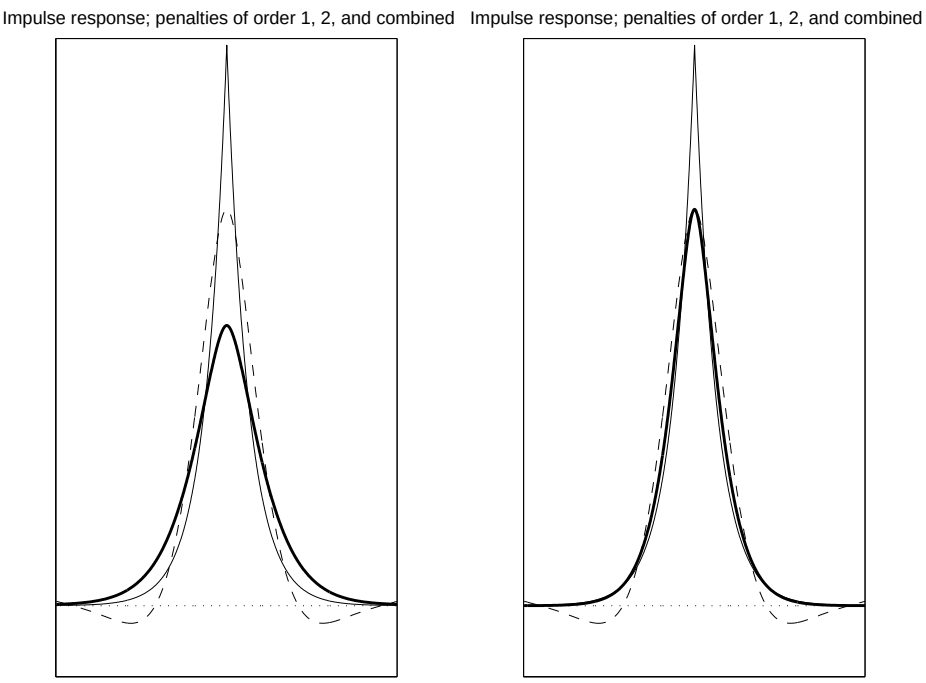

FIGURE 7.5: *Impulse response of the smoother (arbitrary scales). Peaked curve: first-order differences; rounded broken-line curve with negative lobe: second-order differences; thick-line curve: combined differences. Left panel: equal values of λ; right panel: $\lambda/2$ for combined differences.*

Once we have a good smoother for vectors, it is trivial to apply it to all columns of a matrix. The standard algorithms in Matlab, R or S-plus for the solution of linear equations accept a matrix as the right-hand side and return the solution as a matrix. Thus we can write the smoothing of a matrix Y, to get Z as a simple modification of (7.2):

$$(I + 2\lambda D_1' D_1 + \lambda^2 D_2' D_2)\hat{Z} = Y. \tag{7.4}$$

This is the basis for fast smoothing of a two-dimensional histogram H: first smooth the columns of H to get, say, G and then smooth the columns of G' (which are the rows of G) to get F', the transpose of the desired result. It is easily checked that the result is invariant to the order of the smoothing operations: smoothing the rows of H before the columns leads to the same result.

Only a few lines of Matlab are needed to apply the smoother to a histogram given in a matrix `H`: `F = expsm(expsm(H, lambda)', lambda)';`, where `expsm`

is a function defined as

```
function Z = expsm(Y, lambda)
m = size(Y, 1);
E = eye(m);
D1 = diff(E);
D2 = diff(D1);
P = lambda ^ 2 * D2' * D2 + 2 * lambda *  D1' * D1;
Z = (E + P) \ Y;
```

An implementation in S-plus or R would look very similar. Note that Matlab can speed up the computations (about 3 times) by exploiting the sparseness of the system of equations in a very simple way, using the sparse identity matrix `speye` instead of the full `eye`.

7.3 Implementation

Figure 7.6 shows the application to four different displays of one pair of arrays with 12625 expressions. The NW panel shows a scatterplot derived from 12625 log-expressions (base 10). The NE panel shows mean and difference. The SW panel displays mean and absolute value of the difference. This is an example of a skew density with a peak near the origin (for each column of the histogram). The logarithms of the absolute differences (plus a shift of 0.01 to accommodate zeroes) are shown in the SE panel. In each graph a random selection of 1000 data points is also plotted.

The choice of λ partly is a matter of taste. The user should play with it to get a visual appearance to his/her taste. It also depends on the number of bins and the number of data pairs. Our personal experience is that λs in the range from 1 to 100 work well with 100 or 200 bins per dimension and approximately 10^4 data pairs.

The colour scale is also a matter of taste. Using white for zero values and a dark colour for the maximum seems attractive (and saves expensive ink or toner). We advice to take the colour for the maximum not too dark, so that black symbols for the data points will be clearly visible.

7.4 Discussion

We have presented an algorithm for visual enhancement of a scatterplot, using a smoothed histogram. The algorithm is fast: computing and plotting Figure 7.6 takes less than a second, using Matlab 6.5 on a 1000 MHz Pentium III PC. So it can be used in a routine way when exploring scatterplots and one can nearly instantaneously see the effects of changing the amount of smoothing. Even

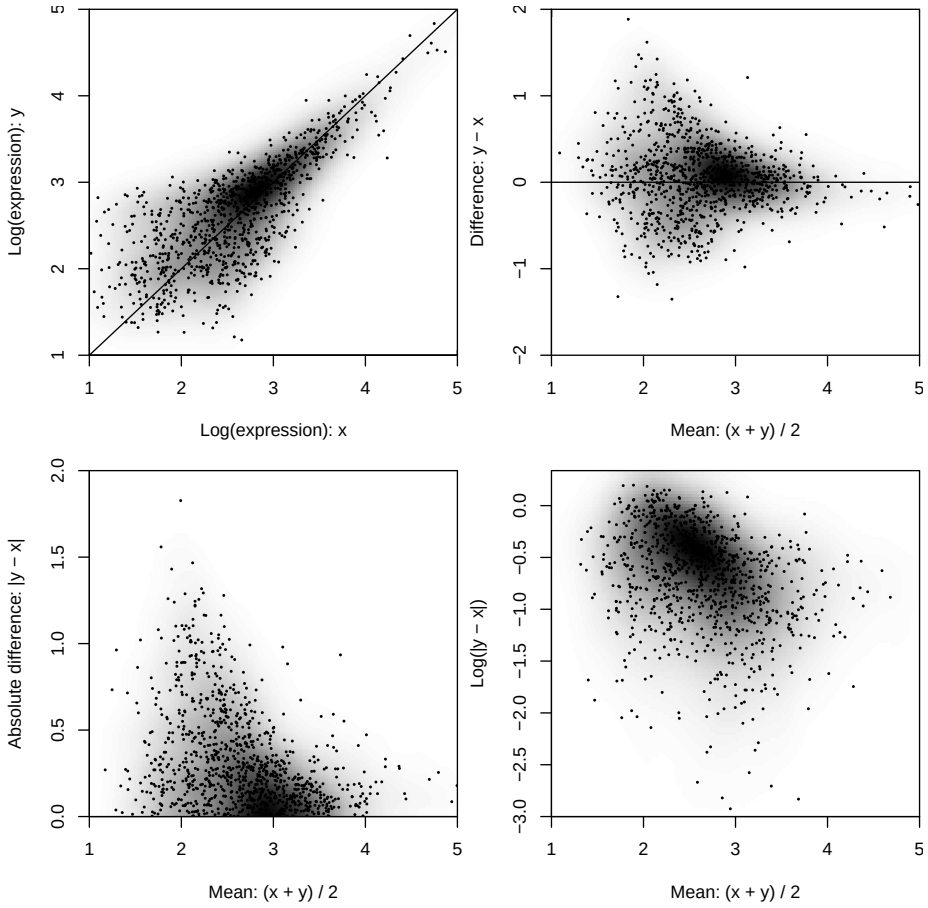

FIGURE 7.6: *Four different displays of a pair of microarrays, using histogram smoothing and plotting of 1000 (of 12625) data points.*

one million data points are handled in less than 10 seconds. R is several times slower; the bottleneck there is the computation of the histogram.

We investigated the performance of the R function `kde2d()` for kernel estimation of two-dimensional densities. With 2000 data points or less it has the same performance as the our algorithm. With 10^4 data points it is over five times slower and above 3×10^4 data points too much memory is needed (on a 256 Mb PC). Either swapping to the hard disk slows done the process, or the computations stop with an error message.

In our experience it is useful to plot part of the dots, to give a good im-

pression of the spread of the raw data. Their number should not be too small, to be representative enough, but also not too large, to not fill the graph with too much ink. A subject for further research is to use the estimated density to determine the probability of plotting a point.

There exist good algorithms for density smoothing, like kernels and local likelihood and they will produce results that look much like ours. Our method is also not very original: penalized likelihood has been used before. But the algorithm presented here has a number of specific advantages:

1. It does not use any special smoothing libraries, but only a few lines of straightforward linear algebra computations, which are easily implemented in high-level languages like Matlab, S-plus or R.

2. It works directly on the two-dimensional histogram matrix, avoiding translation to triples (row, column, count) that other algorithms demand.

3. It respects domain boundaries, which is important when smoothing densities of very skewly distributed data, like variance estimates.

4. It is fast.

5. It can handle extremely large (10^6 or more) numbers of data points.

The four displays of the "scatterquad" in Figure 7.6 help to better understand systematic and random differences between two microarrays. Further refinements seem possible. One could use a smoothing algorithm to estimate and display trends in the upper panels and use trend-corrected differences for the displays of spread in the lower panels and possibly add trends to these plots as well. We will not pursue this issue here further, as it would carry us too far away from the main theme of the paper.

Our approach to visualization of scatterplots is in essence a simplification of density smoothing in two dimensions. Because visual display is the only goal, a refined algorithm is unnecessary. Simonoff (1996) discusses kernel estimation, while Loader (1999) presents algorithms and software for local likelihood.

Chapter 8

Conclusion

The existence of a *curse of dimensionality* is manifest in the analysis of microarray data. It shows itself when researchers are trying to find genes which are correlated with a certain phenotype: the sheer quantity of seemingly correlated genes makes it difficult to find the truly correlated ones. It appears even more strongly in prediction problems: the enormous variety of possible prediction rules completely obscures the underlying biology. In this confusing situation, biologists look to statisticians for guidance, while statisticians look to the biologists. In reality, both parties carry half of the solution, which lies in the incorporation of biological knowledge into the statistical methodology.

Statistical analysis of microarray data started out with explorative methods, which approach the data impartially and try to let the data 'speak for themselves'. Most methods of microarray data now in use are still highly exploratory in nature. This is most notable in unsupervised methods like cluster analysis, but also in prediction methods and methods for finding differentially expressed genes; only rarely do they make any use of biological knowledge. Methods for the analysis of microarray data are mainly directed at generating interesting new hypotheses, which are to be confirmed or disproved at a later stage. Only few of the many hypotheses generated in this way turn out to be meaningful, however, and the task of sifting these out is left to the biologists.

Much can be gained, therefore, by switching to a more knowledgeable way of looking at the microarray data, incorporating biological knowledge into the analysis instead of reserving its use for the interpretation stage only. As learning about genes accumulates, blindly searching for new hypotheses without making use of the knowledge already gained will prove increasingly unsatisfactory. Furthermore, hypotheses about biological mechanisms that arise from exploratory data analysis have to be tested somehow. This requires non-explorative statistical methodology to be developed for microarray data analysis.

This thesis has explored ways of making use of biological information to improve the analysis of microarray data. In the GlobalTest methodology it has provided methods for non-explorative hypothesis-driven research, allowing re-

searchers to test hypotheses about the involvement of biological processes in a certain phenotype. The same methodology can be used as a more informed type of exploratory data analysis, by incorporating the extensive knowledge about pathways into the data analysis. Similarly, in the factor analysis model for prediction in Chapter 6 it was shown how basic knowledge about the nature of microarray data can be used as guidance for the choice of a dimension reduction method.

The use of biological knowledge to improve statistical methods for analyzing microarray data is a promising new development, whose potential has not yet been exhausted. Intelligent use of this information can lead both to more powerful statistical methodology and to more interpretable results. Much work is still to be done. The pathway information which has been exploited for use in testing procedures in this thesis also has good potential for use in prediction methods. A similar challenge is to combine analysis of microarray data analysis with information from linkage studies or proteomics data. It is obvious that close cooperation with biologists is essential for the success of this line of research.

APPENDIX A

Manual of the GlobalTest package

A.1 Introduction

This document shows the functionality of the R-package *globaltest*, whose main function tests whether a given group of genes is significantly associated with a clinical variable. The demonstration in this appendix focuses on practical use of the test. To understand the idea and the mathematics behind the test, and for more details on how to interpret a test result, we refer to the papers (Goeman et al., 2005, 2004).

In recent years there has been a shift in focus from studying the effects of single genes to studying effects of multiple functionally related genes or pathways (Al-Shahrour et al., 2004; Beissbarth and Speed, 2004; Boyle et al., 2004; Mootha et al., 2003; Smid and Dorssers, 2004; Zeeberg et al., 2003; Zhang et al., 2004). Most of the current methods for studying pathways involve looking at increased proportions of differentially expressed genes in pathways of interest. These methods do not identify pathways where many genes have altered their expression in a small way. The package *globaltest* was designed to address this issue.

The *globaltest* package tests whether a group of genes is associated with a clinical variable. A group of genes can be any pre-defined set, for example based in function (KEGG, GO) or location (chromosome, cytogenetic band). The clinical variable may be a phenotypic variable or an experimental condition. It may take the form of a 0/1 group indicator, of a continuous measurement or of a survival time.

The null hypothesis to be tested is that the expression pattern of the genes in the group is not related to the clinical variable. A significant test result has three parallel interpretations.

This chapter is the manual of the R package *globaltest*, that has been published on BioConductor as: J. J. Goeman and J. Oosting (2005). Globaltest: testing association of a group of genes with a clinical variable. R package, version 3.2.0. www.bioconductor.org.

- If a pathway is significantly associated with the clinical variable, the genes in the pathway are, on average, more associated with the clinical variable than would be expected if the null hypothesis were true. One can expect a sizeable proportion of genes to be associated with the clinical variable, but these associations might not be individually significant.
- If a pathway is significantly associated with the clinical variable, samples which have similar values of the clinical variable tend also to have similar expression pattern over the pathway.
- If a pathway is significantly associated with the clinical variable, there is good potential for predicting part of the variance of the clinical variable using the genes in the pathway.

In the examples below we use data sets that are available through the BioConductor web site. All the packages necessary to repeat the examples below are available from www.bioconductor.org. We use the AML/ALL data set (Golub et al., 1999) for illustration.

```
> library(globaltest)
> library(golubEsets)
> library(hu6800)
> library(vsn)
> data(golubMerge)
> golubM <- update2MIAME(golubMerge)
> golubX <- vsn(golubM)
```

This gives us a data set `golubX`, which is of the format *exprSet*, the standard format for gene expression data in BioConductor. It has 7,129 genes for 72 samples. We used *vsn* (Huber et al., 2002) to normalize the data. Any other normalization method may be used instead. Several phenotype variables are available with `golubX`, among them "ALL.AML", the clinical variable that interests us.

In this document we use the globaltest based on BioConductor *exprSet* input. For examples using simple vector or matrix input, see `help(globaltest)`.

A.2 Global testing of a single pathway

Suppose we are interested in testing whether AML and ALL have a different gene expression pattern for certain pathways from the KEGG database.

First we load all KEGG pathways. We will use the rest in the next section.

```
> kegg <- as.list(hu6800PATH2PROBE)
> cellcycle <- kegg[["04110"]]
```

This creates a sorted list `kegg` of 140 pathways, each a vector of gene names. The vector `cellcycle` is one of them. It corresponds to the Cell Cycle pathway, "04110" in the KEGG database, which corresponds to 94 probe sets on the hu6800 chip. Suppose we are predominantly interested in this pathway. We want to know whether this group of genes is associated with the clinical outcome AML versus ALL.

It is advisable to always first test all genes to see if the overall gene expression pattern is different for different clinical outcomes. We can do this by saying

```
> gt.all <- globaltest(golubX, "ALL.AML")
```

The first input X should be the *exprSet* object, the second input Y the name of the clinical variable in `pData(X)`. Alternatively we can give a matrix of expressions as X and a vector as Y.

The test result is stored in a *gt.result* object, which also contains all the information needed to draw the plots.

```
> gt.all

Global Test result:
Data: 72 samples with 7129 genes; 1 pathway tested
Model: logistic

  genes tested Statistic Q Expected Q sd of Q    p-value
1  7129   7129      53.992         10  1.9035 5.1616e-35
```

We conclude that there is ample evidence that the overall gene expression profile for all 7,129 genes is associated with the clinical outcome: samples with similar AML/ALL status tend to have similar expression profiles. In cases such as this one, in which the overall expression pattern is associated with the clinical variable, we can expect most pathways (especially the larger ones) also to be associated with it.

Because `golubX` is an *exprSet*, we could simply give the name of the phenotype variable "AML.ALL" as our Y input. Alternatively, we can give a vector here.

The Global Test allows three different kinds of clinical variables to be tested.

- A clinical variable defining two groups, i.e. having two values (using the logistic model). For a multi-valued clinical variable, the option *levels* can be used to set which groups are to be tested against each other.
- A continuously distributed measurement (using the linear model).

- A survival time (using the Cox model). In that case Y should contain the last observation time of each individual, and an extra argument d should be supplied which contains the event indicator, which has value *event* if an event occurred.

The function `globaltest` will automatically choose an appropriate model based on Y. To override the automatic choice, use the option *model*.

Now we test the Cell Cycle pathway that interests us:

```
> gt.cc <- globaltest(golubX, "ALL.AML", cellcycle)
> gt.cc

Global Test result:
Data: 72 samples with 7129 genes; 1 pathway tested
Model: logistic

  genes tested Statistic Q Expected Q sd of Q    p-value
1    94     94      69.443     10.312  3.2901 1.0166e-18
```

We conclude that the expression pattern of the cellcycle pathway is notably different between AML and ALL samples. However, as the test on all genes was significant we can generally expect most pathways to be significant as well. To get an impression of how "special" this pathway is, one can use the function `sampling`.

```
> gt.cc <- sampling(gt.cc)
> gt.cc

Global Test result:
Data: 72 samples with 7129 genes; 1 pathway tested
Model: logistic

  genes tested Statistic Q Expected Q sd of Q    p-value comp. p
1    94     94      69.443     10.312  3.2901 1.0166e-18   0.285
```

This gives an extra output column "comparative p", which is the fraction of random genesets of the same size as the cell cycle pathway (94 genes) which have a lower p-value than cell cycle itself. In this case around 28 % of 1,000 random 'pathways' of size 94 have a lower p-value than the Cell Cycle pathway. By default 1,000 random sets are sampled; this number can be changed with the option *ndraws*.

By default the p-value of globaltest is calculated using approximate formulas which are accurate for large sample size, but may be inaccurate for very

small sample size. For 72 arrays they should be accurate enough. For very small sample sizes an alternative is to use the permutation version of globaltest. This recalculates the p-value on the basis of 10,000 permutations of the clinical variable.

```
> permutations(gt.cc)

Global Test result:
Data: 72 samples with 7129 genes; 1 pathway tested
Model: logistic
Using 10000 permutations of Y

  genes tested Statistic Q Expected Q sd of Q p-value
1    94     94      69.443     10.533  3.3604       0
```

The permutation p-value is not so accurate in the lower range as it is always a multiple of one over the number of permutations and also has some sampling variance. If desired, the number of permutations can be changed with the option *nperm* to get more accurate p-values.

It is also possible to adjust the globaltest for confounders or for known risk factors. For example in the Golub Data set we may be afraid for a disturbance due to that the fact that some samples were taken from peripheral blood while others were taken from bone marrow. We can correct for this using the option *adjust*. The option *adjust* can also be used when the study design is different from the simple 'two independent samples' design of the standard global test. In a paired design, for example, put the pair-identifier (as `factor`) in *adjust*.

The user may supply one or more names of covariates in the option *adjust* or supply *adjust* as a *data.frame*. The easiest way of adjustment, however, is by using a *formula* object as input for Y, as follows:

```
> globaltest(golubX, ALL.AML ~ BM.PB, cellcycle)

Global Test result:
Data: 72 samples with 7129 genes; 1 pathway tested
Model: logistic, ALL.AML ~ BM.PB
Adjusted: 99.8 % of variance of Y remains after adjustment

  genes tested Statistic Q Expected Q sd of Q    p-value
1    94     94      69.811      10.25  3.3189 1.6258e-18
```

The test result now also gives the percentage of the variance in Y that was left after the adjustment. It is a crude measure like $1 - R^2$. If the percentage is

low, the adjustment already explained most of the variance of the outcome Y and there was not much residual variance left to test the influence of the genes. To see an example, adjust for "Source" instead of "BM.PB".

The option *adjust* may again be combined with the function *sampling*, but not with *permutation*.

A.3 Multiple global testing

It is also possible to test many pathways at once. To test all KEGG pathways we proceed as follows:

```
> gt.kegg <- globaltest(golubX, "ALL.AML", kegg)
```

The result `gt.kegg` can be displayed and prints a matrix whose rows correspond to the KEGG pathways. It gives the test results for each pathway. We can also display only some of them:

```
> gt.kegg[1:10]

Global Test result:
Data: 72 samples with 7129 genes; 10 pathways tested
Model: logistic

      genes tested Statistic Q Expected Q sd of Q    p-value
00271    10     10      10.103     8.0539  5.7226 2.8564e-01
00272    11     11      51.496    16.9070 12.0600 1.6643e-02
00628     2      2      18.066    22.5560 29.5330 3.8852e-01
00330    51     51      30.768     9.2072  2.9245 6.5854e-07
00920     6      6      12.558     6.5985  4.6089 1.0505e-01
05060    13     13      35.394     8.3092  4.4675 1.1091e-04
00450    14     14      39.648     9.8767  5.3131 2.2772e-04
04010   244    244      43.726    10.2410  2.3327 1.6381e-17
00510    26     26      39.145    10.2670  4.6621 4.5571e-05
04070    82     82      32.255     7.5731  2.0705 9.8340e-13
```

```
> gt.kegg["04110"]

Global Test result:
Data: 72 samples with 7129 genes; 1 pathway tested
Model: logistic

      genes tested Statistic Q Expected Q sd of Q    p-value
04110    94     94      69.443     10.312  3.2901 1.0166e-18
```

The same options described above for the single pathway `globaltest` can be applied to the multiple pathway version of `globaltest` as well.

Two functions allow further processing to be done on the test results. The function `result` extracts the whole matrix of test results, while the function `p.value` only extracts the vector of p-values. The latter function can be used for example when a correction for multiple testing is to be done. Note however that due to the extremely high correlations between the tests for different pathways, many multiple testing procedures are inappropriate for the Global Test. See the *multtest* package for details.

We might want to sort the pathways by their p-value, and show the top five. This can be done as follows

```
> sort.gt.kegg <- sort(gt.kegg)
> sort.gt.kegg[1:5]
```

```
Global Test result:
Data: 72 samples with 7129 genes; 5 pathways tested
Model: logistic

      genes tested Statistic Q Expected Q sd of Q    p-value
04060   246    246      77.853     9.9558  2.7046 1.4526e-30
04610    82     82     112.110     8.8155  3.3998 2.0881e-29
04510   169    169      61.849     9.2011  2.3298 2.4397e-28
04020   205    205      37.144     7.6385  1.6212 2.1932e-24
00590    31     31     213.070    13.5480  6.7527 1.5357e-23
```

A.4 Diagnostic plots

There are various types of diagnostic plots available to help the user interpret the `globaltest` result. The plot `permutations` can serve as a check whether the sample size was large enough not to use the permutation version of `globaltest`. The `geneplot` visualizes the influence of individual genes on the test result. The three plots `sampleplot`, `checkerboard` and `regressionplot` all visualize the influence of individual samples. Of these three, `sampleplot` is probably the most useful.

Permutations histogram

The permutations histogram plots the values of the test statistic Q calculated for permutations of the clinical outcome in a histogram. The observed value of Q for the true values of the clinical outcome is marked with an arrow.

```
> hist(permutations(gt.cc))
```

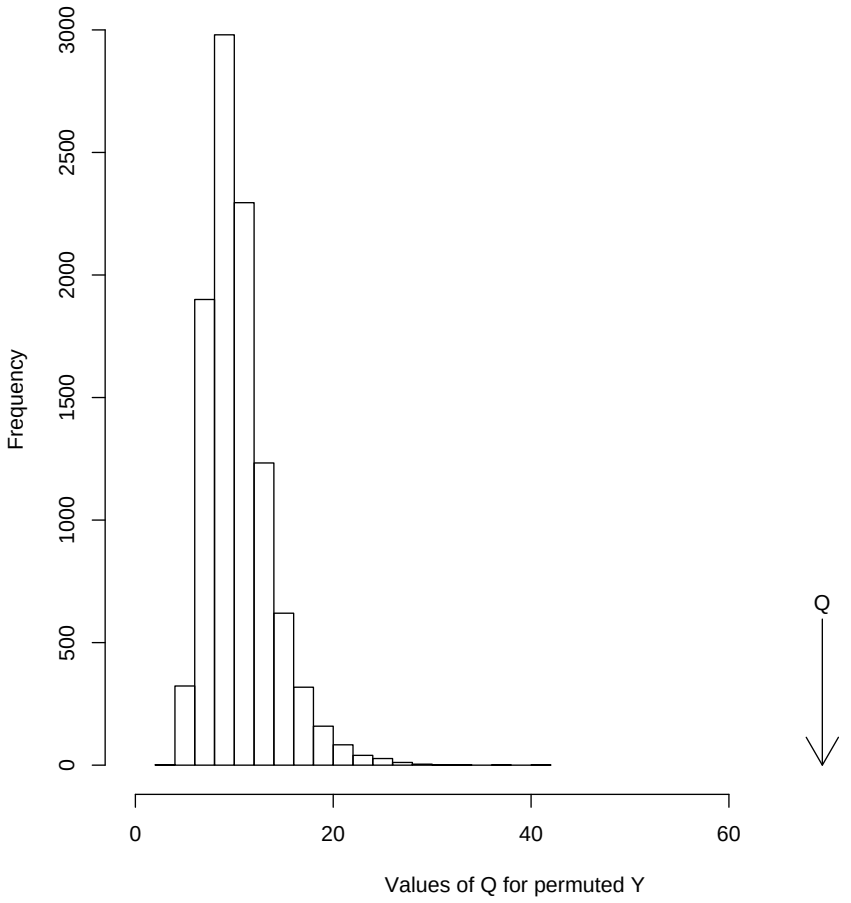

FIGURE A.1: *Histogram of values of the GlobalTest statistic Q for 10,000 permutations of the outcome variable, compared to the value of Q for the observed data.*

The output can be interpreted as a plot of the distribution of the test statistic under the null hypothesis that the pathway is not associated with the clinical variable. Strictly speaking, however, the permutation version of the Global Test is a different test with different properties (especially for survival data). It may give different p-values for small samples.

The function `permutations` may not be used when the adjusted version of `globaltest` was used.

Gene plot

The second diagnostic plot is the Gene Plot, which can be used to assess the influence of each gene on the outcome of the test. The Gene Plot gives a bar and a reference line for each gene tested. The bar indicates the influence of each gene on the test statistic.

A reference line for each bar gives the expected height of the bar under the null hypothesis that the gene is not associated with the clinical outcome (except in a survival model, where the expected height is zero). Marks indicate with how many standard deviations (under the null) the bar exceeds the reference line. Finally the bars are coloured to indicate a positive or a negative association of the gene with the clinical outcome.

The geneplot bars have two interpretations. In the first place, the bars are the Global Test statistic for the single gene pathway containing only that gene. A positive bar that is many standard deviations above the reference line therefore indicates a gene that is significantly associated with the clinical variable in Y. Secondly, the bars indicate the influence of the gene on the test result of the whole pathway (the test statistic for the group is the average of the bars for the genes). Removing a gene with a low bar (relative to the reference line) or a negative bar from the pathway will result in a lower p-value for the pathway, removing a gene with a tall positive bar will have the opposite effect.

To plot the geneplot, use any of the commands below:

```
> geneplot(gt.cc)
> geneplot(gt.kegg, "04110")
> geneplot(gt.kegg["04110"])
```

For a large number of genes the plot might become overcrowded. Use the option *genesubset* to plot only a subset of the genes, *labelsize* to resize the gene labels or *drawlabels = FALSE* to remove them. Alternatively, we can plot part of the geneplot later, as follows

```
> gp.cc <- geneplot(gt.cc)
> plot(gp.cc[1:40])
```

This allows one to look at subsets of a large pathway more closely. The return of the `geneplot` is an object of type *gt.barplot* containing the numbers and names appearing in the plot:

```
> gp.cc[1:10]
```

```
                   influence  expected        sd     z-score
U07563_cds1_at   179.10159883 26.469792 36.932786  4.13269143
```

```
X16416_at          8.49445145  7.459717 10.377423  0.09971017
U33841_at         30.67349290  6.143978  8.543243  2.87121827
U67092_at          0.06648514  4.909773  6.697347 -0.72316514
U67092_s_at        0.03786093  4.317985  5.653305 -0.75710132
X91196_s_at        3.61779645  2.516948  3.409389  0.32288738
U49844_at         73.44818990  6.217861  8.665433  7.75844974
HG4433-HT4703_at  21.67168778  9.114009 12.659503  0.99195670
X59798_at          1.09258726  7.284671  8.735287 -0.70885864
X51688_at         54.98210529 14.400044 20.094406  2.01957013
```

The option *scale* can be used to rescale the bars to have unit standard deviation.

Sample plot

The Sample Plot looks very similar to the Gene Plot and visualizes the influence of the individual samples on the test result. It has a bar and a reference line for each sample tested. The bar indicates the influence of each sample on the test statistic, similar to the `geneplot`. The direction of the bar (upward or downward) indicates evidence against or in favour of the null hypothesis. If a sample has a positive bar, its expression profile is relatively similar to that of samples which have the same value of the clinical variable and relatively unlike the profile of the samples which have a different value of the clinical variable. If the bar is negative, it is the other way around: the sample is more similar in expression profile to samples with a different clinical variable. A small p-value will therefore generally coincide with many positive bars. If there are still tall negative bars, these indicate deviating samples: removing a sample with a negative bar would result in a lower p-value.

If the null hypothesis is true the expected influence is zero. Marks on the bars indicate the standard deviation of the influence of the sample under the null hypothesis. Finally the bars are coloured to distinguish the samples. In a logistic model the colours differentiate between the original groups, in an unadjusted linear model they differentiate the values above the mean from the values below the mean of Y. In an adjusted linear or the survival model they distinguish positive from negative residuals after fitting the null model.

Again, either of the commands below gives the same output.

```
> sampleplot(gt.cc)
> sampleplot(gt.kegg, "04110")
> sampleplot(gt.kegg["04110"])
```

The options of `sampleplot` and the resulting *gt.barplot* object are be handled in the same way as described under “geneplot”.

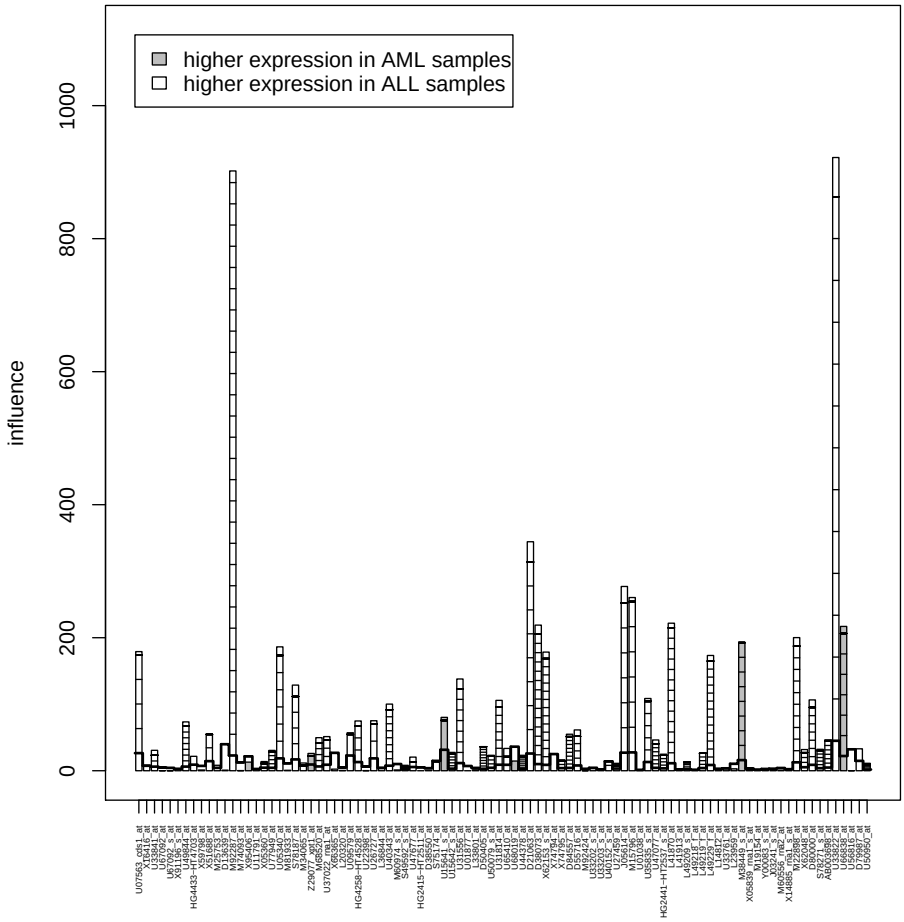

FIGURE A.2: *Gene Plot of 40 genes from the cell cycle pathway in the AML/ALL data. The height of the bar measures association of the expression of that gene with the outcome variable.*

Checkerboard plot

The fourth and fifth diagnostic plot can both also be used to assess the influence of each of the samples on the test result. The checkerboard plot visualizes the similarity between samples. It makes a square figure with the samples both on the X and on the Y-axis, so that it contains all comparisons between the samples. Samples which are relatively similar are coded white and samples which are relatively dissimilar are coded black.

For easier interpretation the samples are sorted by their clinical outcome. If

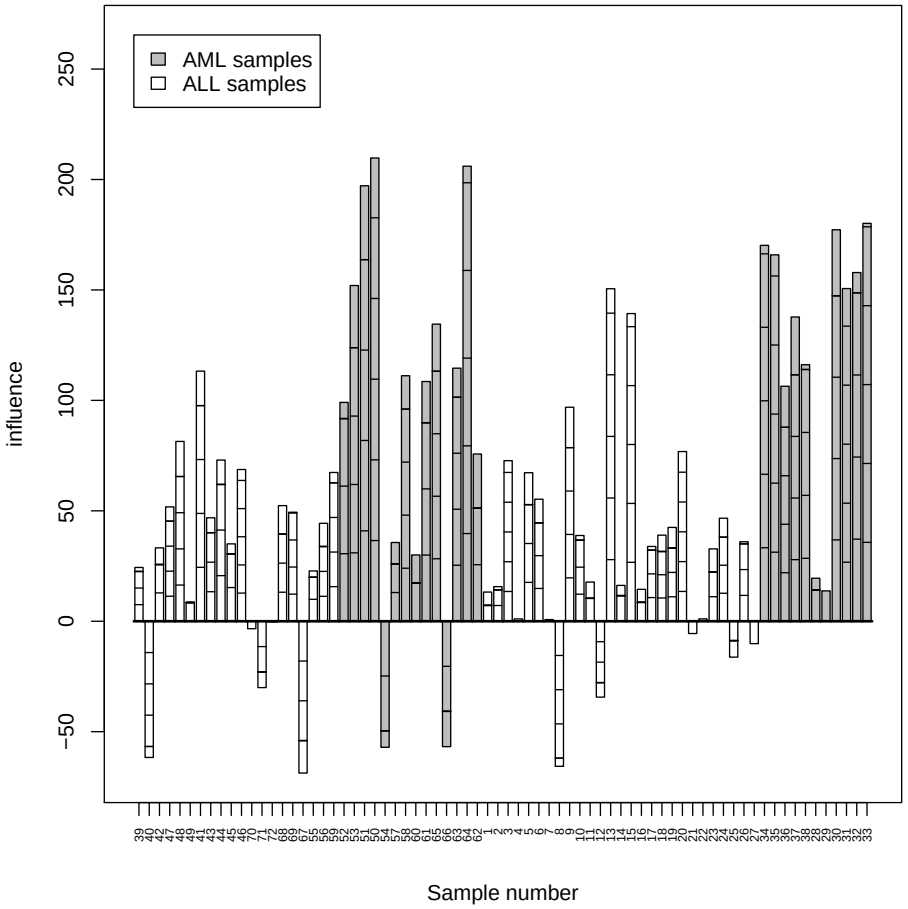

FIGURE A.3: *Sample Plot of the samples in the AML/ALL data set, based on the expression profile of the cell cycle pathway. Positive bars indicate samples whose expression profile is similar to the other samples in the same group; Negative bars indicate samples whose expression profile is similar to samples in the opposite group.*

the test was (very) significant and the clinical outcome has two values, a typical block-like structure will appear. If the clinical outcome was continuous and the test is significant, the black squares will tend to stick together around the upper left and lower right corners. By looking at these patterns some things can be learned about the structure of the data. For example, by looking at samples which deviate from the main pattern, outlying samples can be detected.

```
> checkerboard(gt.cc)
```

```
> checkerboard(gt.kegg, "04110")
```

The function `checkerboard` also has options *labelsize* and *drawlabels*. It returns a legend to link the numbers appearing in the plot if *drawlabels = FALSE* to the sample names.

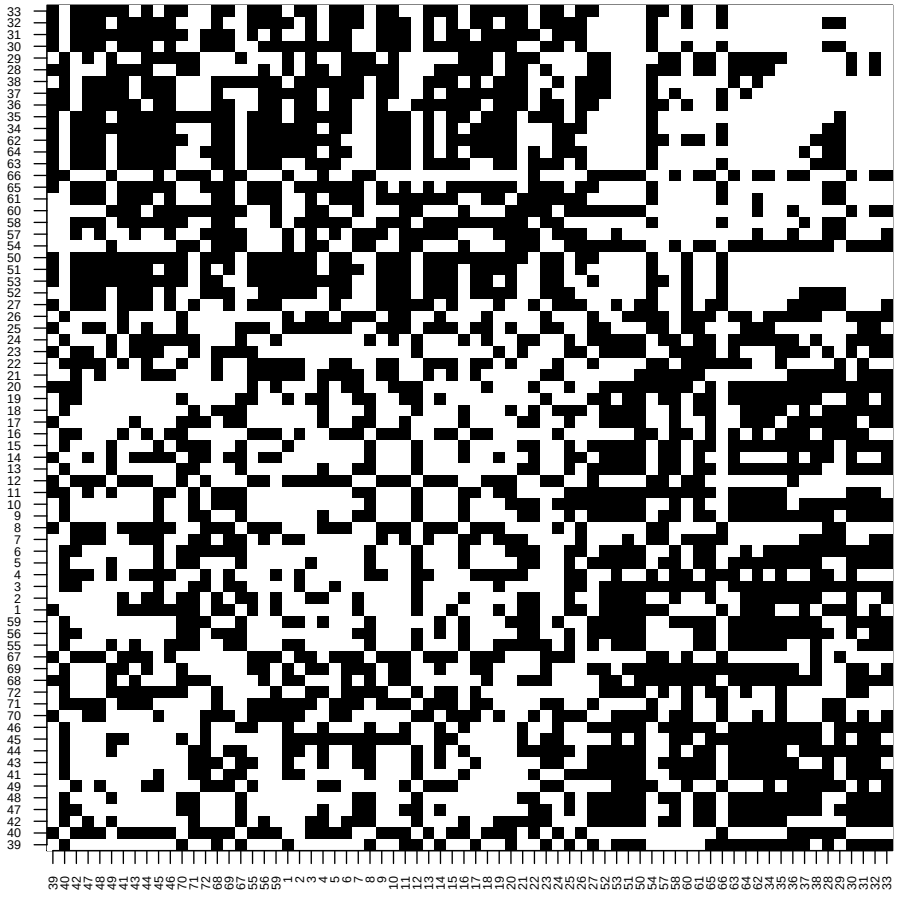

FIGURE A.4: *Checkerboard plot of the samples in the AML/ALL data set, based on the cell cycle pathway. White blocks indicate that samples have similar expression profile, black indicates dissimilar expression profile.*

Regression plot

Using the regression plot an assessment can be made of the influence of each sample on the result of the test. It is an alternative visualization of the

`sampleplot`.

The regression plot plots all pairs of samples, just like the checkerboard plot, but now showing the covariance between their clinical outcomes on the X-axis and the covariance between their gene expression patterns on the Y-axis. The comparisons of each sample with itself have been excluded.

The test statistic of the Global Test can be seen as a regression-coefficient for this plot, so it is visualized by drawing a least squares regression line. If this regression line is steep, the test statistic has a large value (and is possibly significant).

The influence of specific samples can be assessed by drawing a second regression line through only those points in the plot, which are comparisons involving the sample of interest. For example if we are interested the sample with sample name `"1"`, we take the points corresponding to the pairs (1,2) up to (1,72). If the regression line drawn through only these points deviates much from the general line, the sample deviates from the general pattern. This is especially the case if this line has a negative slope, which means that the sample is more similar in its gene expression pattern to the samples with a different clinical outcome than to samples with a similar clinical outcome.

If we want to test sample `"1"`, we say:

```
> regressionplot(gt.cc, sampleid = "1")
> regressionplot(gt.kegg, "04110", sampleid = "1")
```

We can also use this plot for a group of samples, saying for example:

```
> regressionplot(gt.cc, sampleid = c("1", "2"))
```

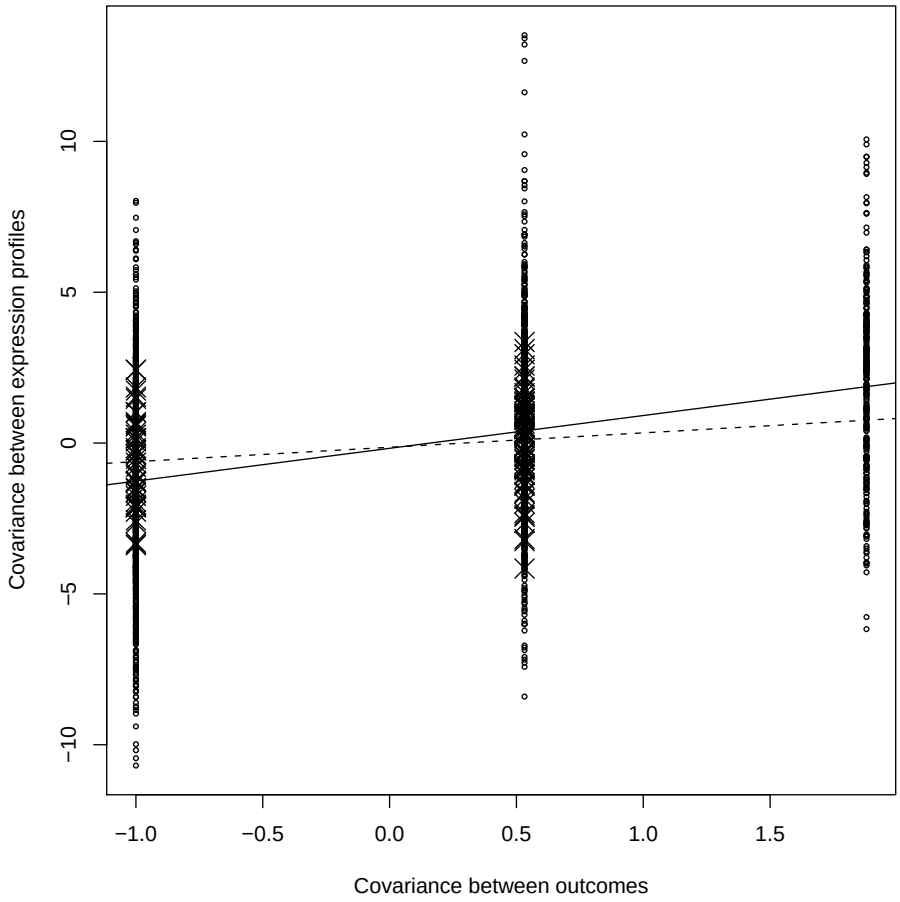

FIGURE A.5: *Regression Plot of the AML/ALL data set, based on the cell cycle pathway, showing the regression of covariance in expression profile on covariance of outcome measure. The dotted line is based only on pairs of samples involving samples "1" and "2".*

Samenvatting

In dit proefschrift worden statistische methoden ontwikkeld voor het analyseren van *microarray*-data. De microarray is een nieuwe technologie uit de moleculaire biologie, die onderzoekers in staat stelt metingen te doen aan genexpressie. Dit is het proces waarmee de informatie die in de genen ligt opgeslagen wordt gebruikt voor de productie van eiwitten. De activiteit van genexpressie kan worden gemeten via het RNA, de belangrijke tussenstap tussen gen en eiwit. Een microarray meet de concentratie van RNA behorend bij ieder specifiek gen en doet dit tegelijkertijd voor tienduizenden genen. Met een microarray is dus het patroon te zien van de gen-expressie van grote aantallen genen in een weefsel of een opgekweekte cellijn.

Door microarrays te vergelijken tussen verschillende typen weefsel, tussen weefsels van verschillende patienten of tussen cellijnen die verschillend behandeld zijn, kunnen allerlei wetenschappelijke vragen beantwoord worden. Interessante vragen zijn er bijvoorbeeld op het gebied van diagnose en prognose. Het vinden van gen-expressiepatronen die onderscheid maken tussen ernstige en minder ernstige vormen van ziekte kan de kwaliteit van diagnoses verbeteren, en daarmee de kwaliteit van de behandeling laten toenemen. Als de microarray bijvoorbeeld gebruikt kan worden om de overleving van borstkankerpatiënten nauwkeuriger te voorspellen zou een groot aantal patiënten een onnodige chemotherapie bespaard kunnen worden. Andere onderzoeksvragen die mogelijk worden gemaakt door microarrays gaan over de functie van genen: door uit te vinden van welke genen de gen-expressie verandert als cellijnen een bepaalde behandeling krijgen, kan iets worden afgeleid over de functie van die genen.

Een statistisch probleem bij het beantwoorden van deze vragen is de hoge dimensionaliteit van de microarray, gekoppeld aan de kleine steekproefgrootte. In een typisch klinisch onderzoek worden microarrays gemaakt van enkele tientallen tot hoogstens enkele honderden patiënten, terwijl voor iedere patiënt de gen-expressie gemeten is van tienduizenden genen. Deze hoge dimensionaliteit leidt tot problemen bij het toepassen van klassieke statistische methoden. Bij het zoeken naar genen die een verschillende gen-expressie hebben onder verschillende experimentele condities moet een zo groot aantal statistische toetsen worden uitgevoerd, dat bekende methoden om te corrigeren voor meervoudig toetsen niet meer goed functioneren. Bij het zoeken naar voorspelregels

om kenmerken van patienten te voorspellen, treedt het verschijnsel van *overfit* op: er zijn vele voorspelregels te vinden die de kenmerken van de onderzochte patiënten perfect voorspellen, maar waarvan de prestaties op nieuwe gevallen allerminst gegarandeerd zijn. De uitdaging die het oplossen van deze problemen biedt, heeft al geleid tot een groot aantal nieuwe statistische methoden.

De statistische methoden die in dit proefschrift ontwikkeld worden, maken zoveel mogelijk gebruik van inhoudelijke kennis uit de biologie, in het bijzonder annotatie van genen, om de kwaliteit en interpreteerbaarheid van de conclusies the verhogen. Annotatie koppelt genen aan de informatie die reeds over deze genen in de literatuur bekend is, bijvoorbeeld in welke celprocessen het gen betrokken is, met welke functies, organen of ziekten het gen is geassocieerd of op welk chromosoom het gen gelokaliseerd is. Een belangrijk concept hierbij is het begrip *pathway*: een *pathway* is een groep genen die met dezelfde functie geassocieerd wordt.

De belangrijkste nieuwe methode in dit proefschrift is de *GlobalTest*-methodologie. Deze wordt uiteengezet in de hoofdstukken 2 tot en met 5. Deze methode biedt een statistische toets die onderzoekers in staat stelt om microarray data te analyseren op het niveau van pathways, in plaats van op het niveau van individuele genen. De onderzoeker gaat dan niet op zoek naar genen waarvan de expressie geassocieerd is met bepaalde kenmerken van patiënten, maar naar pathways waarvan de expressie met deze kenmerken geassocieerd is. Dit is een andere manier van werken, die vaak een tegengestelde onderzoeksvraag heeft. Methoden die zoeken naar individuele genen hebben meestal tot doel de functie van het gen af te leiden uit het kenmerk waarmee de expressie van dat gen associatie vertoont. Als bijvoorbeeld de expressie van genen in gekweekte cellen sterk verandert na kortdurend verhitten van deze cellen, zullen die genen waarschijnlijk een functie hebben bij het herstellen van celschade na hitte. Omgekeerd probeert een methode die zoekt naar pathways juist iets te leren over biologie achter een geobserveerd kenmerk, vanuit de bekende functies van de pathways. Als bijvoorbeeld blijkt dat de expressie van de apoptose-pathway (die de geprogrammeerde celdood regelt) in tumorweefsel dat zich heeft uitgezaaid duidelijk anders is dan in tumorweefsel dat zich niet heeft uitgezaaid, kan geconcludeerd worden dat een storing in de apoptose een stap is in het proces van uitzaaien van tumoren.

Hoofdstuk 2 introduceert de GlobalTest-methodologie die kan toetsen of het gen-expressiepatroon van een bepaalde pathway geassocieerd is met een bepaalde responsvariabele. De details van de methode worden uitgewerkt voor het geval de respons ofwel twee mogelijke waarden aanneemt, ofwel een normaal verdeelde grootheid is.

Hoofdstuk 3 geeft een uitbreiding van dezelfde methodologie naar de si-

tuatie waarin gezocht wordt naar pathways die geassocieerd zijn met overlevingsduur. Het introduceert bovendien de mogelijkheid te corrigeren voor de effecten van verstorende variabelen, wat van groot belang is bij observationeel onderzoek.

Hoofdstuk 4 werkt de wiskunde uit die nodig is om de GlobalTest-methode toe te passen op een responsvariabele die meer dan twee ongeordende waarden aanneemt. De toets die dit artikel presenteert, wordt niet beschreven als een toets voor microarray data, maar in de vorm van een *goodness-of-fit* toets voor het mutinomiale logistische regressiemodel. Dit is een toets waarmee kan worden onderzocht of een dergelijk multinomiaal logistisch model een dataset adequaat beschrijft. Wiskundig gezien is deze toets dezelfde als de toets die nodig is om de GlobalTest-methode te generaliseren naar responsvariabelen met meerdere uitkomstcategorieën.

Hoofdstuk 5 plaatst de toetsen van de vorige drie hoofdstukken in een algemener kader door te laten zien dat ze deel uitmaken van een brede klasse van toetsen die een eenvoudige nulhypothese toetsen tegen een hoogdimensionaal alternatief. Het laat bovendien zien dat dit soort toetsen gemiddeld in een omgeving van de nulhypothese een optimaal onderscheidend vermogen heeft.

Hoofdstuk 6 staat buiten de GlobalTest-methodologie. Het behandelt het probleem hoe een klinische variabele van een patiënt te voorspellen uit de microarray data van die patiënt. Ook hier wordt zoveel mogelijk gebruik gemaakt van kennis over de microarray-data om een goede voorspelregel te construeren. Hiertoe wordt een model van de simultane verdeling van de gen-expressiemetingen en de te voorspellen uitkomstvariabele geconstrueerd. Dit model is gebouwd op de aanname dat er een klein aantal onobserveerbare onderliggende variabelen bestaat, dat zowel de gen-expressiemetingen beinvloedt als de uitkomstvariabele, en dat alle gemeten waarden gepaard gaan met ruis. Op basis van deze eenvoudige aannamen wordt een voorspelregel geconstrueerd die goede eigenschappen heeft in dit model.

Hoofdstuk 7, tenslotte, gaat in op het belangrijke onderwerp van visualisatie van microarray data. Een veelgebruite visualisatiemethode als de puntenwolk geeft al snel een vertekend beeld als er duizenden punten in één diagram weergegeven moeten worden. Het is dan beter om in plaats van een puntenwolk een kleurenweergave van de dichtheid te presenteren, omdat een dergelijke weergave veel duidelijker aangeeft waar de massa van de punten zich bevindt. In het hoofdstuk wordt een snel algoritme gegeven om een dergelijke visualisatie te genereren.

Bibliography

Abraham, B. and G. Merola (2005). Dimensionality reduction approach to multivariate prediction. *Computational Statistics & Data Analysis 48*(1), 5–16.

Al-Shahrour, F., R. Díaz-Uriarte, and J. Dopazo (2004). FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics 20*(4), 578–580.

Albert, A. and E. Harris (1987). *Multivariate interpretation of clinical laboratory data*. New York: Marcel Dekker.

Andersen, P., O. Borgan, R. Gill, and N. Keiding (1993). *Statistical models based on counting processes*. New York: Springer.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics 25*(1), 25–29.

Azzalini, A. and A. Bowman (1993). On the use of nonparametric regression for checking linear relationships. *Journal of the Royal Statistical Society Series B-Methodological 55*(2), 549–557.

Bair, E., T. Hastie, D. Paul, and R. Tibshirani (2004). Prediction by supervised principal components. Technical report, Dept. of Statistics, Stanford University.

Bartholomew, D. J. and M. Knott (1999). *Latent variable models and factor analysis* (2nd ed.). Arnold.

Beer, D. G., S. L. R. Kardia, C. C. Huang, T. J. Giordano, A. M. Levin, D. E. Misek, L. Lin, G. A. Chen, T. G. Gharib, D. G. Thomas, M. L. Lizyness, R. Kuick, S. Hayasaka, J. M. G. Taylor, M. D. Iannettoni, M. B. Orringer, and S. Hanash (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine 8*(8), 816–824.

Beissbarth, T. and T. P. Speed (2004). GOstat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics 20*(9), 1464–1465.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Methodological 57*(1), 289–300.

Benjamini, Y. and D. Yekutieli (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics 29*(4), 1165–1188.

Bernardo, J. M. and A. F. M. Smith (1994). *Bayesian theory*. Chichester: Wiley.

Bolstad, B. M., R. A. Irizarry, M. Astrand, and T. P. Speed (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics 19*(2), 185–193.

Boyle, E. I., S. A. Weng, J. Gollub, H. Jin, D. Botstein, J. M. Cherry, and G. Sherlock (2004). GO-TermFinder: open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics 20*(18), 3710–3715.

Brown, P. J. (1993). *Measurement, regression, and calibration*. Oxford: Oxford University Press.

Burnham, A. J., J. F. MacGregor, and R. Viveros (1999a). Latent variable multivariate regression modeling. *Chemometrics and Intelligent Laboratory Systems 48*(2), 167–180.

Burnham, A. J., J. F. MacGregor, and R. Viveros (1999b). A statistical framework for multivariate latent variable regression methods based on maximum likelihood. *Journal of Chemometrics 13*(1), 49–65.

Burnham, A. J., J. F. MacGregor, and R. Viveros (2001). Interpretation of regression coefficients under a latent variable regression model. *Journal of Chemometrics 15*(4), 265–284.

Burnham, A. J., R. Viveros, and J. F. MacGregor (1996). Frameworks for latent variable multivariate regression. *Journal of Chemometrics 10*(1), 31–45.

Cox, D. R. and D. V. Hinkley (1974). *Theoretical statistics*. Boca Raton: Chapman & Hall.

De Menezes, R. X., J. M. Boer, and J. C. van Houwelingen (2004). Microarray data analysis: a hierarchical t-test to handle heteroscedasticity. *Applied Bioinformatics 3*, 229–235.

Díaz-Uriarte, R. (2005). Supervised methods with genomic data: a review and cautionary review. In F. Azuaje and J. Dopazo (Eds.), *Data Analysis and Visualization in Genomics and Proteomics*, pp. 193–214. Chichester: Wiley.

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

Dudoit, S., J. P. Shaffer, and J. C. Boldrick (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science 18*(1), 71–103.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *Annals of Statistics 32*(2), 407–451.

Efron, B., R. Tibshirani, J. D. Storey, and V. Tusher (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association 96*(456), 1151–1160.

Eilers, P. (1994). Smoothing and interpolation with finite differences. In P. Heckbert (Ed.), *Graphics Gems*, Volume IV, pp. 241–250. London: Academic Pres.

Eilers, P. (2003). A perfect smoother. *Analytical Chemistry 75*(14), 3631–3636.

Eilers, P., J. Boer, G. van Ommen, and J. C. van Houwelingen (2001). Classification of microarray data with penalized logistic regression. In M. L. Bittner, Y. Chen, A. N. Dorsel, and E. R. Dougherty (Eds.), *Proceedings of SPIE*, Volume 4266, pp. 187–198.

Eilers, P. H. C. and J. J. Goeman (2004). Enhancing scatterplots with smoothed densities. *Bioinformatics 20*(5), 623–628.

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

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

Ewis, A. A., Z. Zhelev, R. Bakalova, S. Fukuoka, Y. Shinohara, M. Ishikawa, and Y. Baba (2005). A history of microarrays in biomedicine. *Expert Review of Molecular Diagnostics 5*(3), 315–328.

Fleming, T. R. and D. P. Harrington (1991). *Counting processes and survival analysis*. New York: Wiley.

Ge, Y. C., S. Dudoit, and T. P. Speed (2003). Resampling-based multiple testing for microarray data analysis. *Test 12*(1), 1–77.

Gentleman, R. C., V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. C. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. H. Yang, and J. H. Zhang (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology 5*(10), R80.

Goeman, J. J. and J. Oosting (2005). *Globaltest: testing association of a pathway with a clinical variable*. R package version 3.2.0.

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

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

Goeman, J. J., S. A. van de Geer, and J. C. van Houwelingen (2006). Testing against a high-dimensional alternative. *Journal of the Royal Statistical Society Series B-Statistical Methodology 68*, to appear.

Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science 286*(5439), 531–537.

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The elements of statistical learning*. Springer.

Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics 12*(1), 55–67.

Hosmer, D. W. and S. Lemeshow (2000). *Applied logistic regression* (2nd ed.). New York: Wiley.

Houwing-Duistermaat, J. J., B. H. F. Derkx, F. R. Rosendaal, and J. C. van Houwelingen (1995). Testing familial aggregation. *Biometrics 51*(4), 1292–1301.

Huber, W., A. Von Heydebreck, H. Sültmann, A. Poustka, and M. Vingron (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics 18*(Suppl. 1), S96–S104.

Hwang, J. T. G. and D. Nettleton (2003). Principal components regression with data-chosen components and related methods. *Technometrics 45*(1), 70–79.

Imhof, J. P. (1961). Computing distribution of quadratic forms in normal variables. *Biometrika 48*(3-4), 419–426.

Irizarry, R. A., B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics 4*(2), 249–264.

Jenssen, T. K., W. P. Kuo, T. Stokke, and E. Hovig (2002). Associations between gene expressions in breast cancer and patient survival. *Human Genetics 111*(4-5), 411–420.

Jolliffe, I. T. (2002). *Principal component analysis* (2nd ed.). New York: Springer.

Kerr, M. K., M. Martin, and G. A. Churchill (2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology 7*(6), 819–837.

Klein, J. P. and M. L. Moeschberger (1997). *Survival analysis: techniques for truncated data*. New York: Springer.

Le Cessie, S. and J. C. Van Houwelingen (1991). A goodness-of-fit test for binary regression models based on smoothing methods. *Biometrics 47*, 1267–1282.

Le Cessie, S. and J. C. van Houwelingen (1992). Ridge estimators in logistic regression. *Applied Statistics 41*(1), 191–201.

Le Cessie, S. and J. C. van Houwelingen (1995). Testing the fit of a regression-model via score tests in random effects models. *Biometrics 51*(2), 600–614.

Ledoit, O. and M. Wolf (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis 88*(2), 365–411.

Lesaffre, E. and A. Albert (1989). Multiple-group logistic regression diagnostics. *Applied Statistics 38*, 425–440.

Loader, C. (1999). *Local regression and likelihood*. New York: Springer.

Magnus, J. R. and H. Neudecker (1999). *Matrix differential calculus with applications in statistics and econometrics* (revised ed.). Wiley.

Mansmann, U. and R. Meister (2005). Testing differential gene expression in functional

groups: Goeman's global test versus an ANCOVA approach. *Methods of Information in Medicine 44*(3), 449–453.

Marcus, R., E. Peritz, and K. R. Gabriel (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika 63*, 655–660.

McCullagh, P. and J. A. Nelder (1989). *Generalized linear models* (2nd ed.). Boca Raton: Chapman & Hall.

McLachlan, G. J., R. W. Bean, and D. Peel (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics 18*(3), 413–422.

Mootha, V. K., C. M. Lindgren, K. F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstrale, E. Laurila, N. Houstis, M. J. Daly, N. Patterson, J. P. Mesirov, T. R. Golub, P. Tamayo, B. Spiegelman, E. S. Lander, J. N. Hirschhorn, D. Altshuler, and L. C. Groop (2003). PGC-1 alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics 34*(3), 267–273.

Nguyen, D. V. and D. M. Rocke (2002a). Partial least squares proportional hazard regression for application to DNA microarray survival data. *Bioinformatics 18*(12), 1625–1632.

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

Ogata, H., S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research 27*(1), 29–34.

Pawitan, Y. (2001). *In all likelihood: statistical modelling and inference using likelihood*. Oxford: Clarendon.

Pawitan, Y., J. Bjohle, S. Wedren, K. Humphreys, L. Skoog, F. Huang, L. Amler, P. Shaw, P. Hall, and J. Bergh (2004). Gene expression profiling for prognosis using Cox regression. *Statistics in Medicine 23*(11), 1767–1780.

Pigeon, J. G. and J. F. Heyse (1999). An improved goodness-of-fit statistic for probability prediction models. *Biometrical Journal 41*, 71–82.

R Development Core Team (2005). *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Reiner, A., D. Yekutieli, and Y. Benjamini (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics 19*(3), 368–375.

Schena, M., D. Shalon, R. W. Davis, and P. O. Brown (1995). Quantitative monitoring of gene-expression patterns with a complementary-DNA microarray. *Science 270*(5235), 467–470.

Shevade, S. K. and S. S. Keerthi (2003). A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics 19*(17), 2246–2253.

Simon, R., E. Korn, L. McShane, M. Radmacher, G. Wright, and Y. Zhao (2003). *Design*

and analysis of DNA microarray investigations. New York: Springer.

Simonoff, J. (1996). *Smoothing methods in statistics*. New York: Springer.

Smid, M. and L. C. J. Dorssers (2004). GO-Mapper: functional analysis of gene expression data using the expression level as a score to evaluate Gene Ontology terms. *Bioinformatics 20*(16), 2618–2625.

Smyth, G. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology 3*(1), article 3.

Sohler, F., D. Hanisch, and R. Zimmer (2004). New methods for joint analysis of biological networks and expression data. *Bioinformatics 20*(10), 1517–1521.

Solomon, H. and M. A. Stephens (1978). Approximations to density functions using Pearson curves. *Journal of the American Statistical Association 73*(361), 153–160.

Speed, T. E. (2003). *Statistical analysis of gene expression microarray data*. Boca Raton: Chapman & Hall.

Stone, M. and R. J. Brooks (1990). Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least-squares, partial least-squares and principal components regression. *Journal of the Royal Statistical Society Series B-Methodological 52*(2), 237–269.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B-Methodological 64*, 479–498.

Storey, J. D. and R. Tibshirani (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America 100*(16), 9440–9445.

Tan, Y. X., L. M. Shi, W. D. Tong, and C. Wang (2005). Multi-class cancer classification by total principal component regression (TPCR) using microarray gene expression data. *Nucleic Acids Research 33*(1), 56–65.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society Series B-Methodological 58*(1), 267–288.

Tibshirani, R. (1997). The LASSO method for variable selection in the Cox model. *Statistics in Medicine 16*(4), 385–395.

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

Tusher, V. G., R. Tibshirani, and G. Chu (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America 98*(9), 5116–5121.

Van de Vijver, M. J., Y. D. He, L. J. van 't Veer, H. Dai, A. A. M. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen,

A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards (2002). A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine 347*(25), 1999–2009.

Van Houwelingen, J. C. (1984). Principal components of large matrices with missing elements. In P. Mandl and M. Hukov (Eds.), *Asymptotic statistics 2: Proceedings of the 3rd Prague symposium on asymptotic statistics 29 August-2 September 1983*, pp. 295–302. North Holland.

Van Houwelingen, J. C. (2001). Shrinkage and penalized likelihood as methods to improve predictive accuracy. *Statistica Neerlandica 55*(1), 17–34.

Van Houwelingen, J. C., T. Bruinsma, A. A. M. Hart, L. J. van t Veer, and L. F. A. Wessels (2005). Cross-validated Cox regression on microarray gene expression data. *Statistics in Medicine 24*, to appear.

Van Houwelingen, J. C. and R. M. Schipper (1981). The efficiency of a test based on the asymptotic distribution of the MLE for a linear functional relationship. *Mathematische Operationsforschung und Statistik, Series Statistics 12*(1), 21–30.

Van 't Veer, L. J., H. Y. Dai, M. J. van de Vijver, Y. D. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature 415*(6871), 530–536.

Verweij, P. J. M., J. C. van Houwelingen, and T. Stijnen (1998). A goodness-of-fit test for Cox's proportional hazards model based on martingale residuals. *Biometrics 54*(4), 1517–1526.

Vingron, M. (2001). Bioinformatics needs to adopt statistical thinking. *Bioinformatics 17*(5), 389–390.

Wall, M. M. and R. F. Li (2003). Comparison of multiple regression to two latent variable techniques for estimation and prediction. *Statistics in Medicine 22*(23), 3671–3685.

Wentzell, P. D., D. T. Andrews, D. C. Hamilton, K. Faber, and B. R. Kowalski (1997). Maximum likelihood principal component analysis. *Journal of Chemometrics 11*(4), 339–366.

Westfall, P. H. and S. S. Young (1989). P-value adjustments for multiple tests in multivariate binomial models. *Journal of the American Statistical Association 84*(407), 780–786.

Whittaker, E. (1923). On a new method of graduation. *Proceedings of the Edinburgh Mathematical Society 41*, 63–75.

Wigle, D. A., I. Jurisica, N. Radulovich, M. Pintilie, J. Rossant, N. Liu, C. Lu, J. Woodgett, I. Seiden, M. Johnston, S. Keshavjee, G. Darling, T. Winton, B. J. Breitkreutz, P. Jorgenson, M. Tyers, F. A. Shepherd, and M. S. Tsao (2002). Molecular profiling of non-small cell lung cancer and correlation with disease-free survival. *Cancer Research 62*(11), 3005–3008.

Wold, S., A. Ruhe, H. Wold, and W. J. Dunn (1984). The collinearity problem in linear regression: the partial least-squares (PLS) approach to generalized inverses. *SIAM*

Journal on Scientific and Statistical Computing 5(3), 735–743.

Wright, G. W. and R. M. Simon (2003). A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics 19*(18), 2448–2455.

Wu, Z. J., R. A. Irizarry, R. Gentleman, F. Martinez-Murillo, and F. Spencer (2004). A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Association 99*(468), 909–917.

Yang, Y. H., S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research 30*(4), e15.

Zeeberg, B. R., W. M. Feng, G. Wang, M. D. Wang, A. T. Fojo, M. Sunshine, S. Narasimhan, D. W. Kane, W. C. Reinhold, S. Lababidi, K. J. Bussey, J. Riss, J. C. Barrett, and J. N. Weinstein (2003). GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biology 4*(4), R28.

Zhang, B., D. Schmoyer, S. Kirov, and J. Snoddy (2004). GO Tree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *BMC Bioinformatics 5*, 16.

Zhang, J. (2004). *GO: a data package containing annotation data for GO*. R package version 1.6.5.

Curriculum Vitae

De auteur van dit proefschrift werd geboren op 24 juni 1976 in Leiderdorp. Hij bezocht vanaf 1988 het Stedelijk Gymnasium te Leiden, waar hij in 1994 zijn V.W.O. diploma behaalde. In datzelfde jaar begon hij aan de Universiteit Leiden met de studie wiskunde, die hij in 2001 afrondde. De doctoraalcriptie die hieruit voortkwam, getiteld *Using survival to predict survival*, werd in 2002 bekroond met de scriptieprijs van de Vereniging voor Statistiek, en is in verkorte vorm gepubliceerd in *Statistica Neerlandica*. In dezelfde periode studeerde de auteur ook geschiedenis aan de Universiteit Leiden. Deze studie werd in 2001 *cum laude* afgerond met een doctoraalscriptie op het gebied van de historische methodologie, getiteld *Grondslagen van de vergelijkende methode*.

Het onderzoek dat leidde tot dit proefschrift werd uitgevoerd tussen 2001 en 2005, toen de auteur als promovendus verbonden was aan de afdeling Medische Statistiek en Bioinformatica van het LUMC en aan het Mathematisch Instituut van de Universiteit Leiden. De resultaten van het onderzoek werden eerder gepresenteerd op verschillende conferenties, workshops en colloquia, onder andere in Freiburg, Aarhus, Kaunas, Diepenbeek, Wye, Heidelberg, Boston, Marseille en Leicester. De presentatie op de laatste conferentie werd met een prijs bekroond. In deze periode heeft de auteur ook meegewerkt aan de organisatie van de workshop *On high-dimensional data*, die gehouden werd aan het Lorentz Center in Leiden in september 2002. Daarnaast was hij van 2003 tot 2005 secretaris van het Leids Promovendi Overleg.

Op dit moment werkt de auteur als wetenschappelijk onderzoeker aan de afdeling Medische Statistiek en Bioinformatica van het LUMC.