ABSTRACT

Title of dissertation: ANTI-PROFILES FOR ANOMALY CLASSIFICATION AND REGRESSION

Wikum Dinalankara, Doctor of Philosophy, 2015

Dissertation directed by: Professor Hèctor Corrada-Bravo
Department of Computer Science

Anomaly detection is a classical problem in Statistical Learning with wide-reaching applications in security, networks, genomics and others. In this work, we formulate the anomaly classification problem as an extension to the detection problem: how to distinguish between samples from multiple heterogenous classes that are anomalies relative to a well-defined, homogenous, normal class. Our formulation of this learning setting arises from studies in cancer genomics, where this problem follows from prognosis and diagnosis applications. Anomaly classification is also a natural model for applications in electronic health record data, network intrusion-detection, and other areas.

Standard binary and multi-class classification schemes are not well suited to the anomaly classification task since they attempt to directly model these highly unstable, heterogeneous classes. In this work, we show that robust classifiers can be obtained by modeling the degree of deviation from the normal class as a stable characteristic of each anomaly class. To do so, we formalize the anomaly classification problem, characterize it statistically and computationally via kernel methods

and propose a class of robust learning methods, anti-profiles, specifically designed for this task.

We focus on an open area of research in cancer genomics which motivates this project: the classification of tumors for prognosis and diagnosis. We provide experimental results obtained by applying the anti-profile method to gene expression data. In addition we extend the anti-profile approach to use kernel functions, and develop a support-vector machine (SVM) based method for classification of anomalies based on their deviation from a stable normal class. We provide experimental results obtained by applying this method to genetic data to classify different stages of tumor progression, and show that this method provides much more stable classifiers than the application of regular classifiers. In addition we show that this approach can be applied to anomaly classification problems in other application domains.

We conclude by developing an SVM for censored survival information and demonstrate that the anti-profile method can produce stable classifiers for modeling the clinical outcome of clinical studies of cancer.

ANTI-PROFILES FOR ANOMALY CLASSIFICATION AND REGRESSION

by

Wikum Dinalankara

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2015

Advisory Committee:
Professor Hèctor Corrada-Bravo, Chair/Advisor
Professor Mihai Pop
Professor Sridhar Hannenhalli
Professor Ramani Duraiswami
Professor Zia Khan

Preface

Portions of the material presented here have either been published in peer-reviewed journals or are being prepared for submission to peer-reviewed journals. Chapter 3 of this thesis has been published in Cancer Informatics under the title "Gene Expression Signatures Based on Variability can Robustly Predict Tumor Progression and Prognosis" [15]. Chapters 2 and 5 are currently under preparation for submission.

Dedication

I dedicate this dissertation to my parents - without their love and support and faith in my abilities this dissertation would not have been possible.

Acknowledgments

I am indebted to many people, mentors, colleagues and friends, without whom I would not have been able to complete this dissertation.

First and foremost I would like to thank my advisor Hèctor Corrada-Bravo for this guidance and support. He patiently advised and mentored me as I slowly trudged through my graduate career, and the skills and knowledge I have acquired I owe it to him. I would also like to thank my dissertation committee for their support and advice.

My special thanks go to Nick Thieme, who collaborated with me on the anomaly SVM project.

I would also like to thank my friends - Kwame Okrah, Joyce Hsiao, Joseph Paulson, Hisham Talukder and everyone in the Bravo Lab. They have never been just colleagues in my life, but dear friends. They're the best lab mates you could ask for, and without their help my graduate student life would have been very difficult.

My thanks also go to all my colleagues at the Center for Bioinformatics and Computational Biology at the University of Maryland.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Anomaly detection is a classical problem in Statistical Learning with wide-reaching applications in security, networks, genomics and others. In this dissertation, we formulate the anomaly classification problem as an extension to the detection problem: how to distinguish between samples from multiple heterogenous classes that are anomalies relative to a well-defined, homogenous, normal class. Our formulation of this learning setting arises from studies in cancer genomics, where this task is a natural application. It is also a natural model for applications in electronic health record data, network intrusion-detection, and other areas.

Anomaly classification remains a largely unfocused area of research. Usually an ad-hoc approach is taken resulting in solutions specific to the problem domain that are not generalized to multiple domains. These approaches usually consist of either applying a standard classification method, e.g., Support Vector Machines, to directly distinguish the anomalous classes, or applying an anomaly detection technique [13]. The former approach does not integrate relevant information that could be obtained from the normal class while the latter approach makes it difficult to incorporate differences between the anomalous classes themselves. Attempts to combine both approaches have usually resulted in bootstrapping; while this might

provide good solutions for the problem at hand, it cannot be considered as a general approach to anomaly classification [34].

In this dissertation we present an approach that provides an alternative to the above methods. We formalize the anomaly classification problem as a general learning setting and characterize it statistically and computationally using kernel learning methods. Since kernel functions can be interpreted as pairwise similarity functions (satisfying certain conditions), they serve as a natural framework to measure within normal class similarity and anomalous deviation from this class. We also develop kernel-based classification approaches specifically designed for the anomaly classification task and show that they produce accurate classifiers with higher stability than standard classification methods.

The main inspiration behind this approach comes from a widely studied problem in cancer genomics: the development of microarray-based methods for diagnosis and prognosis of cancer [37]. The main difficulty in developing such techniques have been the lack of stable markers in cancer gene expression profiles. Apart from a few exceptions, many gene signatures developed have provided poor results when tested on independently obtained data, indicating that the signature is not adequately robust to be deployed in a clinical setting.

We suggest that more robust analyses can be performed by taking gene expression variability into account. While variability in gene expression has traditionally been ignored or discarded from consideration for biological relevance, recent studies have demonstrated that variability analyses can yield important biological insights. In the first chapter we present a survey of recent studies that analyze variability in

gene expression and methylation.

In particular, recent work by Corrada-Bravo et al. [11] demonstrates that by modeling the increased gene expression variability in cancer, a statistical method can be developed which provides a stable and robust universal predictor of cancer. The underlying observation behind this approach is that certain genes will consistently show higher across sample variability in cancer as compared to normal samples. The hyper-variability among these genes can be leveraged to measure deviation from a stable profile of healthy samples, resulting in a cancer anti-profile.

Here we further advance this approach by demonstrating that it can be used as a prognosis predictor of cancer survival. We demonstrate that using the same genes that universally exhibit hyper-variability in cancer, the degree of heterogeneity in tumor gene expression as measured with respect to healthy samples can be used as a measurement of potential malignancy or risk of relapse.

These insights form the principle behind the anomaly classification approach presented here: that by primarily modeling the solution based on a stable, cohesive non-anomalous class while incorporating the deviations of the anomaly classes for discrimination results in a more robust and stable classifier. To demonstrate this, we present the anomaly SVM (aSVM), a kernel method based on the same principles as the anti profile method for classification of tumor prognosis. We compare this method with other SVM methods: the regular SVM and the one-class SVM which is generally used for anomaly or novelty detection. These are applied to genomic datasets consisting of gene expression and DNA methylation levels, as well as other classification problems to demonstrate their broad utility.

An important facet of bioinformatic study of diseases is the analysis of survival information of patient obtained through clinical trials. Since a number of data sources analyzed in this dissertation contain such information, and the usual formulation of the SVM being unsuitable for a thorough analysis of such data, we examine the survival SVM. The survival SVM is an SVM formulation designed for censored survival data. We demonstrate how the anti-profile approach can be applied to derive a survival SVM for anomaly classification. We present two SVM formulations for utilizing both normal and anomalous survival data, and demonstrate their efficacy.

In summary, the contributions of this dissertation are the following: 1) formalization of the anomaly classification task and its statistical and computational characterization based on kernel methods, 2) compiling a literature survey of recent advances in gene expression variability analysis 3) the development of classification methods designed specifically for the anomaly classification task, 4) application of this methodology to produce stable prognosis and diagnosis tools for cancer genomics, 5) the application of this methodology to censored survival data to make efficient predictions of prognosis.

Chapter 2: Gene Expression Variability Analysis: A Survey

2.1 Introduction

It is common for gene expression studies to limit their focus to quantifying and characterizing gene expression levels without paying significant attention to characterizing variability. In some instances studies may make assumptions regarding the variability of gene expression levels that may be found to be unsustainable under scrutiny, or any patterns observed in variability may be discarded as a confounding noise factor introduced solely by error [33].

However, in the last decade a number of researchers have sought to make gene expression variability a central focus of a study. These studies present intriguing new avenues for research in gene expression analysis. Given that certain patterns of gene expression variability have been linked to a diseases, deriving new methods and techniques for estimating and characterizing, and performing differential analysis with regard to variability can provide considerable insight into characterizing the behavior of diseases. EV is already known to be related HIV susceptibility [39] and to a number of neurological diseases; for certain genes EV is decreased in Schizophrenia and increased in Parkinson's [46]. Cancer research in particular, as we will see below, have much to gain from understanding how to model and account for expression

variability (EV), as it may increase the potential for developing successful targeted therapeutics.

In this survey we explore some recent studies that have made significant contributions to EV analysis. We present how researchers have quantified and estimated EV in gene expression data, how the correlation of EV with certain gene level factors has been characterized, and how differential analysis with regard to EV has been carried out by different research groups. In addition we present how similar analysis can be performed with respect to expression variability in methylation as well. We also discuss some of the conclusions resulting from these studies and assess their biological implications.

2.2 Estimating variability

The recent study by Alemu et al. [3] forms a significant starting point for exploring gene expression variability. The main aim of this study was to deduce what can be established as general characterizations of expression variability (EV) with respect to a number of genomic and epigenomic features. The authors argue that tissue specific expression variability of a gene can be considered a property of a gene, and provide a number of reasons as to why this is justifiable. Among these are the known relationships between expression variability and genes associated with certain biological functions or diseases. For example, low expression variability has been observed in transporters, channels and metabolizing enzymes [70], whereas the converse has been observed in certain genes associated with diabetes [68]. An additional

justification for the authors' reasoning would be the evidence provided by previous studies on Lymphoblastoid cell gene expression levels, which have been shown to have significantly correlated EV levels, even when the actual gene expression levels were not.

Alemu et al. follow this with a comprehensive statistical analysis demonstrating how gene expression variability can be estimated and can be analyzed with respect to various gene level characteristics, yielding a number of interesting biological insights.

Afsari, Geman and Fertig [2] have demonstrated a different route for exploring EV, where the main focus is the detection of signaling pathways in relation to cancer. The value behind such a study is based on possibility that certain chains of protein-protein interactions occurring within a patient may be indicative of malignant tumor activity. Then primary motivation for discovering such molecular pathways is that a drug targeted at such a pathway would yield an effective treatment.

However, detection of tumor activity in this manner is a significantly difficult task that needs careful consideration in such an approach. The authors point out that despite the advances in methods that directly detect phosphorylation changes in a protein network, any studies that utilize the large collections of existing transcriptional data need to design statistical techniques to yield useful and interesting results. Thus a key step in their approach is the selection of genes that are differentially expressed as a result of tumor activity. The authors asses and comment on the suitability of a number of exiting techniques for this task, which we shall explore further below.

Alemu et al. have used microarray data for their analysis of EV. They curated a set of 688 microarray samples from Gene Expression Omnibus (GEO). This set was obtained by filtering out undesirable samples from an initial repository of 7741 samples spanning over 175 studies. In addition to consideration of batch effects in such experiments, when assembling such a large collection from diverse sources it is imperative that appropriate pre-processing steps be taken so that comparisons across samples can be made. For this the authors used fRMA, a single chip normalization procedure, followed by employing the Gene Expression Barcode procedure for calculation of standardizes gene expression measurements. Further filtering of samples was carried out by detecting and removing samples that were too similar to each other (via computing Euclidean distance between two given samples) and removing samples with extremely large GNUSE values. The final pre-processing step was to remove disease samples, as the focus of the analysis was to characterize the gene EV of normal human tissue. Since many diseases are linked to changes in EV, the inclusion of disease samples would have made it extremely difficult to separate changes in EV due to any disease conditions apart from changes in EV due to other factors considered. After these steps the samples were deemed ready for analysis by measuring correlation between a tissue specific EV measure and a given feature associated with a gene. This collection was comprised of 74 tissues of which 41 were the primary focus of the analysis as these tissues contained five or more samples.

For estimating gene EV, the mean and standard deviation was calculated for each probeset and tissue, where consideration was limited to probesets that were

considered to be expressed for each tissue. This latter condition was applied through selecting probesets that yielded a standardized expression value greater than 2.54. Following this, a local likelihood estimation method is used for modeling variance as a function of the mean. This would allow the measurement of excess variance after accounting for variance that would be expected for a given mean, which the authors utilize to form a measurement of EV.

Alemu et al. justify their use of such a modeling procedure by demonstrating the inadequacy of standard noise measurements for their purpose. A standard approach to measuring the extent of variability present with respect to the mean is to measure the coefficient of variation (CV), defined as the ration of the standard deviation to the mean. Computing CV for this data demonstrates that the CV is biased towards low mean expressions, making it unsuitable for EV analysis.

A gamma model was used to estimate the expected variability of a probeset in each tissue given the overall expression. After computing the expected variance given the mean, EV is measured as the ratio of observed variance to expected variance. Unlike CV, this estimate was not found to have any biases with respect to the mean expression level. This EV measurement obtained for each probeset and tissue was aggregated to form a gene specific measurement of EV by taking the median probeset specific EV from probesets mapped to a gene. Following this, the gene specific EV estimate was used to assessing to what extent it correlated with the various genetic and epigenetic determinants that were considered in the study. The following types of properties of genes were considered:

1. genomic properties (gene size, gene structure, regulatory elements in the gene's vicinity)
2. epigenomic marks such as DNAse hypersensitivity
3. interacting partners of the gene's protein product
4. pathways and biological functions
5. disease associations
6. regions of natural structural variations in human populations

While Alemu et al. have retained their focus to assessing EV under normal (i.e. healthy) conditions, other researchers have sought to explore EV under diseases conditions. While a number of diseases are known to be exhibit certain patterns of gene expression variability, the most widely studied among these is cancer. Recently, Irizzary and Feinberg [19] have explored the possibility that increased variability within a given phenotype might lead to increased fitness. They hypothesize that increased variance in the genotype may increase fitness via increased variability of the phenotype, regardless of any significant change of the mean phenotype. This avenue of research has been further continued by Hansen et al. [30]. Both studies extend their exploration to exploring epiegenetic variability in addition to gene EV.

Irizzary and Feinberg have used CHARM microarrays for estimating DNA methylation levels. After quantile normalization and age correction, the raw data from the arrays were translated into methylation percentages which were used to obtain a methylation profile. This was accomplished by smoothing the individual

probe level methylation percentages across genomic regions. In addition they also carried out whole genome bisulfite sequencing for three colon cancer samples, their matched normal mucosa, and two colon adenomas.

Hansen et al. used a custom Illumina array for methyation where 151 cDMRs (cancer-specific DNA methylated regions) were analyzed over 290 samples from both matched normal and cancer tissues for colon, breast, lung, thyroid and Wilms' tumor. The Illumina array consisted of 384 probesets over 139 regions, with 111 normal samples, 122 cancers, 20 colon premalignant adenomas and 27 additional normal samples. Similar to Irizzary and Feinberg, this study also used smoothing of individual probe level methylation percentages to obtain a comprehensive methylation profile for a given samples over a genomic region. This allowed the authors to compare the shifting of highly methylated region boundaries between cancer and normal samples. They further supplement this analysis by making comparisons between the variability in methylation levels and variability in gene expression levels via publicly available microarray data.

2.3 Differential analysis of variability

A significant aspect of exploring gene expression level or methylation level variability consists of carrying out a differential analysis to gain insight into how EV can be characterized under disease conditions.

In their analysis of methylation level variability in tumors across multiple cancer types, Hansen et al. first carried out comparisons between the normal and

cancer variance based on the differentially methylated regions previously observed by Irizzary and Feinberg. For each cancer type, they plotted the standard deviation of methylation percentage of cancer samples against that of normal samples, for each probeset. This provided an initial assesment of the amount of variability observed in each of cancer and normal states. They observed that in all cancers, the majority of probesets demonstrated higher variability than in normals.

Cellular heterogeneity, age of the individual, and genetic heterogeneity were ruled out as causes of the observed hypervariability of methylation in cancers. By examining the normal methylation profile obtained from publicly available methylation data for normal samples the authors were able to conclude that their choice of CpG sites did not fully account for the hypervariability observation, suggesting that it is rather a general property of cancer methylation.

As mentioned above, in the next stage of the study Hansen et al. estimated methylation profiles are each samples by smoothing individual methylation levels across nearby CpG sites. A comparison of such profiles obtained via whole genome bisulfite sequencing enabled to authors to locate discrepancies in the boundaries of highly-methylated or highly-unmethylated regions between normal, cancer and adenoma samples.

In the final step of the study, a complementary analysis was carried out for gene level expression variability between normal and cancer. Using microarray gene expression data from both normal and cancer samples, the authors were able to map 6869 genes to the differentially methylated regions identified in previous steps. A gene ontology enrichment analysis for differentially expressed genes revealed a

number of cell cycle genes as being associated with methylation region boundary shifts between cancer and normal. We examine these conclusions further in the next section.

Afsari et al. [2] have provided a different route for differential analysis of gene expression variability. The main focus of their analysis is pathway regulation for which they utilize EV analysis. They point out that the detection of differentially expressed pathway regulation based on transcription data falls into two major approaches: over-representation methods and enrichment methods. The former involves detection of genes that are annotated to pathways and also differentially expressed to a significant degree. The latter usually involves obtaining a statistic that characterizes the differential expression of genes in a pathway relative to a null distribution.

However, while both approaches can yield robust results, given that individual gene expression can be highly variable in tumors these methods may not be adequate to fully characterize these complex changes in variability. The main contribution of this study is to review solutions that can be applied to this issue, particularly Differential Rank Conservation (DIRAC) and expression variation analysis (EVA). These solutions - collectively referred to as differential variability analysis, form a third approach in addition to over-representation methods and enrichment analysis.

The curated databases that contain gene levels annotations for pathways play a crucial role in identifying significantly perturbed pathways (which is expected as a result of tumor heterogeneity) in gene level comparisons between phenotypes.

Over-representation:

1. Select a set of genes that are differentially expressed (DE) between two phenotypes.

2. Calculate a gene set statistic for each pathway that compares each set of pathway genes to the set of DE genes identified in step 1.

3. Pathways whose members are significantly expressed for DE genes are considered significant.

Enrichment:

1. Calculate differential expression of genes between two phenotypes (e.g. t-statistic or Z-statistic).

2. Calculate a cumulative statistic from the values in step 1 to characterize differential activity of a pathway which allows comparison with a null distribution (e.g. Kolmogorov-Smirnov test, sum, mean, maxmean). The null hypothesis may be implemented using an alternative set of genes or by permuting sample labels.

Both methods infer coordinated, average expression changes between phenotypes in sets of genes annotated to a pathway. Since enrichment methods look at cumulative changes rather than using hard thresholds as in over-representation, enrichment methods are more sensitive to detection of coordinated expression changes, albeit at the cost of yielding more false positives. However, both methods are best suited when the changes in the pathways occur consistently within a given phenotype, and this assumption may not hold for certain cases, especially in cancer

which is known for its highly heterogeneous nature. This makes the third family of methods - analysis of differential variability in gene expression, a relatively novel but extremely important contribution. This approach is characterized by assessing the variability for a given pathway and compares it between phenotypes to determine whether the given pathway as a whole is differentially expressed between the phenotypes in terms of variability.

Since heterogeneous alterations that occur in tumors can increase variability of expression of genes, over-representation and enrichment methods can also be applied using variability statistics. However, complex methods that employ multivariate statistics may be more robust in detecting alterations in expression that depend on interactions among the genes in a pathway.

The first differential variability expression method published was by Zhang et al. [89], which computed correlation between all pairs of genes for a given pathway, and then calculated a z-score to assess the difference in pairwise interactions for this pathway between two phenotypes. This method was later improved upon by Watkinson et al. [87]. A recently method published by Liu et al. [41] follows a similar but more complex approach, where instead of simply correlation sum, difference, max and min are used respectively to measure gene pair cooperation, competition, redundancy and dependency. This is followed by applying a maxmean statistic over all pairs, which is the same procedure as followed by Zhang et al. Ochs et al. [54] in contrast provides a method for measuring pathway dysregulation via outlier analysis.

Afsari et al. argue that rank based techniques are superior to techniques that use normalized gene expression values in the following ways:

- They are more robust to the preprocessing and normalization of data.
- They can perform as efficiently as other classification methods for distinguishing between phenotypes.
- They are simpler to explain and interpret in biological terms.

The DIRAC method proposed by Eddy et al. [16] is one of the variability methods that is rank-based. For a given pathway and phenotype, DIRAC generates a binary template for the ordering of the expression values for the genes in the pathway, and then the distance between the training samples and the template is calculated as the Hamming distance over the gene pairs. Permutation of labels are used to estimate a p-value for estimating the significance of the difference in variability. Eddy et al. have shown that highly dysregulated pathways revealed by this method are associated with bad prognosis.

Afsari et al. propose an alternative to DIRAC in order to increase computational feasibility, EVA. For a given pathway and phenotype, EVA calculates the Kendall-tau distance between the rank vectors corresponding to two randomly chosen expression profiles. Averaging the Kendal-tau distance between each pair of samples from that phenotype provides a variability statistic the phenotype. A p-value is computed analytically from the difference between the empirical Kendall-tau statistics.

In addition they provide an R package, GSReg where the EVA method is implemented. They analyze the same data that had been previously analyzed using DIRAC to show that the results given by both methods are highly similar, and show

that the EVA method is more computationally efficient.

So far the methods explored have all carried out their analysis in a supervised setting; i.e. the phenotypical labels of the relevant data has been fully known. Ghosh and Li [22] have proposed a further addition to these procedures: a method by which these tests can be applied in the absence of group labels (e.g. healthy vs. diseased). Their testing approaches are based on the $C(\alpha)$ principle, originally proposed by Neyman and Scott [53] and applied for rare variant detection.

In discovering genes that are differentially expressed between two conditions, typically finding genes that are up-regulated or down-regulated in diseased tissue with respect to non-diseased tissue is the goal. The test of differential expression used is usually mean-based, such as the t-test. However it has been observed that other patterns of differential expression may exist as well. For example, Tomlins et al. [79] have observed that only a fraction of samples in one group were over-expressed relative to those in the other group. Based on this line of research Tomlins et al. have developed a ranking method for outlier analysis using gene expression data, and further Tibshirani et al. [75] have provided a statistical method for calculating a p-value for assessing significance for this model. Ghosh and Chinnaiyan [21] have adapted these ideas to develop a similar framework which is non-parametric.

The basic model considered by Ghosh and Chinnaiyan is as follows:

$$Y_{gi}|Z_i = 0^{ind} F_{0g}(y)$$

$$Y_{gi}|Z_i = 1^{ind} \pi_{0g} F_{0g}(y) + (1 - \pi_{0g}) F_{1g}(y)$$

where for gene g and sample i, Y_{gi} is the expression measurement and Z_i is

a binary indicator for the group. F_{0g} and F_{1g} denote the gene-specific distribution functions for the expression in the non-differential and differentially expressed genes, and π_{0g} is the proportion of samples that show no differential expression for gene g.

When Z is available this can be considered as supervised outlier profile analysis. To apply this for a situation where Z is not available, Ghosh and Li first integrate the model to remove Z:

$$Y_{gi}^{ind} F_{0g}(y) + c_g F_{1g}(y)$$

Here $c_g = (1 - \pi_{0g})(1 - P(Z_i = 0))$. This can be generalized for infinite subtypes as follows:

$$Y_{gi}^{ind} \sum_{k=0}^{\infty} c_k F_k(y_{gi}; \theta_{gk}, \sigma_{gk}^2)$$

where F_k is a normal distribution.

With this model the null hypothesis is that there are no subtypes for gene g, while the alternative hypothesis is that there exists one or more subtypes. Therefore this model can be used for the detection of overdispersion. The $C(\alpha)$ test method can now be applied, which in fact corresponds to calculating the overdispersion score, which in turn is the second derivative of the probability density of the data divided by the density.

The authors applied three tests for each gene to calculate a set of gene specific test statistics: a test based on skewness, a test based on kurtosis, and the K2 test, which combines both skewness and kurtosis. If multiple types of information were available for the same data (e.g. gene expression and copy number) are available, a bivariate extension was developed based a similar extension for the B-H procedure

developed by Philips and Ghosh [59]. Here two sets of p-values are calculated for the samples for the two data platforms. These p-values are placed on a two-dimensional grid and a Voronoi tessellation is calculated. The area of the Voronoi cell of each sample, together with the location of the point on the grid is used for assessment, as samples that show evidence against the null hypothesis are expected to cluster near the origin and be enclosed in small Voronoi cells.

These methods were tested by the authors first with simulated data and then with prostate cancer with copy number and transcript mRNA microarray data. The data contains both samples labeled as cancer and non-cancer, and using all these samples does not yield useful results due to large differences in expression patterns between the two labels. By limiting the analysis only to cancer samples, using the gene expression data revealed about 20% of the genes as statistically significant (out of 7534 genes) based on an outlier transcript profile, and using the copy number data revealed about 45% of the genes to be significant. Combining both types of data jointly yielded more useful results.

This study demonstrates that higher order moment based tests are useful for identifying biologically releveant signals in high throughput genomic data, especially when significant differential expression between phenotypes are variability based.

2.4 Biological insights derived from modeling variability

As discussed previously, Alemu et al. followed a statistical approach to analyze EV accounting for batch effect, multi-experiment comparisons and expression level

of a gene. Their analysis provided a number of useful biological insights.

Overall, EV showed positive correlation with gene size associated features (gene length, number of exons, longest transcript length), number of TF clusters cis elements represented in promoter, miRNA, and disease genes. Other features such as number of transcripts, CG ratio in 2kb promoter, DHS (hypersensitivity), number of interacting partners and inherent disorder showed a negative correlation. While the positive correlation between gene length related features may be attributed to the higher noise level expected with longer transcripts (length bias has already been shown to affect differential expression [56]), further analysis carried out by the authors has shown that this is not the case. The number of transcripts for a gene was found to be negatively correlated with EV. To further verify this, the top, bottom, and middle level genes with respect to variability were selected for each tissue and Wilcoxon tests were used to compare the selected genes with the background. This revealed that the length related features exhibited a non-monotonic relationship with EV. Genes with top and bottom levels of variability (hyper-variable and hypo-variable respectively) have shorter genes and transcripts and fewer exons relative to the background genes.

With regard to features associated with the proximal (2kb) promoter, evolutionary conservation, GC fraction and number of nonredundant regulatory motifs were tested. This has shown that EV is probably not an evolutionarily conserved property, since it was not significantly associated with evolutionary conservation. Genes with GC-rich promoters were more likely to show less EV (hypo-variable), which is consistent with the fact that GC-rich promoters correspond to genes that are

usually expressed and are usually related to house-keeping functions. These genes tend to have constrained levels of expression due to greater number of interactions and other homeostasis requirements.

For assessing the relationship between EV and epigenetic features, DHS (DNase hypersensitive sites) and three histone modifications were selected (H3K4me3, H3K27me3, H3K36me3 - these histone modifications have been shown to be associated with expression [7, 8]). H3K4me3 was shown to have a positive correlation with EV, while the other features showed a negative correlation. Since the expected EV was quantified as a function of average expression, this finding is not biased by the absolute expression level. This suggests that transcriptionally active loci tend to have lower EV.

For characterizing the relationship between chromatin state and EV, promoters characterized as strong or weak according to ChromHMM were selected. Comparing genes with high or low EV (hyper-variable and hypo-variable respectively) that were annotated strong or weak relative to the background using a Fisher exact test showed that genes with high EV are under weak regulatory control (i.e. significantly depleted for strong promoters) and vice versa. Further, a linear regression analysis revealed that in 19 tissues, EV was negatively correlated with inherent disorder score. This is expected, since inherently disordered proteins lack its own stable structure and their structure is highly dependent on their partners. An analysis of 9692 genes known to be related to human diseases revealed that EV was positively correlated with disease for 32 out of 41 tissues. Since EV is related to phenotypical plasticity which is a molecular basis in many diseases, this relationship is expected.

Alemu et al. provide the following reasons for considering EV as an inherent property of a gene: it has been previously shown through eQTL analysis that the amount of EV accounted for by genetic variability is less than 5% [63], and intra-population EV remains similar even when inter-population EV is significantly different [39]. Further, high levels of EV are rare even when consideration is limited to tissues it is expressed in. Finally, gene regulation has been eliminated as a case of EV in general since genes encoding transcription factor regulators have not been shown to be associated with high levels of EV. These factors, as well as the relatedness of EV across various contexts and the heritability of EV together support the idea that EV is an inherent property of a gene encoded wither in the genome or heritable epigenome.

From their analysis Alemu et al. arrive at some general conclusions. From the evidence derived through analyzing correlation of EV with epigenetic features they conclude that the EV of a gene is partially encoded by the its epigenomic context. Since genes with GC-rich promoters tended to be correlated with hypo-variability, promoters that elicit high expression are more likely to transcribe genes which demonstrate low EV. In addition, discovery genes with low EV becomes less likely in regions of high degrees of structural polymorphism, and in turn such genes are more likely to encode inherently disordered proteins. Another conclusion is that genes with high EV are more likely to be associated with diseases.

The authors also state two general characterizations of genes regarding low EV and high EV: that the former are more likely to be involved in house-keeping functions while the latter are more likely to be involved in extracellular processes

(i.e. signaling and response functions).

The studies carried out by Irizzary and Feinberg, and by Hansen et al. have provided significant insights into how gene EV and methylation EV can be characterized between tumor and normal samples. Almost all the cDMRs (cancer-specific DNA methylated regions) tested by Hansen et al. showed alteration among cancer, where increased stochastic variation was noticed for the methylation levels. This suggested a disruption of the mechanisms that maintain epigenetic integrity.

Hansen et al provide a comprehensive set of evidence for the existence of differentially methylated regions among multiple cancer types. It has been observed that DNA methylation at CpG dinucleotides demonstrate hypo-methylation and hypermethylation in cancer for certain genes. However, while many studies on cancer epigenetics focus on high density CpG islands and gene promoters, the research by Hansen et al includes regions with lower CpG density as well. These lower density regions are named 'shores', in contrast with the high CpG density regions referred to as 'islands'. The study reveals that these shores strongly correspond to the same regions that show methylation variation among tissue-specific DNA Methylated Regions(DMRs) - i.e. regions where a substantial variation of methylation levels exist between different types of tissue.

While a difference in mean methylation levels were observed between the normal and cancer samples, the increase in across-sample variability of the cancer samples with respect to the variability among the normal samples was even more striking. A binomial distribution model of methylation level distributions was used to account for the increased level of variability that could be expected with the

mean shifts. While the CpG sites had been initially selected for differences in mean methylation among colon cancer, the increased variability was a novel observation. Of these CpG sites, 157 showed significant hyper-variability among all the cancer types tested. These sites belong to not only islands, but shores and regions much further from islands as well. Increased cellular heterogeneity and the age of the patients were accounted for as possible artifactual causes for the hyper-variability, and ruled out.

Another possible explanation for the increased variability is that the selected CpG sites naturally demonstrate increased variability of cytosine methylation levels. To verify that this observation was linked to cancer, a different publicly available methylation dataset was used as a control. Comparing colorectal cancer to matched normal mucosa on the Illumina Human Methylation 27K BeadChip array demonstrated that 81% of the sites showed hyper-variability among in the custom array, while only 42% contained a significant increased of variability in the control. Thus the methylation stochasticity can be considered a general property of cancer.

As mentioned previously, to further observe the behavior of methylation levels among the lower density CpG regions which are not usually examined by array based method, shotgun bisulfite genome sequencing was performed on three colorectal cancers and matched normal colonic mucosa. This revealed the presence of large contiguous blocks of hypo-methylation in cancer compared to normals. 13540 Such regions of lengths between 5kb to 10Mb were identified.

In totality the hypo-methylated blocks accounted for more than half of the genome. A small fraction of hyper-methylated regions were found as well. Thus the

predominant change in block methylation was a loss in methylation levels, where the mean of 73% among all samples decreased to levels between 50-61% in cancer. In addition, it was noted that these blocks were common across all three cancer samples. From the 157 CpG sites that were identified as being hyper-variable across all cancer types tested, 63% of the hypo-methylated CpGs were within hypo-methylated blocks and 37% of the hyper-methylated CpGs were within the above mentioned hyper-methylated blocks.

The study further observed the relationship between the hyper-variable methylation regions and gene expression levels. It is generally observed that there is an inverse relationship between gene expression levels and methylation. Publicly available microarray gene expression data of colon normal and cancer samples were used. Testing for the expected inverse relationship between gene expression and methylation for each category of small DMRs, the strongest relationship was observed for the hypo-methylated shores caused by methylation boundary shifts. A subsequent gene ontology enrichment analysis for differentially expressed genes categorized by their relation to the small DMRs revealed a number of strongly enriched categories for mitosis, and enrichment for a number of cell-cycle related genes.

Another striking observation from the study was that the degree of hyper-variability between the adenoma samples and the cancer samples. When projected to a lower dimensional space using PCA, the normal samples clustered tightly together with the cancer samples dispersed, and the adenoma samples demonstrated an intermediate degree of variability and an intermediate distance to the normal cluster. Out of the 30 adenoma samples used in the custom array, two samples,

a premalignant colon adenoma with a small methylation based distance from the normals, and another adenoma with a large distance, were used for whole-genome bisulfite sequencing. Using average block-level methylations to compute pairwise Euclidean distances between the samples confirmed the existence of genome wide hyper-variability among adenomas, with the degree of variability being less than that of cancer.

Substantial enrichment of genes with increased expression variability in cancer was found, in comparison to the normal samples when tested for genes in the hypo methylated blocks. The possibility of high cellular heterogeneity as the driving cause behind this observation was ruled out, revealing a statistically significant association between the hyper-variability of gene expression and the placement of the gene within a hypo-methylated block. 52% Of the genes with the largest increase in expression variability were inside these blocks, whereas only 17% could be expected by chance. 25 Out of these 26 genes (out of a total of 50), showed stochastic levels of expression among the cancer samples, while they were never expressed among the normal samples.

These observations suggest increased epigenetic plasticity caused by a varying environment where the cancer cell grows in an environment where it is subjected to a number of variable forces such as varying oxygen tension or through metastasis to a distant site. The results obtained from this study has far reaching conclusions, which we will explore subsequently. From a methodology perspective, this study stands as a useful demonstration of carrying out differential EV analysis, where EV that can be explained due to causes other than the disease in consideration are

eliminated prior to reaching conclusions.

Based on these findings, Corrada-Bravo et al [11] introduce anti profiles as a stable method for screening multiple types of cancer. The principle underlying this model of cancer screening is that certain genes will consistently show higher across samples variability among cancer samples as compared with normal samples. In this study these genes are identified and the hyper-variability is used to predict outcome, where the model is referred to as an anti-profile as it measures variation from normal behavior. The observation by Hansen et al that there are consistently hypo-methylated regions of varying lengths (between 5kb and 10Mb) is utilized here for the selection of genes. The same study also demonstrated that the genes corresponding to epigenetic hyper-variability in cancer are also generally tissue-specific genes, an observation which is utilized to develop a universal anti profile.

Using two independent colon cancer studies, training and testing sets, both containing normal and cancer samples were constructed. Genes that lie inside the differentially methylated blocks were considered to select those genes showing consistent hyper-variability among cancers compared to normals. Using the anti profile scoring method with these genes resulted in AUC (Area Under ROC Curve) values of 0.94 and 1.0 respectively for the two datasets. In addition, it was noted that the normal ranges of expression calculated for two datasets were quite consistent.

By compiling a large cancer dataset consisting of cancer and normal microarray samples from many types of tumors, the study developed a universal cancer anti-profile by restricting the anti profile to tissue specific genes; tissue-specific genes here were considered to be genes which were expressed among at least 95% of samples for

at most three tissues. This anti-profile was verified via cross validation analysis to provide high AUC values. This research forms a significant part of the background for the anti-profile analysis carried out in this thesis (see Chapter 3), where we use the probesets included in the universal anti profile for our applications of predicting and classification of tumors according to relapse risk and progression level.

2.5 Conclusion

We have examined a number of studies which have carried out estimation and differential analysis with gene expression level variability. In addition we have presented how the same ideas can be extended for methylation level analysis as well.

These studies, and the important biological insights gained from them, demonstrate that differential variability analysis has an important role to play in biostatistical research. Further they suggest that complementing an expression level analysis focusing on mean differences with a parallel analysis examining variance differences should be an important consideration in many studies of genomic and epigenomic level gene expression or methylation studies.

Variability based analysis is a complicated task. However the studies we have discussed demonstrate how to overcome some of the potential challenges to derive informative conclusions. For example, there are many factors that influence gene EV, and generally an analysis should be carried out to rule out potentially infringing causes of variability other than the hypothesized cause.

The research presented here by Hansen et al., and followed upon by Corrada-

Bravo et al., form an important part of the background research for this thesis. We discuss the anti-profile method in more detail and examine it's further utilization for cancer prognosis in the next chapter.

Chapter 3: Anti-Profiles for Cancer Progression Analysis

3.1 Background

Despite many advances in cancer treatment, early detection remains the most promising avenue in terms of patient survival [4, 43, 52, 72, 84]. While there have been many attempts at devising early cancer screening techniques, many approaches remain inefficient in clinical settings, or are not pragmatic due to lack of cost effectiveness or due to requirement of invasive procedures [31,36,38]. Genomic screening techniques are a promising approach in this area. Continuing advances in high-throughput technology make these approaches both cost and time effective. Certain types of genomic data, such as gene expression derived from peripheral blood are minimally invasive as well.

The main difficulty in developing such techniques has been the lack of stable markers in cancer gene expression profiles. Apart from a few exceptions [24, 80], many gene signatures have failed to reproduce their results when tested on independently obtained data [37], indicating that the signature is not adequately robust to be deployed in a clinical setting.

However, recent work by Corrada-Bravo et al. [11] demonstrates that by modeling consistent increased gene expression variability across cancer types, a statistical

model can be developed which provides a stable and robust predictor of cancer that works well across multiple cancer types. The underlying observation behind this approach is that certain genes will consistently show higher across sample variability in cancer as compared to normal samples. Hyper-variability in these genes can be leveraged to measure deviation from a stable profile of normal expression, resulting in a cancer anti-profile.

Here we further advance this approach by demonstrating that it can also be used as a predictor of survival or relapse. We demonstrate that using genes that show consistent, or universal, hyper-variability across cancer types, their degree of deviation in gene expression from normal tissue can be used as a measurement of potential malignancy (measured as risk of relapse or death). The results indicate that the anti profile approach can be used as a more robust and stable indicator of tumor malignancy than traditional classification approaches.

3.2 Methods

We extend the observation of consistent hyper-variability in cancer with respect to the normal samples to include tumor progression. Our hypothesis here was that the degree of hyper-variability as measured with respect to the normal samples would increase with tumor progression.

Corrada-Bravo et al. demonstrate how to derive a colon-cancer anti-profile for screening colon tumors by measuring deviation from normal colon samples in [11]. Briefly, to create an anti-profile a set of normal samples and tumor samples are se-

lected; probsets are then ranked by the quantity $\frac{\sigma_{j,tumor}}{\sigma_{j,normal}}$ (where $\sigma_{j,tumor}$ and $\sigma_{j,normal}$ are the standard deviations among the tumor samples and normal samples respectively for probeset j) in descending order and a certain number of probesets(typically 100) with the highest value are selected. Normal regions of expression are calculated for each probeset, and an anti-profile score for a sample is calculated by counting the number of probesets for which expression lie outside the normal region.

We used a number of publicly available microarray datasets and one methylation dataset. The microarray datasets were either Affymetrix Human Genome U133 Plus 2.0 (GPL570) or Affymetrix Human Genome U133A (GPL96). The raw data were collected and processed using fRMA normalization [49] and the barcode algorithm [50] to obtain z-scores. A detailed description of the datasets used and the selection of samples can be found in the supplementary file.

For anti-profile analysis, a variance ratio statistic across multiple tissues is calculated [11]. This statistic is computed as $u_g = log_2 \frac{mean_c s_{gc}}{mean_t s_{gt}}$ where s_{gc} and s_{gt} are the standard deviations of cancer type c and tissue t, and g being the probeset. Probesets are ranked according to u_g. We used the 100 probesets with the highest u_g values for our experiments with anti-profile scores. This calculation was available for the GPL570 platform by Corrada-Bravo et al., and we used a number of cancer and normal samples of many tissue types from GPL96 microarray platform experiments to obtain a similar universal set of probesets for our GPL96 experiments.

The normal regions of expression are calculated based on median and median absolute deviation(mad) statistics as $m_g \pm 5 \times mad_g$ for a probeset g. Here $m_g = median_t(median_{gt})$ for tissue t and $mad_g = median_t(mad_{gt})$. For most of our

experiments our computations were limited to a single tissue type(colon, breast, lung, thyroid or adrenocortical).

The normalized expression values and the selected probesets can be supplied to the *apCount* method of the *AntiProfiles* Bioconductor package [10], which counts the number of probesets for which the expression of the given tumor sample lies outside the normal range of expression. This count is used as the anti-profile score. The *AntiProfileStats* object from the package and the *buildAntiProfile* method were used to compute and use the universal anti-profile signature.

For comparing the anti-profile scoring method against classifiers that do not take into account the hyper-variability of cancer, we compared our approach with PAM, a popular shrunken centroid classifier [77]. PAM extends the regular centroid based classification by computing a standardized centroid for each class. The shrunken centroid represents the class using the average gene expression of that class divided by the within-class standard deviation for that gene. The amount of shrinkage is determined via a threshold parameter which affects the classification by reducing the effects of features that are determined to be noisy.

For shrunken centroid classifications, we used *pamr*, an R package which implements the PAM algorithm. The methods *pamr.train, pamr.cv, pamr.predict* were used for training, cross validation and testing respectively. For any given dataset, a binary classification was attached to the data (either a high risk vs. low risk classification based on patient survival information or a carcinoma vs. adenoma classification based on tumor progression information) which was used for fitting the PAM model, and after cross validation on the training dataset, the threshold

parameter was selected to minimize both training error and the number of probesets used for classification.

3.3 Results and discussion

Anti-profiles capture tumor progression

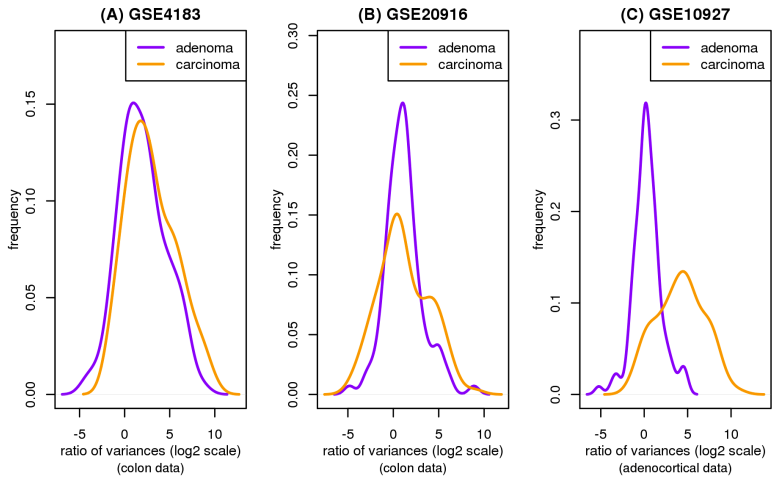

Figure 3.1: **Among probes that exhibit higher variability among cancers than among normals, degree of hyper-variability observed is related to level of progression.**
(A)Distribution of variance ratio statistic($log_2[\sigma^2_{tumor} \div \sigma^2_{normal}]$) for colon dataset(Gyorffy et al.; GSE4183) from anti-profile computed using another colon dataset(Skrzypczak et al.; GSE20916). (B)Distribution of variance ratio statistic for Skrzypczak et al. colon dataset from anti-profile computed using Gyorffy et al. colon dataset. (C)Distribution of variance ratio statistic for adrenocortical data(Giordano et al.; GSE10927) for universal anti-profile probesets.

In our experiments we first extended the anti-profile approach by using colon-cancer anti-profiles for differentiation between tumors according to their progression level. To test our hypothesis we obtained two publicly available microarray datasets with normal, adenoma and cancer colon samples [29, 65].

Based on the finding that consistent decreases in methylation are observed

along large genomic blocks [30], probesets were selected in [11] by selecting genes that lie inside such blocks to create a colon cancer anti-profile. From those probesets we selected the 100 probesets that showed most hyper-variability among cancer samples in comparison to the normal samples. We then plotted the distribution of variance of cancer/adenoma samples to variance of normal samples ratio (in $\log_2$ scale) for these probesets on the other dataset (Figures 3.1A, 3.1B). Both adenomas and cancers show higher variability than normals(region to the right of $x = 0$) while cancers show higher hyper-variability than adenomas. This demonstrates that the hyper-variability property is a stable marker between experimental datasets, and it validates our hypothesis that the hyper-variability measurement can be extended to model tumor progression. A Kolmogorov-Smirnov test between the two distributions with the alternative hypothesis being that the distribution for adenoma samples is less than that of cancer samples yields $p = 1$ for the first dataset. For the second dataset we see that the distribution for cancer samples exhibit both higher and lower degrees of hyper-variability compared to the adenomas. Nevertheless, further experiments with anti-profile scoring demonstrate that these probesets can be useful for differentiating between adenoma and cancer samples (see below).

Next we performed the same analysis using the universal anti-profile signature computed in [11]. We obtained an adrenocortical microarray dataset [25] containing normal, adenoma and cancer samples. For the most hyper-variable 100 probesets from the universal anti-profile, we plotted the distribution of variance of cancer/adenoma samples to variance of normal samples ratio (Figure 3.1C). The same observations as before could be seen: a majority of these probesets show greater

variability among cancer and adenoma samples than among normal samples, and this degree of variability is higher among cancer samples in comparison with adenoma samples (A Kolmogorov-Smirnov test between the two distributions with the alternative hypothesis being that the distribution for adenoma samples is less than that of cancer samples yields $p = 1$). This extends the validation of our hypothesis regarding hyper-variability and tumor progression level to the universal anti-profile signature.

In the next stage of our experiments we applied the anti-profile scoring method. As discussed in [11], the anti-profile scoring method counts the number of hyper variable probesets for which the expression of tumor samples lie beyond the normal region of expression. It has been shown to be an effective measurement in differentiating between tumor samples and normal samples, and our aim was to apply the same scoring method for differentiating between different stages of tumor progression. With the two colon cancer datasets used to derive colon-cancer anti-profiles, we used the hyper-variable probesets and the normal regions of expression for those probesets derived from one dataset to calculate anti-profile scores for the normal, adenoma and cancer samples in the other dataset. The distribution of these scores are plotted in Figures 3.2A and 3.2B: for both datasets, the average anti-profile score increases from the normal group to the adenoma group to the cancer group: for the first dataset [29], the mean scores for these groups are 18.88, 27.93 and 35.33 respectively and for the second dataset [65] the respective mean scores are 32.2, 51.4 and 58.9. Comparing the adenoma scores against the cancer scores yield an area under the receiver operating characteristic curve(AUC) of 0.711 and a p-value of

0.05 from the Wilcoxon rank-sum test for the first dataset; the same comparison for the second dataset yields an AUC value of 0.97 and a p-value $< 10^3$ from the Wilcoxon rank-sum test.

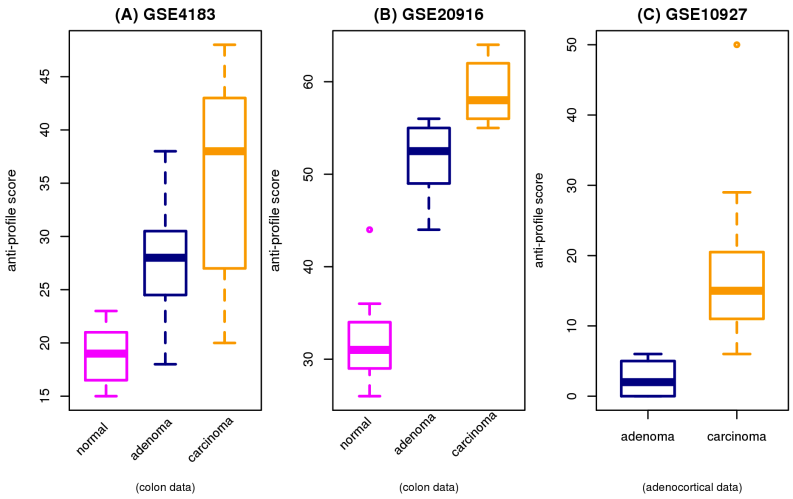

Figure 3.2: **Anti-profile scores calculated for tumors and normals.** (A)Anti-profile scores for colon dataset(Gyorffy et al.) from anti-profile computed using another colon dataset(Skrzypczak et al.); Carcinoma vs. adenoma area under ROC(AUC) = 0.711, Wilcoxon rank-sum test p-value = 0.05. (B)Anti-profile scores for Skrzypczak et al. colon dataset from anti-profile computed using Gyorffy et al. colon dataset; Carcinoma vs. adenoma AUC = 0.97, Wilcoxon rank-sum test p-value $< 10^{-3}$. (C)Anti-profile scores for adrenocortical data(Giordano et al.) from universal anti-profile probesets; AUC = 0.997, Wilcoxon rank-sum test p-value $< 10^{-4}$.

Similarly we applied the anti-profile scoring method for the adrenocortical dataset with the universal anti-profile probesets(Figure 3.2C). The cancer samples have higher anti-profile scores than the adenoma samples: the mean anti-profile scores are 2.5 and 16.84 for the adenoma and cancer groups respectively. The comparison of the two score groups give an AUC value of 0.997 and a Wilcoxon rank-sum test p-value $< 10^4$. In addition we also performed the same experiment on 10 Follicular Thryoid adenomas and 13 Follicular Thyroid carcinomas obtained

from GSE27155 [26], where we used the 100 most hyper-variable probes from a universal anti-profile signature for the GPL96 platform(see 'Application to Breast cancer' section below). This provided an AUC of 0.808 and a Wilcoxon rank-sum test p-value of 0.01. However, only 4 normal samples were available in this dataset, thus limiting the confidence in the experimental result.

Increased expression variability is associated with clinical outcome in colon, lung and breast cancer

Based on the observation that increased expression variability and anti-profile scores in probesets with hyper-variable expression is associated with tumor progression, we hypothesized that it will also be associated with clinical outcomes for tumors: aggressive tumors exhibiting poor clinical outcome would be associated with increased hyper-variability in these specific genes and vice versa.

Application to Colon cancer

We first experimented with a colon-cancer anti-profile as discussed in the previous section. We obtained a microarray dataset of colon tumor samples supplied with survival information(indicating the relapse of a patient within a certain number of years) [47]. Using the other colon cancer microarray datasets used in the previous section [29,65], we constructed a colon-cancer anti-profile using the normal and cancer samples, limiting the probesets to the colon cancer hyper-variable genes from [11]. A set of 100 probesets with the highest variability among cancer samples

with respect to normal samples were selected from this anti-profile. These probesets and the normal regions of expression calculated for them using the normal colon samples from the above mentioned datasets together constituted the colon cancer anti-profile used.

We stratified the samples into high risk and low risk as follows: patients that relapsed within 1 year after diagnosis were classified as high risk, and those that did not relapse within 1 year were classified as low risk. For the selected probesets we calculated the distribution of variance of high/low risk samples to variance of normal samples ratio (Supplementary Figure 3.6A). The hyper-variability of the colon cancer anti-profile probesets is reflected in these results, given that the majority of the probesets have a $\log_2$ variance ratio > 0. We can also see that the high risk samples exhibit slightly higher variability than the low risk samples when compared against the normals, affirming that the hyper-variability observation extends to tumor prognosis as well.

Further, we calculated anti-profile scores for the colon tumor samples. Since the high risk and low risk grouping is not a well defined classification that only tentatively captures tumor progression, we used Kaplan-Meier survival curves to measure the effectiveness of the anti-profile scores. We ordered the tumor samples according to the anti-profile score and stratified them to three equal sized groups and observed the rate of survival in each group using Kaplan-Meier curves. This demonstrated to us that the anti-profile scores clearly correlate with the prognosis of the tumors, with higher anti-profile scores showing higher chances of relapse and vice versa (Figure 3.3A; logrank test score 9.452, p-value 0.008). We also noticed

that the high risk samples have a higher score on average than the low risk samples (Supplementary Figure 3.6B) : mean anti-profile scores for low risk and high risk groups are 41.85 and 48.87 respectively (Wilcoxon rank-sum test p-value of 0.0078 between the two groups).

Application to Lung cancer

Next we applied the anti-profile method to analyze lung cancer survival. Here we tested the universal anti-profile from [11] with two microarray lung cancer datasets containing patient survival information based on patient relapse; the primary dataset containing both normal and tumor samples [55], the second containing only tumor samples [9].

As with the colon dataset, we stratified the samples into high risk and low risk based on patient relapse within 5 years. For the universal anti-profile probesets, we plotted the distribution of variance of high and low risk samples to variance of normal samples ratio (Supplementary Figure 3.7). The majority of the universal anti-profile probesets show higher variability among the tumor samples than the normal samples, indicating that the universal anti-profile manages to capture the hyper-variability property of these datasets.

We used normal lung samples to calculate anti-profile scores for the tumor samples for both datasets. Ordering the tumor samples by anti-profile score, for each dataset we stratified them to three equal sized groups and plotted Kaplan-Meier survival curves (Figure 3.3B). For the first dataset, the tumor samples with the

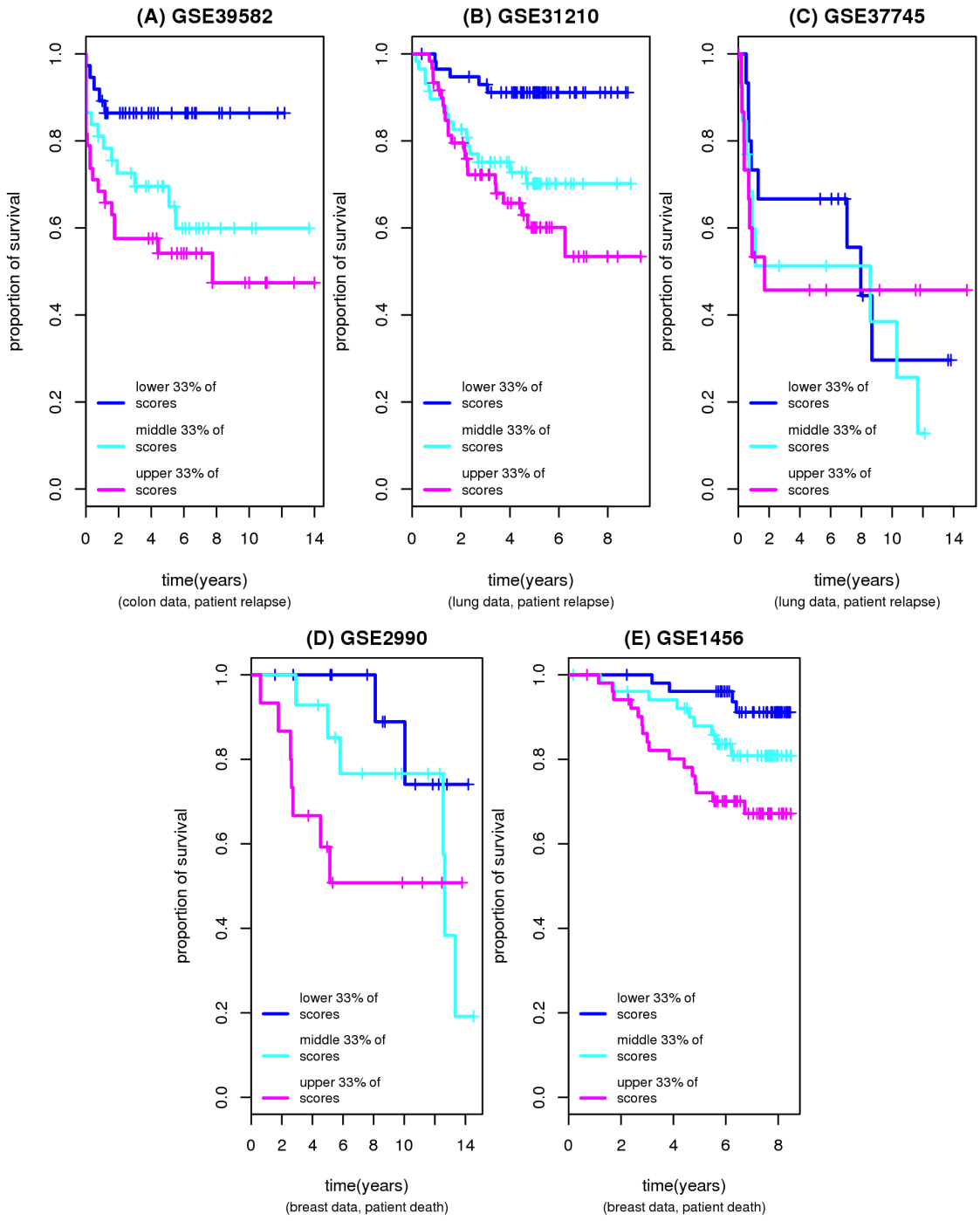

Figure 3.3: **Anti-profile scores correspond to tumor prognosis.** (A)Kaplan-Meier survival curves for colon tumor samples ranked by anti-profile scores are grouped to three equal sized groups; Logrank test score 9.452, p-value 0.008. (B)Survival curves for first lung dataset(Okayama et al.) - samples ranked by anti-profile scores are grouped to three equal sized groups; Logrank test score 15.44, p-value $< 10^{-4}$. (C)Survival curves for second lung dataset(Botling et al.); Logrank test score 0.611, p-value 0.73. (D)Survival curves for first breast dataset(Sotiriou et al.); Logrank test score 3.971, p-value 0.137. (E)Survival curves for second breast dataset(Pawitan et al.); Logrank test score 10.467, p-value 0.005.

highest anti-profile scores show greatest relapse among the three groups, while the tumor samples with lowest scores show the least relapse (logrank statistic for the first dataset 15.44, p-value $< 10^3$). For the second dataset we obtained a logrank statistic 0.611 and a p-value 0.73 from the same procedure. The distribution of the scores for the high risk and low risk samples (Supplementary Figure 3.8) indicate that for both datasets the low risk samples have lower anti-profile scores and vice versa (for the first dataset, mean score for low risk and high risk groups are respectively 14.32 and 19.81 with a Wilcoxon rank-sum test p-value $< 10^3$; for the second dataset, mean score for low risk and high risk groups are respectively 31.64 and 33.32 with a Wilcoxon rank-sum test p-value of 0.58).

The results obtained for the second lung dataset did not show a thorough separation in the Kaplan-Meier survival curves when sorted into 3 groups. Comparing the generalized normalized unscaled standard error (GNUSE) values, a standard metric of microarray quality, to compare the quality of the microarray data for the two lung cancer datasets (Supplementary Figure 3.9), we noticed that this second dataset has a higher GNUSE value distribution in comparison to the first dataset, which might explain the poor performance. However stratifying the samples as top 50% of scores and lower 50% of scores did show some separation of the two groups in terms of survival(logrank statistic 1.418, p-value 0.22). The survival curves show the expected survival difference between the groups for the first 8 years, suggesting that the prognosis predicted by the anti-profile scores may become less relevant over time. In addition with increasing age, there is increased possibility that the health of a patient may deteriorate more aggressively. Further, it should be noted here

that the normal samples came from the first dataset (a Japanese cohort), whereas the second dataset only contained tumors (a Swedish cohort). In an ideal scenario we would compute a population dependent normal profile.

The first dataset also contained information about death of patients. A similar analysis as before with patient death instead of relapse showed a logrank statistic of 8.342 with p-value 0.015 when the samples were ranked by anti-profile scores and stratified to 2 groups (Supplementary Figure 3.10).

These results demonstrate that the universal anti-profile probesets can be used to explore the hyper-variability in lung microarray data and further validate the use of using deviation from normal samples as a measurement of tumor prognosis.

Application to Breast cancer

We next applied the methodology to breast cancer microarray data on Affymetrix Human Genome U133A platform (GPL96). Since the universal anti-profile signature had been derived from Affymetrix Human Genome U133 Plus 2.0 (GPL570) microarray data, we used a number of publicly available GPL96 platform cancer and normal samples (1207 cancer samples and 773 normal samples) of multiple tissue types to re-calculate an anti-profile signature for the GPL96 platform(see 'Methods'). We used the most significant 100 probesets from this signature for our breast cancer anti-profile experiments.

After obtaining two publicly available breast cancer microarray datasets [51, 58], we selected lymph node negative and ER positive samples and verified that

these probesets were able to capture the hyper-variability of cancer samples (Supplementary Figure 3.11). Since relapse information was not available for majority of the samples, we used death within 5 years as our criteria for obtaining a high risk - low risk classification.

We collected breast normal samples from publicly available datasets and calculated anti-profile scores for the two datasets. We drew Kaplan-Meier survival curves by ranking the samples by score and grouping them to 3 equal sized classes (Figure 3.3C). Similar to our observation with colon and lung cancer data, the anti-profile scores showed a correlation with survival of patients(logrank statistic for the first dataset 3.971, p-value 0.137, logrank statistic for the second dataset 10.467, p-value 0.005). The distribution of the scores for the high risk and low risk samples (Supplementary Figure 3.12) demonstrate that high risk samples have higher scores on average, and vice versa (for the first dataset, mean score for low risk and high risk groups are respectively 10.13 and 17.12 with a Wilcoxon rank-sum test p-value of 0.0061; for the second dataset, mean score for low risk and high risk groups are respectively 11.41 and 16.91 with a Wilcoxon rank-sum test p-value $< 10^3$).

The second breast dataset also contained information about patient relapse. Performing a similar analysis using relapse instead of death provided a logrank statistic of 10.755 (p-value 0.004) when the samples were grouped by anti-profile score (Supplementary Figure 3.13).

In addition Supplementary Figure 3.14 show similar results obtained for a third breast cancer dataset with patient death information. With only 9 deaths being recorded, our method of stratifying samples into high risk and low risk classes

was not appropriate for this dataset. However we observed a trend of samples with high anti-profile scores exhibiting a higher rate of relapse and vice versa, as with the other datasets.

In summary, these results obtained for lung and breast cancer data further show the utility of the anti-profile approach as a robust and effective method for modeling tumor prognosis, and validates our hypothesis that deviation from the normal group can be considered a measure of the risk level associated with a tumor.

The anti-profile approach is more stable than standard classification methods

We compared the anti-profile method with PAM using lung cancer data. For this, using the high risk and low risk stratification of samples previously described, we constructed a binary classification problem between low and high risk, and trained the PAM classifier on one dataset and tested the classifier on the other dataset. We used cross validation on the training dataset to determine the threshold parameter which minimizes the misclassification error on the training data. The same experiment was performed between the two breast cancer datasets, and also the two colon cancer datasets used for our analysis based on tumor progression (here the adenoma/carcinoma status was used as the binary stratification). The posterior probabilities obtained for the testing dataset were used to calculate area under the ROC curve values and Wilcoxon rank-sum test p-values.

To compare against this, we applied the anti-profile method to the same train-

Tested Dataset	**Training Dataset**	**Anti-Profile scores**				**PAM**	
		Training set hypervariable probesets		*Universal hypervariable probesets*			
		AUC	**Wilcoxon p-value**	**AUC**	**Wilcoxon p-value**	**AUC**	**Wilcoxon**
Lung1	Lung2	0.739	0.00002	0.716	0.0001	0.66	0.004
Lung2	Lung1	0.44	0.571	0.558	0.584	0.56	0.55
Breast1	Breast2	0.712	0.08	0.788	0.017	0.831	0.004
Breast2	Breast1	0.707	0.0021	0.729	0.0006	0.719	0.001
Colon1	Colon2	0.8	0.0054	0.455	0.692	0.603	0.339
Colon2	Colon1	0.97	0.00042	0.915	0.0018	0.92	0.0007

Table 3.1: **Comparison of prediction results obtained using the anti-profile scoring method and PAM.** For each tissue type of lung, breast and colon, two datasets with tumor samples were obtained and both the anti-profile method and the PAM model were fitted on one dataset and tested on the other dataset. For a binary stratification of samples by risk level, the area under the ROC curve(AUC) and the p-value from the Wilcoxon rank-sum test were calculated from the decision values resulting from each method. Datasets used are: *Lung1*(GSE31210), *Lung2*(GSE37745), *Breast1*(GSE2990), *Breast2*(GSE1456), *Colon1*(GSE4183) and *Colon2*(GSE15960).

ing and testing dataset pairs. For each tissue type we used normal and the tumors of one dataset to select 100 probesets and calculate anti-profile scores for the other dataset. In addition we also calculated the anti-profile scores using the universally hypervariable probesets (100 probesets). A comparison of these results can be seen in Table 1.

From the comparison of area under the ROC curves and the Wilcoxon rank-

sum test p-values, we see that while the shrunken centroid classifier performed at a similar level with the breast cancer datasets, it did not perform as well as the anti-profile method with the other datasets. Particularly with the lung cancer data, in comparison with the anti-profile scoring the PAM classifications failed to obtain high levels of differentiation between the high risk and low risk samples of the testing data. These results show that taking into account the stochastic hyper-variability of cancers with regard to normals can produce more stable classifiers when building predictors across datasets.

Anti-profiles based on DNA methylation also capture tumor progression

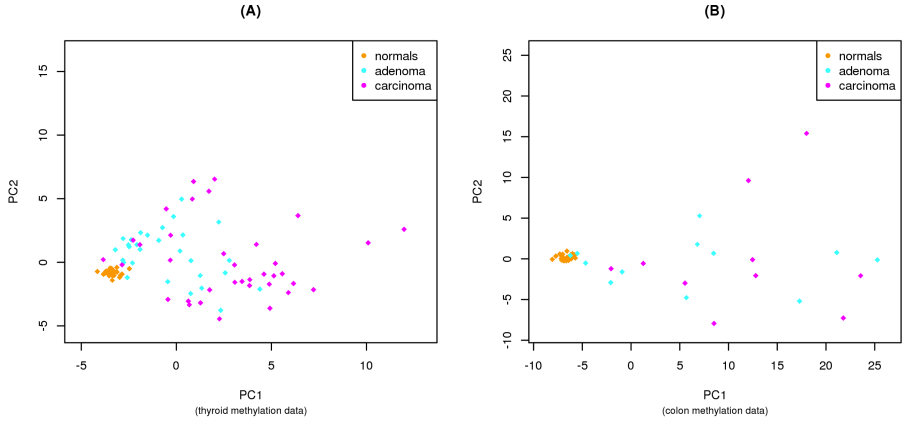

Figure 3.4: **Anti-profiles applied to methylation data.**
(A)First two principal components of thyroid methylation data. (B)First two principal components of colon methylation data.

DNA methylation is one of the primary epigenetic mechanisms by which gene expression is regulated, and is believed to play a particularly important role in cancer. High levels of methylation in promoter regions are usually associated with

low transcription [71], and therefore globally methylation and gene expresison are seen to have an inverse relationship [40]. Abnormal methylation patterns have been observed in cancer, with loss of sharply delimitted methylation levels (in comparison with normal methylation levels) in regions associated with tissue differentiation [30]. This is believed in part to give rise to the aberrant gene expression profiles that are typically associated with cancer. Given these observations, we expected that the anti-profile method would be applicable to methylation data as well.

We applied the anti-profile scoring method to DNA methylation data from thyroid and colon samples [30], where for each tissue type normal, adenoma and cancer samples were available. Using 384 probesets available in their custom Illumina methylation array data, for each cancer type we used the normal samples to define the normal regions of methylation and calculated anti-profile scores by summing the number of features that fell outside the normal methylation region for each cancer sample.

Figure 3.4 shows the distribution of adenoma and carcinoma samples against normal samples on a principal component plot, showing the presence of the hypervariability pattern in methylation data: the normal samples cluster tightly while the adenomas show some dispersion, and the carcinomas show even greater dispersion. Since these behaviors are present for both colon and thyroid data, it again reinforces our notion that the anti-profile approach has wide application for classification in cancer.

Supplementary Figure 3.15 shows the results obtained with the anti-profile scores. As with the gene expression data, the methylation data also shows that

adenomas tend to have lower anti-profile scores than the carcinomas: for the thyroid tumors the median anti-profile score for the adenoma class is 10 while for the carcinoma class it is 17, and for the colon tumors the median score for the adenoma class is 75.5 while for the carcinoma class it is 121.5. We also obtained an Illumina HumanMethylation450 experimental dataset containing DNA methylation levels for normal, adenoma and cancer samples comprised of Thyroid, Breast, Colon, Pancreas and Lung tissues (see Supplementary table S2) [78]. Given the extremely large number of methylation sites available in the dataset, we first used the *minfi* Bioconducter package to aggregate nearby CpG sites and calculate a mean methylation level for each cluster. This produced about 220k clusters, which we then used for our experiment.

For each selected tissue type, we followed a two-stage feature selection process before applying anti-profile scoring. We obtained a list of hypo-methylated blocks for Thyroid, Colon, and Lung cancer from. In the first stage we computed the interesection of hypo-methylated regions as follows: to test Thyroid samples, we computed the intersection of Colon and Lung blocks. For testing Colon samples, we computed the intersection of Thyroid and Lung blocks. For testing Breast and Pancreas samples the intersection of Thyroid, Lung and Colon blocks was computed. For each tissue type, collapsed CpG site clusters lying within the respective intersections were selected. Following this, in the second stage, we selected the normals and cancers of the remaining tissue types and selected the 20 most significant clusters. Using these selected features we built an anti-profile using the normals of the selected tissue and used it to calculate anti-profile scores for the adenoma and

cancer samples of that tissue group. This was repeated for Thyroid, Breast, Colon and Pancreas tissue groups (the Lung tissue group could not be used due to lack of adenoma samples). The results, presented in Supplementary Figure 3.16 show higher anti-profile scores in cancers than in adenomas for all tissues.

The anti-profile approach may be used for prognostic prediction

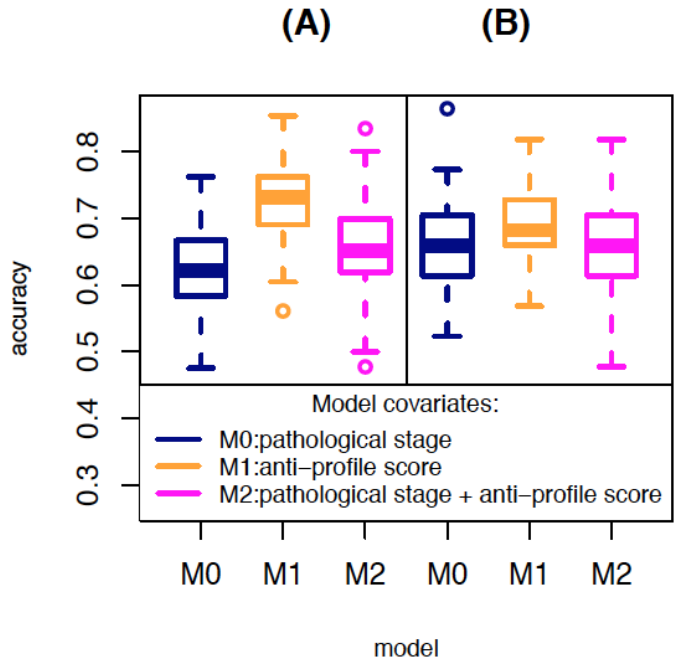

Figure 3.5: **Anti-profiles applied to Cox proportioanl hazard models for survival prediction.**
Cox proportional hazard models with significant clinical covariates and anti-profiles were used to predict patient survival at 5 years for the second breast caner dataset (Pawitan et al.) with (A)patient death and (B)patient relapse. The plots show accuracy of prediction calculated for 100 training and testing subsets randomly selected.

To further examine the prognostic ability of the anti-profile score, we used the anti-profile scores as a covariate for modeling patient survival for some of the datasets. We obtained clinical covariates for the microarray datasets when publicly

available, and fitted each covariate separately to a Cox proportional hazard model to ascertain their prognostic significance. The Cox proportional hazards model is a widely used statistical model for assessing censored survival information [67]. It provides a way for modeling the effect of a particular factor (such as age, severity of disease, etc) on the time taken by a patient to relapse (from the time of entering the clinical trial), or the time at which a patient dies. Here we treated the anti-profile score in the same manner as the other clinical factors.

For the first lung cancer dataset [55], we tested age, sex, smoking status, pathological stage. After fitting each covariate individually to a Cox proportional hazards model (assuming constant covariates) with patient relapse information, only pathological stage provided a p-value < 0.05 from a Wald test. In addition we also fitted the anti-profile score as a covariate, which also yielded $p < 0.05$. Using patient death information instead of relapse, once again both pathological stage and anti-profile score showed significant association with survival (Wald test $p < 0.05$).

For the second breast cancer dataset [51], we tested pathological stage and subtype (Basal, ERBB2, Luminal A, Luminal B, Normal Like) for prognostic relevance with relapse, and found that only pathological stage was significant when fitted to a Cox model (Wald test $p < 0.05$). The anti-profile scores provided $p < 0.05$ as well. Using patient death information instead of survival produced similar results with pathological stage and anti-profile score both showing prognostic significance when fitted independently (Wald test $p < 0.05$).

For the colon cancer dataset with survival information (patient relapse) [47], we tested age, pathological stage, chemotherapy (treatment or lack of it) and location

(distal vs. proximal). Pathological stage and chemotherapy status proved to be significant (Wald test $p < 0.05$, with pathological stage yielding $p < 10^{-5}$) when fitted independently to a Cox model. The anti-profile scores proved to be significant as well ($p < 0.05$).

For these datasets, we also tested the predictive ability of a Cox model fitted with these covariates selected above by predicting whether a given patient would live upto a given time t (we used $t = 5$ years). For this we predicted the survival curve for that patient using a model fitted with a training set and compare the predicted survival curve to the rate of survival for training group patients that did not survival at time t against training group patients that did survive upto time t.

The dataset is split randomly into training(70%) and testing(30%) sets and three Cox models are fitted to the training data: *(a)* a model with only selected clinical covariates, *(b)* a model with only anti-profile scores, and *(c)* a model with both clinical covariates and the anti-profile score. For each model, the mean survival at time t is calculated for training set patients surviving and not-surviving at that time point. For each patient in the testing set, the predicted survival probability at time t is compared to the surviving group mean and not-surviving group mean and the closest group chosen to predict whether the patient will survive or not. These predictions are compared to actual survival to calculate an accuracy rate (patients censored by the time t are not used for the calculation). This process is repeated for a 100 training and testing subsets created from the main dataset, and the distribution of accuracy values were plotted.

For the second breast cancer dataset [51], a Cox model fitted with pathological

stage proved to be less accurate than a model fitted with the anti-profile score (Figure 3.5A) for predicting patient death at 5 years. The mean accuracy level for the model with pathological stage was 0.619, for a model with anti-profile score was 0.726, and for a model with both covariates it was 0.655. A Wilcoxon test between the results from the first and third models yielded $p < 0.005$, showing that the anti-profile score can significantly increase predictive power of a model. A similar test based on patient relapse (Figure 3.5B) showed all three model choices performing at a similar accuracy level, with the anti-profile score based model providing a slightly higher accuracy rate.

We used a similar experiment on the lung and colon cancer datasets mentioned above, but found that adding anti-profile scores to survival models including significant clinical covariates did not improve their performance significantly (Supplementary Figure 3.17).

GO analysis of the universal anti-profile signature

We carried out a gene ontology (GO) enrichment analysis for the 100 most hypervariable probesets used in many of our experiments. We used the *GOstats* package to conduct a hyper-geometric test for these genes and explored the relevant GO terms which yielded $p < 0.001$ from the test (Table 3.2). We observed that a number of terms related to morphogenesis were present in our results, as well as cell differentiation. Other terms included extracellular matrix disassembly; this process of breaking down the extracellular matrix is known to be related to embryonic de-

velopment and tissue remodeling, and in addition is also known for being associated with cancer progression.

	GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
1	GO:0018149	0.00	54.07	0.09	4	30	peptide cross-linking
2	GO:0022617	0.00	14.91	0.39	5	125	extracellular matrix disassembly
3	GO:0007492	0.00	19.20	0.24	4	77	endoderm development
4	GO:0090596	0.00	9.13	0.76	6	245	sensory organ morphogenesis
5	GO:0043062	0.00	7.01	1.17	7	376	extracellular structure organization
6	GO:0060235	0.00	134.38	0.02	2	7	lens induction in camera-type eye
7	GO:0048562	0.00	8.00	0.86	6	278	embryonic organ morphogenesis
8	GO:0048593	0.00	13.85	0.33	4	105	camera-type eye morphogenesis
9	GO:0035987	0.00	25.72	0.13	3	43	endodermal cell differentiation
10	GO:0048705	0.00	8.89	0.64	5	205	skeletal system morphogenesis
11	GO:0045596	0.00	5.09	1.85	8	595	negative regulation of cell differentiation
12	GO:0008593	0.00	20.16	0.17	3	54	regulation of Notch signaling pathway
13	GO:0021879	0.00	20.16	0.17	3	54	forebrain neuron differentiation
14	GO:0001503	0.00	6.35	1.08	6	347	ossification
15	GO:0050910	0.00	55.97	0.04	2	14	detection of mechanical stimulus involved in sensory perception of sound
16	GO:0010629	0.00	3.62	3.70	11	1193	negative regulation of gene expression
17	GO:0048513	0.00	2.82	8.56	18	2761	organ development

Table 3.2: **Gene ontology enrichment analysis.** Results of a hyper geometric test performed for association between universal hypervariable genes and GO terms; terms with p-value < 0.001 from the test have been selected.

Stability of feature selection methods

To ascertian the stability of variance based probeset selection further, we compared feature selection using the t-statistic (as a mean-difference based method) and

the ratio of variances which we have utilized for the anti-profile approach. We used the primary lung microarray dataset and created subsets of it, with each subset containing 50% of high risk samples. For each subset, we computed the t-statistic between the high risk samples and the normals, and also the ratio of high risk variance to normal variance. Following this, we noted which 100 probesets provided the largest mean or variance based statistic. For 10 subsets selected in this way, Figure 3.6 shows the number of probesets that were selected by each method. A similar analysis with the primary breast cancer dataset is seen in Figure 3.7.

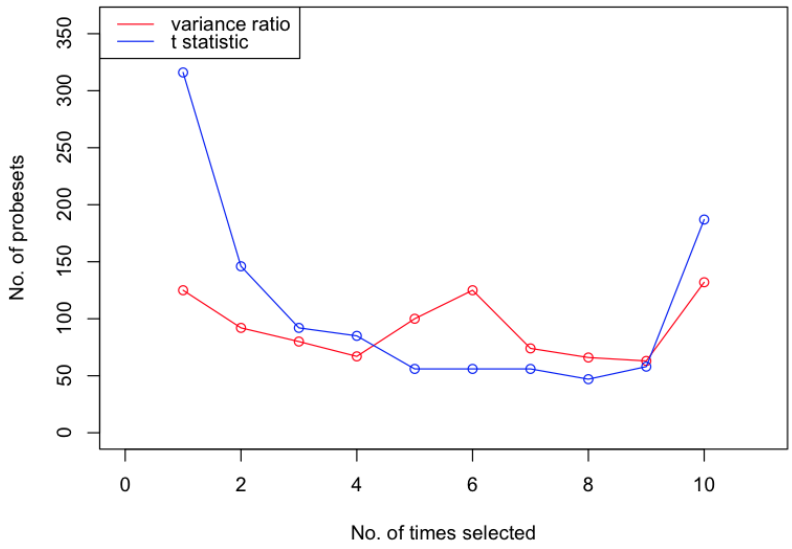

Figure 3.6: **Comparison of feature selection methods:** Probeset selection using the t statistic and the ratio of variances between high risk samples and normal samples from 10 subsets of data for lung microarray data.

Probesets that were selected for only subset do not signify stability in the feature selection method, and we observed that the t-statistic contained a larger share of probesets that were selected only once in comparison with the variance based method. While the t-statistic also showed slightly higher number of probesets that were selected in all the subsets, when we considered the probestes that were

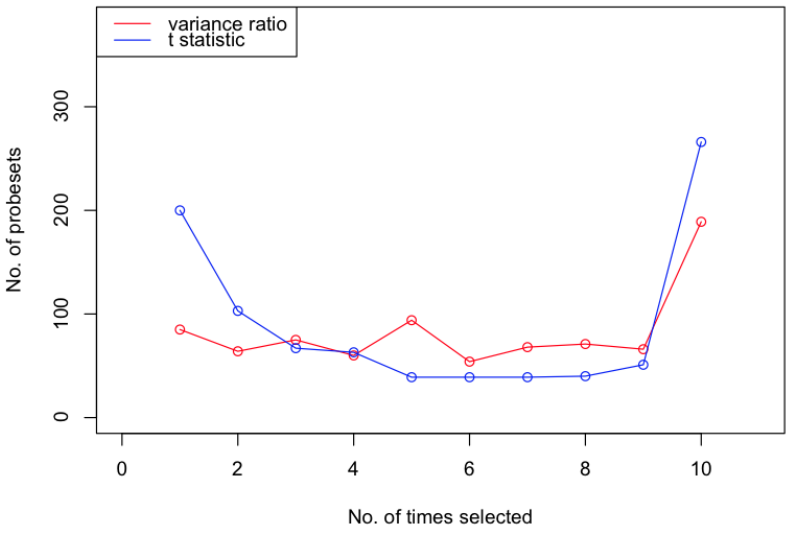

Figure 3.7: **Comparison of feature selection methods:** Probeset selection using the t statistic and the ratio of variances between high risk samples and normal samples from 10 subsets of data for breast microarray data.

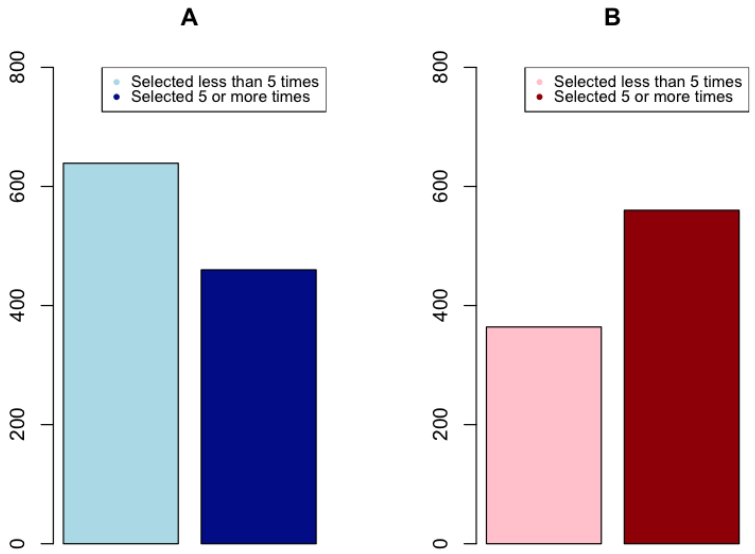

Figure 3.8: **Feature selection with Lung data:** Aggregating the number of probesets selected by (A) the t statistic and (B) variance ratio.

selected 5 or more times in total (Figures 3.8 and 3.9), we obsedved that the variance based selection was more likely to select the same probesets from multiple subsets. These results demonstrate that variance based feature selection can be more stable than mean based feature selections.

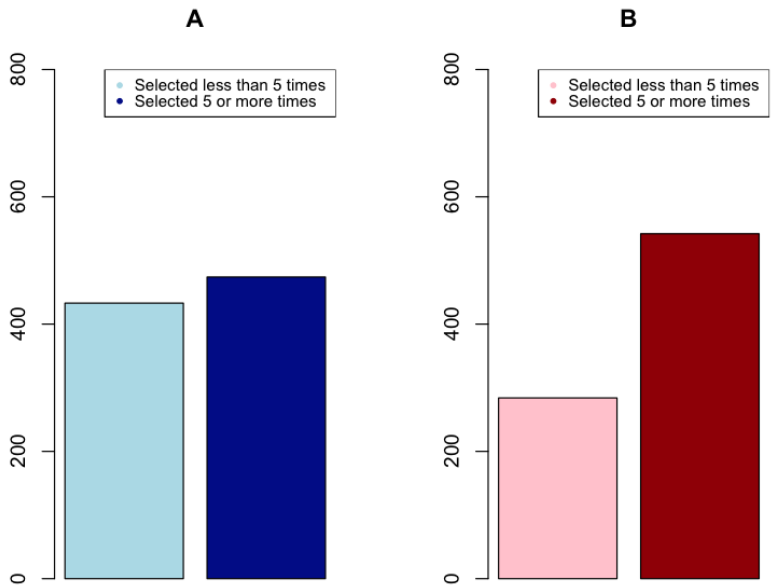

Figure 3.9: **Feature selection with Breast data:** Aggregating the number of probesets selected by (A) the t statistic and (B) variance ratio.

Correlation between universal signature probesets

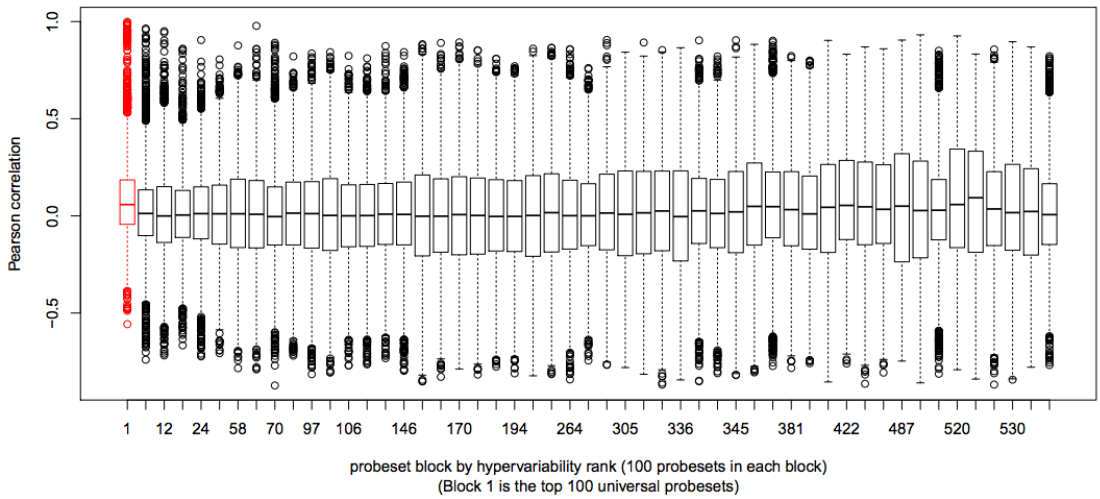

Figure 3.10: **Correlation between features:** Distribution of pairwise correlations for 100-probeset blocks of lung microarray data.

To examine the correlation between the probesets of the universal signature, we ranked the probesets by the hypervariability statistic of the universal signature

and selected successive 100-probeset blocks. This generated 546 blocks, the very first block being the universal signature we have utilized for experiments. For each block, we calculated the pairwise Pearson's correlation values for the probesets. Figure 3.10 shows the distribution of these values for the first block and a set of randomly selected blocks. While the first block has a mean correlation greater than zero, we observed that various other blocks regardless of their position in the hypervariability ranking alsow showed higher means. In addition we observed that the blocks with higher hypervariability (i.e. blocks close to 1), are more likely to contain some outlier pairs with high levels of correlation.

From the top 100 probeset universal signature we used, we selected the pairs that show high levels of correlation by selecting the pairs in the 98th percentile and removing the probeset with lesser hypervariability in those pairs. This reduced the size of the top 100 signature from 100 to 61 probesets. From this reduced signature we calculated anti-profile scores for the lung, colon and adrenocortical microarray datasets used previously. We observed some changes in the results obtained for the lung data: for the first lung dataset, the AUC calculated from the anti-profile scores changed from .716 with the 100 probeset signature to .699 with the reduced signature, while for the second lung dataset the AUC reduced from .558 to .515. For an adenoma against carcinoma classification from a colon dataset, the AUC changed only slightly, from .91 to .9. A similar test for adreoncortical dataset showed no change in the results, where with both signatures the AUC remained at .99.

These observations suggest that for datasets that perform quite well with regard to classification (in particular adenoma against carcinoma classification), the

highly correlated probesets of the universal signature contains superflous information which ultimately do not affect the performance of the anti-profile method negatively. However, these probesets do provide useful information for other classification tasks that are not easily separable, thus we believe our use of the full 100 probeset signature is an optimal choice under these considerations.

3.4 Conclusions

Our aim has been to develop a robust and stable approach for classification of tumor samples. We have demonstrated that the anti-profile scoring method which was initially applied for classification between tumor and normal samples can be extended to classification between tumor samples as well. This method has the particular advantage that tumor samples are only used to select probesets, but given that, the anti-profile score is based strictly on normal tissue samples. The ability of the anti-profile score to successfully provide a ranking of tumor samples which correspond to their risk of relapse (or death), and the robustness of the method across experimental datasets demonstrate that the universal anti-profile signature provides a robust basis to develop feature selection methods for tumor prognosis and diagnosis related microarray experiments. In addition it confirms our hypothesis behind the extension of the anti-profile approach to tumor prognosis: that the measurement of deviation from a set of normal samples which are likely to be more cohesive, is a more stable and robust indicator of the risk level of a tumor sample as opposed to direct comparisons between the highly variable tumor

samples.

High-throughput technologies for gene expression measurements, especially microarrays, have progressed to the point that the use of gene expression data to develop gene expression based cancer signatures is quite common in cancer research [69]. However, despite a number gene expression profile based signatures being published and even commercially utilized, in many instances the developed signature has performed inadequately under subsequent validations. Validations of such signatures should ideally be carried out on populations completely independent from the population selected for the derivation of the signature. Only a few gene signatures produced, such as the Amsterdam 76-gene signature [86], have proven to be reliable for clinical use.

Heterogeneity among multiple types of tumors have been a well known observation [37]. While the proliferating ability of tumor cells is a widely used principle behind many prognostic gene signatures, this is usually measured via a mean-shift based differential expression measurement. However, Irizarry and Feinberg [19] demonstrate that increased variance in the genotype may increase fitness via increased variability of the phenotype, regardless of any significant change of the mean phenotype. This shifts the focus of measuring tumor heterogeneity from a mean shift to a variance shift.

As part of a comprehensive study of the colon cancer methylome, the degree of hyper-variability in DNA methylation between the adenoma and the cancer samples was observed to increase significantly [30]. When projected to a lower dimensional space using PCA, the normal samples clustered tightly together with the cancer

samples dispersed, and the adenoma samples demonstrated an intermediate degree of variability and an intermediate distance to the normal cluster. Based on these findings, Corrada-Bravo et al [11] introduced anti profiles as a stable method for screening multiple types of cancer. The principle underlying this model of cancer screening is that certain genes will consistently show higher across samples variability among cancer samples as compared with normal samples. In this study these genes are identified and the hyper-variability is used to predict outcome, where the model is referred to as an anti-profile as it measures variation from normal behavior. The same study also demonstrated that the genes corresponding to expression hyper-variability in cancer are also generally tissue-specific genes, an observation which is utilized to develop a universal anti-profile. Recent studies have looked at gene expression variability in the context of geneset and pathway discovery [1] and unsupervised construction of profiles in prostate cancer based on outlier analysis [23].

The anti-profile methods developed are applications and extensions to the predictive setting of ideas in existing statistical methods developed to identify and model outliers in gene expression due to cancer [42, 76], and other extensions are in active development [85]. These ideas have are in increasingly used in the analysis of epigenetic data [12, 73]. The general idea of using deviation from a stable class to classify between groups of anomalies is underdeveloped in the Machine Learning field, but should prove to be fertile ground for the development of general methodology [14].

The results presented here confirm that an anomaly classification based approach to gene expression and methylation based experiments of tumor prognosis

and diagnosis can be highly valuable. In summary, our work shows by application to lung cancer, breast cancer, colon cancer and adrenocortical tumor gene expression datasets, and also to thyroid and colon methylation data, that the anti-profile approach does in fact produce models that are accurate, robust and stable.

3.5 Supplementary Information

3.6 Methods and implementation

3.6.1 Microarray dataset analysis

We used the following publicly available microarray datasets and one methylation datasets: for colon cancer anti-profile analysis of carcinoma vs. adenoma comparisons, we used two microarray datasets with GEO access numbers GSE4183 [29](8 normals, 15 adenomas, and 15 carcinomas) and GSE20916 [61](10 normals, 10 adenomas, and 10 carcinomas). Here we used each dataset to generate a colon cancer anti-profile and used it to analysis the hyper-variability of cancer and adenoma samples of the other dataset and to calculate anti-profile scores for those samples.

For carcinoma vs. adenoma comparisons with the universal anti-profile probesets, we used an adrenocortical dataset: GSE10927 [25](10 normals, 22 adenomas, and 33 carcinomas). For this dataset we used the universal anti-profile signature from [11] and the normals in the dataset to calculate anti-profile scores for the adenoma and carcinoma samples. We also used Thyroid data from GSE27155 [26] which is a GPL96 platform microarray dataset for a similar experiment(4 normals, 10 Follicular adenomas, 13 Follicular carcinomas).

Colon tumor data with patient survival information was obtained from GSE39582 [47] for analysis of anti-profiles as tumor prognosis using anti-profiles. For stratifying the samples into low risk and high risk classes we determined whether any given

sample exhibited relapse(instead we used death of patient for some of the other experiments) within a given time period, in which case it was classified as high risk, or if no relapse(or death) was observed within the given period it was classified as low risk. We used a cutoff period of 12 months for the colon dataset and applied an age cutoff of 55 years. This yielded 112 samples, out of which 23 were high risk samples and 74 were low risk samples. We used Kaplan-Meier curves generated from the *survival* R package to analysis the relationship between anti-profile scores and survival of patients.

For a similar analysis using the universal anti-profile probesets with lung cancer, we used two lung datasets: GSE31210 [55] containing both normals and tumors, and GSE37745 [9], containing tumor samples. Both datasets contained patient relapse information and we used relapse within 5 years for high risk and low risk classification. Applying an age cutoff of 65, the first dataset contained 14 normals and 176 tumor samples (42 high risk and 73 low risk), and the second dataset contained 102 tumor samples (19 high risk and 14 low risk).

Similarly we used two breast cancer datasets for prognosis analysis: GSE2990 [66] and GSE1456 [58]. Since patient relapse information was not available for a large number of samples we used death within 5 years to obtain a low risk - high risk classification. In addition we selected only lymph node negative and ER positive samples, leaving 43 samples from the first(8 high risk and 23 low risk) dataset and 156 samples from the second(22 high risk and 120 low risk). We used four datasets with normal breast samples, GSE15852 [57], GSE16873 [17], GSE20437 [27] and GSE21947 [28] which provided 85 samples in total. We also obtained an additional

breast cancer dataset, GSE3494 [51] which contained survival information regarding death of patients. However, almost all the patients were censored within a small time window(10-12 years), thus we could not obtain any useful categorization of the samples which assigned a risk level to each patient; as a result we were unable to obtain any useful analysis from this dataset(the dataset contains 59 samples; a stratification based on patient death or survival within 8 years provides 48 low risk samples, but only 6 high risk samples).

In addition we used DNA methylation data obtained by Hansen et al [30]. The experiment was performed using a custom nucleotide-specific bead array on the Illumina GoldenGate platform to analyze 151 colon cancer-specific differentially DNA-methylated regions (cDMRs). The data contains DNA methylation levels for 384 probesets and 36 cancer, 29 adenoma and 26 normal thyroid tissue samples, and the colon data consist of 28 normal samples, 12 adenoma samples, and 10 cancer samples.

A summary of the datasets used are provided in tables S1 and S2.

3.6.2 Software

We used the R language (version 3.0.2) for our statistical analysis [60]. For anti-profile probeset selection and score calculation, we used the *antiProfiles* [11] package(version 1.2.0) for the R platform. Other R packages used include *ROCR*(version 1.0.5) [64], *survival*(version 2.37.7) [74], *Biobase*(version 2.22.0) [20], *minfi*(version 1.13.8) [5], and *pamr*(version 1.54.1) [77].

Name	**GEO Reference**	**Tissue**	**Normals**	**Adenomas**	**Cancer**	**Survival information**
Gyorffy et al.	GSE4183	Colon	8	15	15	No
Skrzypczak et al.	GSE20916	Colon	10	10	10	No
Marisa et al.	GSE39582	Colon	0	0	566	Yes
Giordano et al.	GSE10927	Adrenocortical	10	22	33	No
Okayama et al.	GSE31210	Lung	20	0	226	Yes
Botling et al.	GSE37745	Lung	0	0	196	Yes
Sotiriou et al.	GSE2990	Breast	0	0	101	Yes
Pawitan et al.	GSE1456	Breast	0	0	156	Yes
	GSE15852	Breast	43	0	0	No
	GSE16873	Breast	12	0	0	No
	GSE20437	Breast	12	0	0	No
	GSE21947	Breast	18	0	0	No

Table 3.3: **A summary of the gene expression microarray datasets used.**

3.7 Supplementary figures

Name	GEO Reference	Tissue	Normals	Adenomas	Cancer
Hansen et al.		Thyroid	26	29	36
Hansen et al.		Colon	28	12	10
Timp et al.	GSE53051	Thyroid	12	21	14
Timp et al.	GSE53051	Colon	18	10	9
Timp et al.	GSE53051	Breast	10	4	10
Timp et al.	GSE53051	Pancreas	12	6	23
Timp et al.	GSE53051	Lung	12	6	23

Table 3.4: **A summary of the DNA methylation datasets used.**

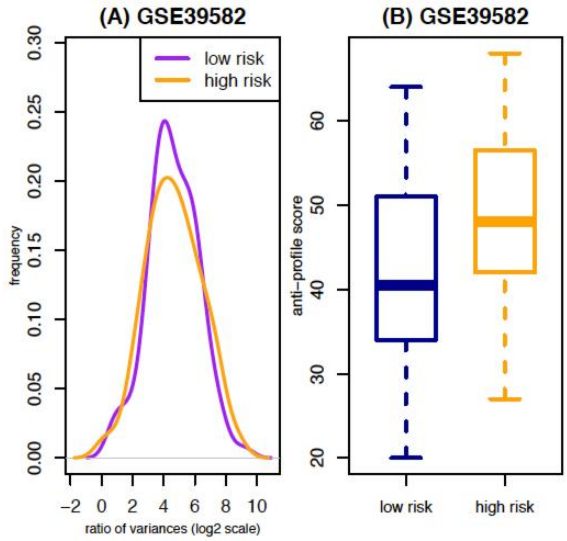

Figure 3.11: **Colon cancer survival analysis based on patient relapse.** (A)Distribution of variance ratio statistic for high risk and low risk samples from colon dataset(Marisa et al.; GSE39582) from an anti-profile computed using another colon data. (B)Distribution of anti-profile scores among low risk and high risk samples; AUC = 0.684, Wlcoxon rank-sum test p-value = 0.0078.

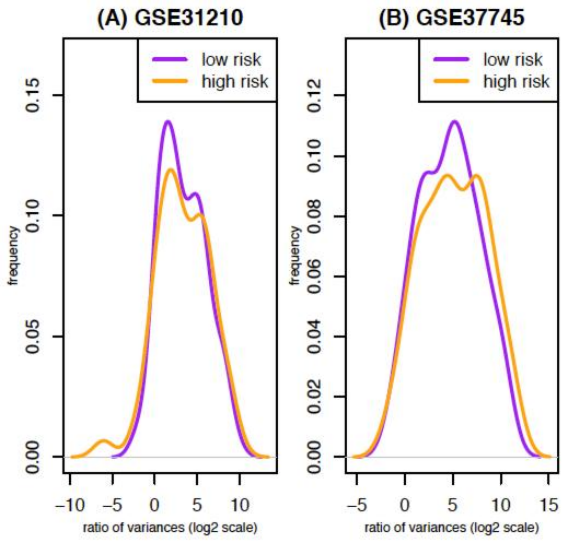

Figure 3.12: **Lung cancer survival analysis based on relapse.** (A)Distribution of variance ratio statistic for high risk and low risk samples from first lung dataset(Okayama et al.; GSE31210) for 100 universal anti-profile probests with the highest hyper-variability.(B)Distribution of variance ratio statistic for high risk and low risk samples from second lung dataset(Botling et al.; GSE37745) for 100 universal anti-profile probests with the highest hyper-variability.

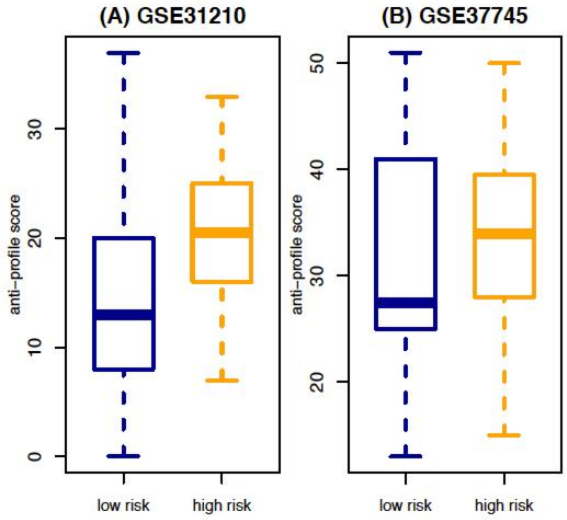

Figure 3.13: **Lung cancer prognosis is related to the anti-profile score.** (A)Anti-profile scores for first dataset(Okayama et al.) high and low risk samples from universal anti-profile probesets; AUC = 0.716, Wlcoxon rank-sum test p-value $< 10^{-3}$. (B)Anti-profile scores for second dataset(Botling et al.) high and low risk samples; AUC = 0.558, Wlcoxon rank-sum test p-value = 0.58.

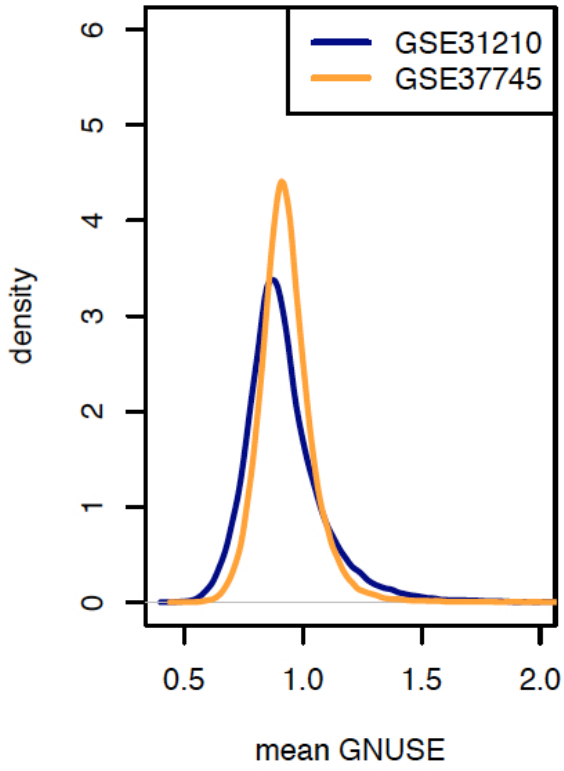

Figure 3.14: **GNUSE value comparison.**
Distribution of generalized normalized unscaled standard error values for the two lung cancer datasets, GSE31210 (Okayama et al.), and GSE37745 (Botling et al.). The second dataset has a higher standard error profile.

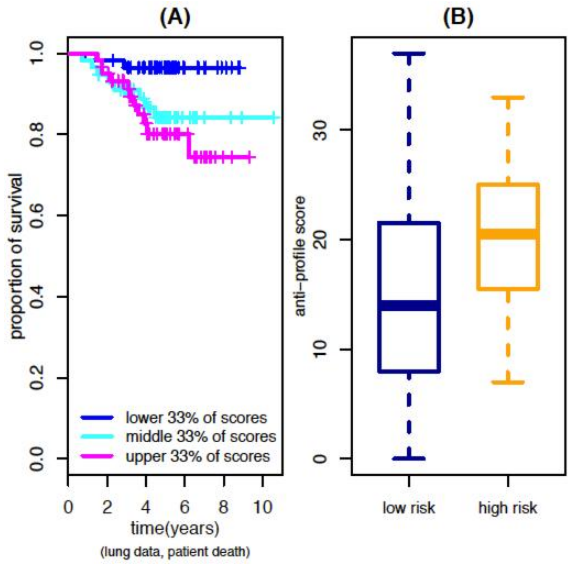

Figure 3.15: **Additional lung cancer survival results.**
(A)Kaplan-Meier survival curves based on patient death for first lung dataset(Okayama et al.); samples ranked by anti-profile scores are grouped to three equal sized groups; Logrank test score 8.342, p-value 0.015. (B)Distribution of same anti-profile scores for high and low risk classification based on patient death within 5 years; AUC = 0.685, Wlcoxon rank-sum test p-value 0.01.

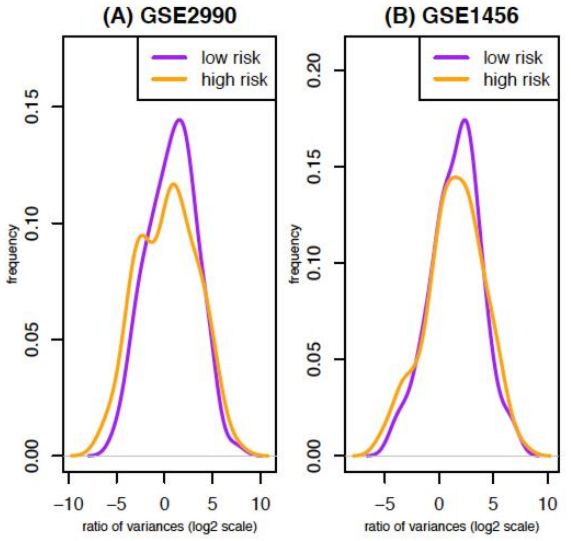

Figure 3.16: **Breast cancer analysis based on patient death.**
(A)Distribution of variance ratio statistic for high risk and low risk samples from first breast dataset(Sotiriou et al.; GSE2990) for 100 universal anti-profile probests with the highest hyper-variability.(B)Distribution of variance ratio statistic for high risk and low risk samples from second breast dataset(Pawitan et al.; GSE1456) for 100 universal anti-profile probests with the highest hyper-variability.

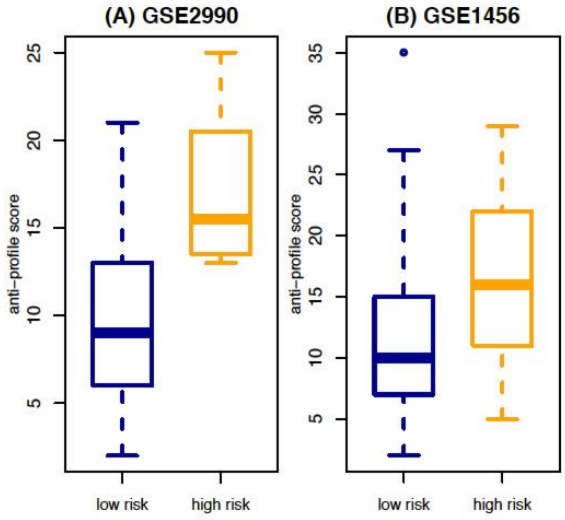

Figure 3.17: **Breast cancer prognosis is related to the anti-profile score.** (A)Anti-profile scores for first dataset(Sotiriou et al.) high and low risk samples from universal anti-profile probesets; AUC = 0.832, Wlcoxon rank-sum test p-value 0.0061. (B)Anti-profile scores for second dataset(Pawitan et al.) high and low risk samples; AUC = 0.743, Wlcoxon rank-sum test p-value $< 10^{-3}$.

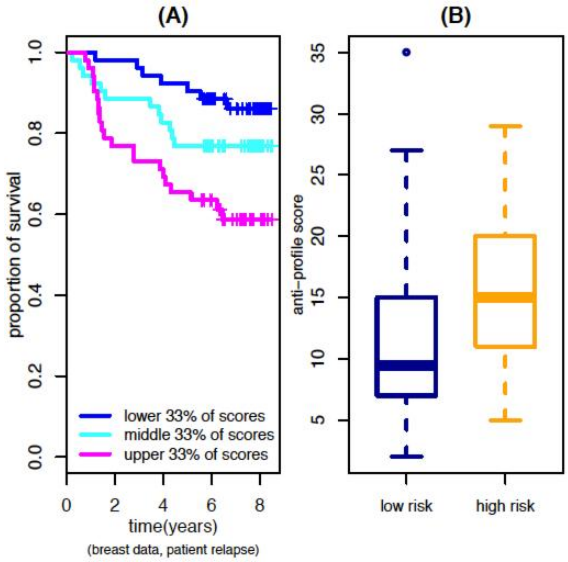

Figure 3.18: **Additional breast cancer survival results.** (A)Kaplan-Meier survival curves based on relapse for second breast cancer dataset(Pawitan et al.); samples ranked by anti-profile scores are grouped to three equal sized groups; Logrank test score 10.755, p-value 0.004. (B)Distribution of same anti-profile scores for high and low risk classification based on relapse within 5 years; AUC = 0.703, Wlcoxon rank-sum test p-value $< 10^{-3}$.

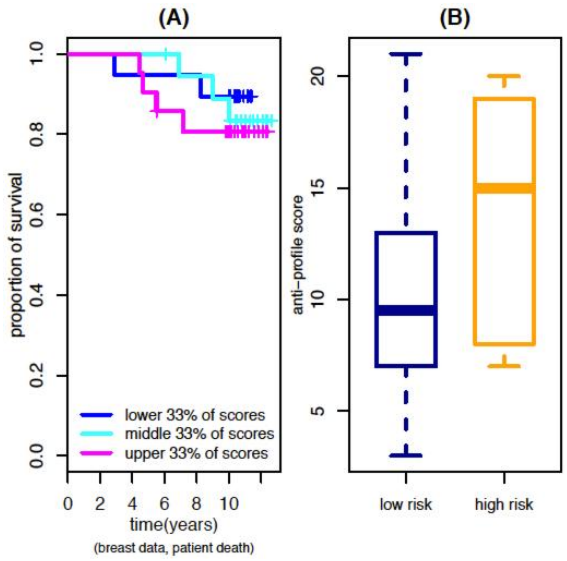

Figure 3.19: **Additional breast cancer survival results.**
(A)Kaplan-Meier survival curves based on patient death for an additional breast cancer dataset(Miller et al.; GSE3494); samples ranked by anti-profile scores are grouped to three equal sized groups; Logrank test score 0.696, p-value 0.706. (B)Anti-profile scores obtained for the third breast cancer dataset for a risk classification based on patient death or survival within 8 years(48 low risk and 6 high risk samples); AUC = 0.694, Wlcoxon rank-sum test p-value 0.1257. Our method of stratifying samples into risk groups was not well applicable for this dataset.

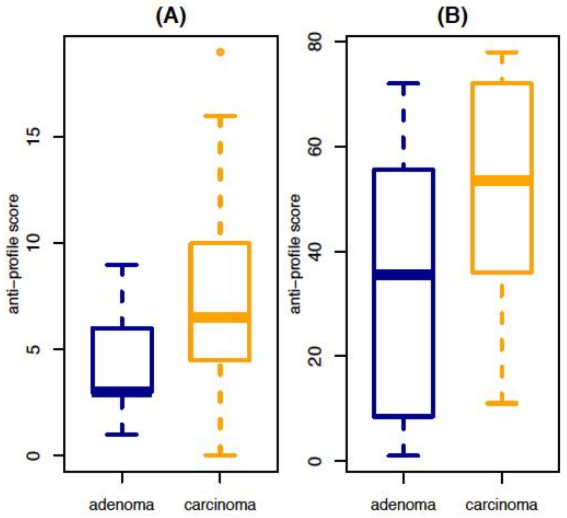

Figure 3.20: **Anti-profiles applied to methylation data.**
(A)Distribution of anti-profile scores for adenoma and carcinoma for thyroid tumor samples from methylation data(AUC = 0.784, Wilcoxon rank-sum p value $< 10^{-5}$); (B)Distribution of anti-profile scores for adenoma and carcinoma for colon tumor samples from methylation data(AUC = 0.717, Wilcoxon rank-sum p value = 0.093).

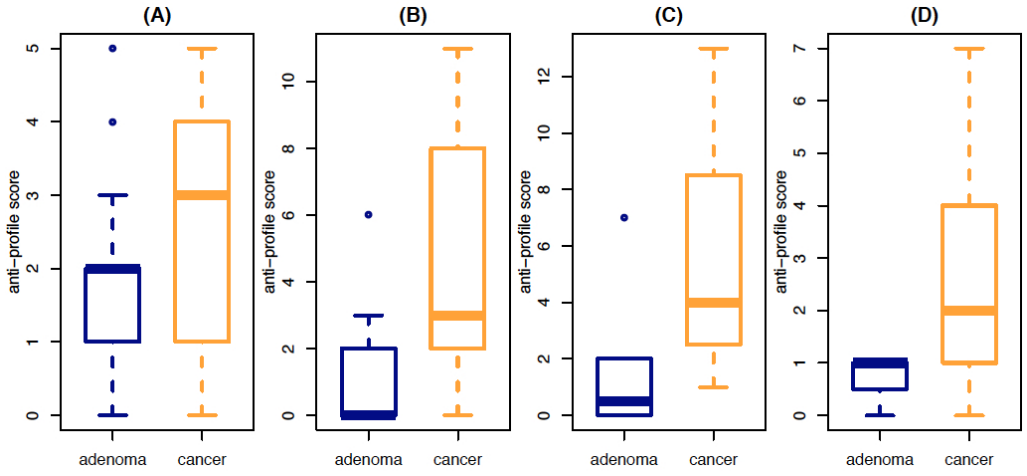

Figure 3.21: **Anti-profiles applied to Illumina HumanMethylation450 data.** Distribution of anti-profile scores for adenoma and carcinoma for (A)Thyroid (AUC = 0.682, Wilcoxon rank-sum p value = 0.068) (B)Colon (AUC = 0.816, Wilcoxon rank-sum p value = 0.018) (C)Pancreas (AUC = 0.840, Wilcoxon rank-sum p value = 0.011) and (D)Breast (AUC = 0.837, Wilcoxon rank-sum p value = 0.058) tissue samples.

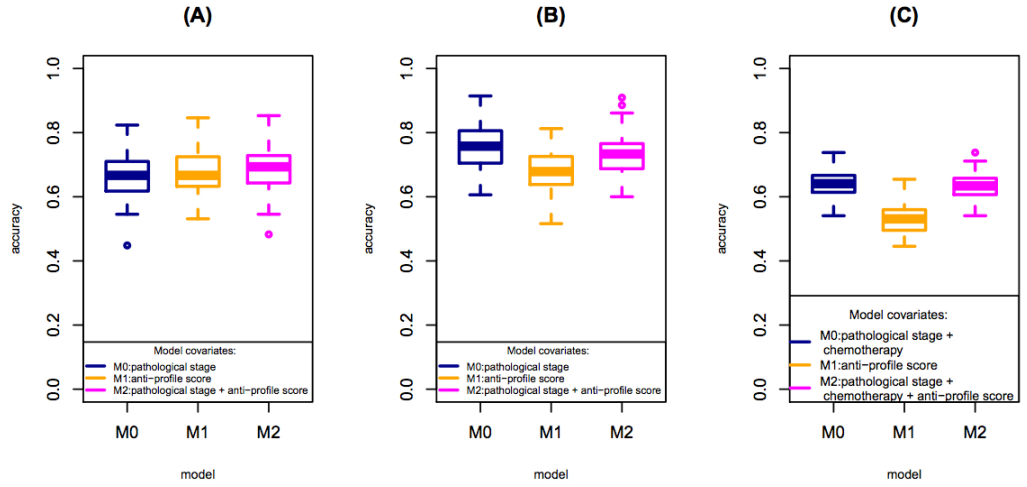

Figure 3.22: **Anti-profiles applied to Cox models for survival prediction.** Cox proportional hazard models with significant clinical covariates and anti-profiles were used to predict patient survival at 5 years for (A)first lung cancer dataset (Okayama et al.) with patient relapse, (B)first lung cancer dataset with patient death, and (C)colon cancer dataset (Marisa et al.) with patient relapse. The plots show accuracy of prediction calculated for 100 training and testing subsets randomly selected. Model M0 only contains significant clinical covariates, model M1 contains only the anti-profile score, and model M2 contains both selected clinical covariates and the anti-profile score.

Chapter 4: Support Vector Machines

4.1 Introduction

We begin out discussion on support vector machines (SVMs) by introducing kernel functions. We follow it up with a derivation of the SVM and then present how SVMs are generally used for anomaly detection by presenting the one class SVM.

4.2 Kernels

For measuring the similarity between normal and anomalous samples, we use kernel functions. Kernels functions provide an efficient way to compute similarities between vectors and have a wide range of applications in statistical learning. The overview of kernel methods presented here is based on the material presented by Scholkopf and Smola in [62].

If we are given a set of vectors $x \in X$ where $X \in \mathbb{R}^N$, then rather than applying a learning technique in a two dimensional feature space, we can augment the feature space by creating artificial features which are the products of various monomials. For example, for $N = 2$, we can construct a mapping Φ which transforms the two

dimensional feature space to a higher dimensional feature space as follows:

$$\Phi : \mathbb{R}^2 \to \mathbb{R}^3$$

$$([x]_1, [x]_2) \to ([x]_1^2, [x]_2^2, [x]_1[x]_2)$$

Figure 4.1 provides an example for a mapping from $\mathbb{R}^1$ to $\mathbb{R}^2$. While the points belonging to the two classes(red and blue) on the line cannot be separated by a single point, when projected to a two dimensional space using the map $x \to (x, x^2)$, they can indeed be separated by a single line.

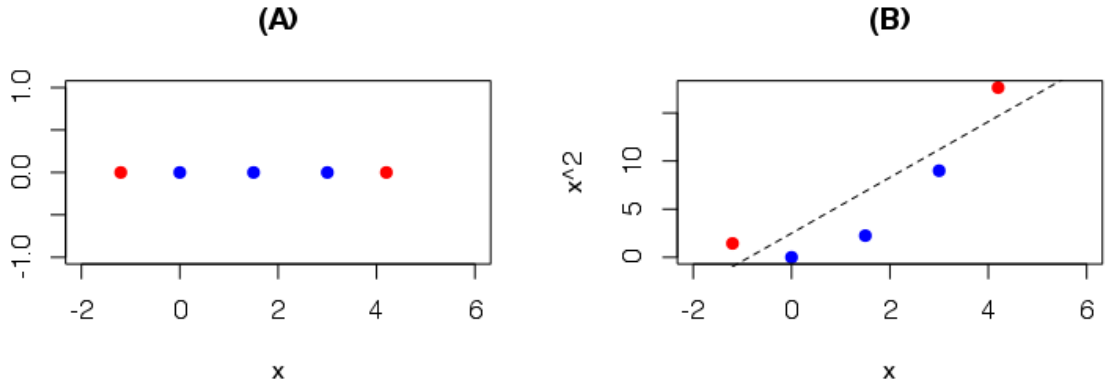

Figure 4.1: **Seperation of classes through projection to a higher dimensionality.**

A one-dimensional set of points that cannot be separated(A) can be made separable by a straight line when mapped to a two-dimensional space(B).

However, for large N, creating all possible feature products and using these in the learning application can be extremely computationally expensive. This is where the kernel trick comes to be applied.

For certain cases, dot products in high dimensional spaces can be computed without explicitly constructing the mapping and then applying multiplication. For

these functions, if the learning method relies solely on dot products of the sample vectors, then replacing the dot products with the value of the kernel function would be equivalent to mapping the vectors to a high dimensional space and applying vector multiplication, without the additional computing cost.

Given the implicit mapping in the kernel to be Φ, we can write the kernel function k as:

$$k(x_i, x_j) =< \Phi(x_i), \Phi(x_j) >$$

If we consider the mapping given above, we can write the dot product of the mapping as follows:

$$< \Phi(x_i), \Phi(x_j) >= [x_i]_1^2[x_j]_1^2 + [x_i]_2^2[x_j]_2^2 + 2[x_i]_1[x_j]_1[x_i]_2[x_j]_2 =< x_i, x_j >^2$$

Therefore the mapping is performed implicitly by using the kernel function

$$k(x_i, x_j) =< x_i, x_j >^2$$

For a given symmetric kernel function $k : X^2 \to \mathbb{R}$, for patterns $x_1, x_2, ..x_m \in X$, we can define a matrix K with elements

$$[K]_{ij} = k(x_i, x_j)$$

which is referred to as the Gram matrix of the function$k$ for $x_1, x_2, ..x_m$. If the Gram matrix has the property

$$\sum_i^m \sum_j^m c_i c_j [K]_{ij} \geq 0 \ \forall c_i, c_j \in \mathbb{R}$$

then the kernel function is considered to be positive semi-definite. Therefore for a given function k on $X \times X$ which induces a positive semi-definite Gram matrix for

all $x_1, x_2, ..x_m \in X$ (and for any $m \in N$), the kernel function is said to be positive semi-definite.

The positive semi-definite property of the Gram matrix implies that all the eigenvalues of the matrix are non-negative. In addition, this implies that the diagonal values of the Gram matrix are non-negative as well - i.e. for any $x \in X$, $k(x, x) \geq 0$. In general usage a given kernel function is assumed to be positive semi-definite.

For a given kernel function, we can construct an equivalent dot product space as follows: the implicit mapping Φ can be denoted as a function of the following form:

$$\Phi : X \to \mathbb{R}^X$$

$$x \to k(., x) \; \forall x \in X$$

$$\Phi(x)(.) = k(., x)$$

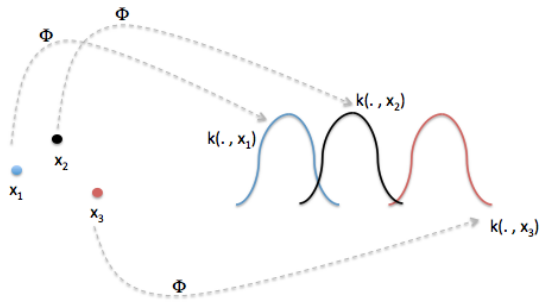

Figure 4.2: **Mapping data via kernel functions.**
The mapping $\Phi()$ transforms a point to a function $k(., x)$ via the associated kernel.

Here $\Phi(x)$ can be thought of as a function which, for a given input from X, measures the similarity of the given input to x (Figure 4.2).

k Is also referred to as the representer of evaluation:

$$< k(.,x), k(.,x') >= k(x,x')$$

Considering these properties, k can be said to be a reproducing kernel [83].

The dot product space constructed in this manner is also referred to as a pre-Hilbert space. By adding a norm operation to this space, it can be turned into a reproducing kernel Hilbert space. This can be formally defined as follows [62]:

Let X be a non-empty set and $\mathcal{H}$ a Hilbert space of functions $f : X \to \mathbb{R}$. Then $\mathcal{H}$ is called a reproducing kernel Hilbert space endowed with the dot product $< .,. >$ if there exists a function $k : X \times X \to \mathbb{R}$ with the following properties:

1. k has the reproducing property:

$$< f, k(x,.) >= f(x) \;\; \forall f \in \mathcal{H}$$

2. k spans $\mathcal{H}$.

In particular, the reproducing property implies that $< k(.,x), k(.,x') >= k(x,x')$.

4.3 The Support Vector Machine

Support Vector Machines(SVMs) are one of the most widely used classification tools available today. Here we present an overview of the derivation of the SVM method for binary classification. In section 3 we present a modification to the SVM which can be used for anomaly classification.

SVMs provide the advantage that little or no prior knowledge about the particular problem is required for its effective application. This follows from the theo-

retical underpinning of SVMs, where the SVM is formulated as a maximum margin classifier([81]): if a set of points belonging to two classes in Euclidean space are given(though SVMs can be used for categorical datasets as well), the goal of the SVM is to obtain the hyperplane that separates the two classes with the largest possible margin, irrespective of the underlying statistical properties of the set of points. If the classes are not linearly separable, then the hyperplane that separates the two classes to the best possible extent with the minimum amount of misclassification is selected, in which case we say that the SVM is a soft margin classifier. An outline of the derivation of the SVM method is presented in the following, based on Hastie, Tibshirani and Friedman [32]:

Consider a data set consisting of N pairs $(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)$ with $x_i \in \mathbb{R}^p$ and $y_i \in \{-1, 1\}$. If the given two classes are class A and class B, then for any data point i, $y_i = -1$ indicates that the point belongs to class A and $y_i = 1$ indicates that the point belongs to class B. Define a hyperplane by $f(x)$ as follows:

$$f(x) = x^T w + b = 0$$

Here w is a vector of length p with $||w|| = 1$, and b is a constant. It can be shown that for a given point x, $|f(x)|$ is equal to the perpendicular distance from that point to the hyperplane defined by $f(x) = 0$. The sign of $f(x)$ indicates which side of the hyperplane the given point lies in. Therefore, if the two classes are separable, there should exist a function $f(x)$ such that $y_i f(x_i) > 0$ for all the points.

Let us consider the simpler scenario in which the two classes are separable. For a given hyperplane $f(x) = 0$, let M be the margin between the two classes and the

hyperplane which separates the two classes. Here the distance from the hyperplane to each class is defined as the shortest distance from the hyperplane to any point in that class. Then the calculation of the hyperplane becomes an optimization problem of the form

$$\begin{aligned} \underset{w,b}{\text{minimize}} \quad & \frac{1}{2}||w||^2 \\ \text{subject to} \quad & y_i(x_i^T w + b) \geq M, i = 1..N. \end{aligned}$$

Now consider the case in which the two classes are not separable by a hyperplane. In this scenario, we have no other recourse but to allow some amount of error - i.e. we must allow a certain number of points to lie on the incorrect side of the margin. Thus the optimization problem becomes one of maximizing M while minimizing the sum total of error generated by each misclassified point. For a given point x_i, let ξ_i be the error generated. The magnitude of the error is the distance from the point to the hyperplane, measured positively, with the exception that for a point that is correctly classified (i.e. the point lies on the correct side of the hyperplane) the error is zero. Hence for correctly classified points $\xi_i = 0$.

Using the former constraint expression, we can proceed as before. To minimize the error induced, we need to bound the error by adding the minimization of $\sum_i \xi_i$ to the optimization problem. If, we before we use $M = \frac{1}{||w||}$ then the constraint becomes

$$y_i(x_i^T w + b) \geq 1 - \xi_i \ \ , \xi_i \geq 0 \ \ , \ \ \forall i$$

At this point, $y_i(x_i^T w + b)$ measures the positive distance from the point to the hyperplane, and therefore when $\xi_i > 1$, a misclassification has occured.

Then the final optimization problem is as follows:

$$\begin{aligned} \underset{w,b}{\text{minimize}} \quad & \frac{1}{2}||w||^2 + C\sum_{i=1}^{N}\xi_i \\ \text{subject to} \quad & y_i(x_i^T w + b) \geq 1 - \xi_i,\ \xi_i \geq 0\ i = 1..N. \end{aligned}$$

Here $C \in \mathbb{R}$ is a constant that measures how lenient we would like to be with the amount of error allowed. Large C implies that we would like the minimize the error to much larger degree while sacrificing the margin size, and vice versa. As $C \to \infty$, the problem approximates the separable scenario.

Now it is possible to form the Lagrangian of the optimization to obtain an unconstrained problem:

$$L = \frac{1}{2}||w||^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i[y_i(x_i^T w + b) - (1 - \xi_i)] - \sum_{i=1}^{N}\mu_i\xi_i$$

where α_i and μ_i are the Langrange multipliers. Then the optimization is

$$\underset{w,b,\xi_i}{\text{maximize}} \quad L$$

This is referred to as the primal problem.

Commonly, from this point onwards the focus shifts to forming the dual problem and its optimization. The Wolfe dual can be derived by differentiating the Langrangian with respect to variables w, b and ξ_i. This provides the following:

$$w = \sum_{i=1}^{N}\alpha_i y_i x_i$$

$$\sum_{i=1}^{N}\alpha_i y_i = 0$$

$$\alpha_i = C - \mu_i\ \forall i$$

Applying these, the dual problem can be written as

$$L' = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j < x_i . x_j >$$

Thus the optimization problem is

$$\underset{\alpha_i}{\text{maximize}} \quad \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j < x_i . x_j >$$

$$\text{subject to} \quad 0 \leq \alpha_i \leq C \;\; i = 1..N.$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0$$

In addition, the Karush-Kuhn-Tucker conditions can be applied to obtain further constraints the describe the dual problem.

What provides the SVM with its true effectiveness is not simply the ability to obtain a maximum margin separating hyperplane, but the ability to use kernel functions with very little additional cost. Since the given data points only appear as products in the dual problem, we can augment this formulation to obtain a classifier that functions effectively in a different feature space.

As discussed in the previous section, if the given data is not separable in its original space, say $\mathbb{R}^p$, then we can project the data to a higher dimensionality $\mathbb{R}^q$ where $q > p$, such that the two classes might become separable. Consider a mapping $\phi(.)$ which performs this operation: i.e. for $x_i \in \mathbb{R}^p$, $\phi(x_i) = z_i$ where $z_i \in \mathbb{R}^q$. Then If we were to map all the given data using ϕ, then dual problem would be

$$L' = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \phi(x_i) \phi(x_j)$$

If there is a kernel function $k(.,.)$ such that $k(x, y) = \phi(x)\phi(y)$, then we do not need to do the computational work to project each data point, nor do we need to perform

any costly vector operations in high dimensional spaces. Therefore, assuming that k performs more efficiently than performing the actual vector projection and product calculation, we can obtain the same effectiveness of performing the classification in a higher dimension with a much lower cost.

To decide which class a new point (say $\hat{x}$) belongs to, we need to evaluate $f(\hat{x})$. Depending on whether the function value is positive or negative, we can predict which class it belongs to. Since $w = \sum_{i=1}^{N} \alpha_i y_i x_i$, we can write the decision function $g()$ as

$$g(\hat{x}) = sign(\sum_{i=1}^{N} \alpha_i y_i k(x_i, \hat{x}) + b)$$

where $sign(x) = 1$ if $x >= 0$ and -1 otherwise. b Can be calculated by solving $y_i f(x_i) = 1$ for any point with $0 < \alpha_i < C$. The hyperplane is solely defined by the points for which $\alpha_i > 0$ and these points are called support vectors.

In our experiments we mainly use the Gaussian Radial Basis function, one of the most commonly used kernel functions:$k(x, y) = exp(-\frac{||x-y||^2}{2\sigma^2})$, or more commonly written as $k(x, y) = exp(-\gamma||x - y||^2)$. Therefore in addition to the C parameter, the kernel parameter γ also has to be selected. Usually these parameters are selected via cross validation over a set of data specifically set aside(i.e. not used for for testing). Cross validation involves dividing the given dataset into a number of equal sized portions (in the simplest case 2 groups) and leaving out one group for testing while fitting the SVM to the other remaining groups. Alternatively, a separate dataset can be kept aside apart from the training and testing set, and the parameters that best perform on this set can be selected. Both approaches have

been used in the experiments described here. In both cases, a large number of pre-selected parameters are used, and the SVM is fitted and tested over all possible parameter combinations.

It can be shown that a solution to the problem of the following form, where $h \in \mathcal{H}_K$

$$\text{minimize} \quad \sum_i (1 - y_i f_i(x)) + \lambda ||h||^2_{\mathcal{H}_K}$$

can be written in the form of [83]:

$$f(x) = b + \sum_i \alpha_i k(x, x_i)$$

Thus the SVM belongs to a larger class of classifiers which can be generally represented using reproducing kenrel Hilbert spaces.

4.4 The Single-Class SVM

The single class SVM is a particular formulation of the SVM problem designed to address anomaly detection. While we provide a modified SVM for anomaly classification in section 3, here we present the single class SVM for comparison, and discuss why the single class SVM is difficult to use directly for anomaly classification.

The objective of the single class SVM is to obtain a hyperplane such that all the given training data provides a positive value from the resulting decision function [62]. It is assumed that the training data is free of anomalies, and hence any novel or anomalous instances will be identified by being placed on the opposite side of the hyperplane.

The one class SVM formulation is as follows: let $x_1..x_m \in \mathbb{R}^N$ be the samples that we are using for fitting the SVM model, m being the number of samples, with dimensionality N. Let ϕ be a function mapping the samples from $\mathbb{R}^N$ to $\mathcal{H}$. Here $\mathcal{H}$ is the dot-product space of ϕ. Then the dot products lying in $\mathcal{H}$ can be computed via a kernel function k as

$$k(x_i, x_j) =< \phi(x_i), \phi(x_j) >$$

To obtain the one class SVM, we formulate a function f which provides positive value within the region containing the normal samples, and a negative value outside this region. As with the regular SVM, the strategy is to map the samples from the input space to the feature space of the kernel and and then separate them from the origin with a maximum margin. This is possible because in the kernel space, the mappings for the normal samples should lie close to each other and close to the margin while the mappings for the outliers would be farther away. For a new observation x, the value of $f(x)$ is determined by evaluating which side of the hyperplane it falls on. Due to the mapping through ϕ, these regions may correspond to various non-linear estimators in the input space.

To obtain the separation the following quadratic maximization problem has to be solved [62]:

$$maximize_{w \in H, \xi \in \mathbb{R}^m, \rho \in \mathbb{R}} \ \frac{1}{2}||w||^2 + \frac{1}{\nu m}\sum_{i=1}^{m}\xi_i - \rho$$

$$s.t. \ \ < w, \phi(x) >\geq \rho - \xi_i, \xi_i \geq 0, \ \ \forall i$$

The ξ_i are non-zero slack variable that are penalized in the objective function.

Using Lagrange multipliers $\alpha_i \geq 0$ and $\beta_i \geq 0$ we can write the Lagrangian

$$L = \frac{1}{2}||w||^2 + \frac{1}{\nu m}\sum_{i=1}^{m}\xi_i - \rho - \sum_{i=1}^{m}\alpha_i(< w, \phi(x_i) > -\rho + \xi_i) - \sum_{i=1}^{m}\beta_i\xi_i$$

By setting derivatives with respect to w, ξ and ρ to zero we obtain the following dual problem:

$$minimize_{\alpha \in \mathbb{R}^m} \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j k(x_i, x_j)$$

$$s.t.\ 0 \leq \alpha_i \leq \frac{1}{\nu m}, \sum_{i=1}^{m}\alpha_i = 1$$

The decision function can now be written as

$$f(x) = sign\left(\sum_{i=1}^{m}\alpha_i k(x_i, x) - \rho\right)$$

Thus the single class SVM requires all the points given for defining the boundary. To calculate ρ, use any of the α_i values as:

$$\rho = \sum_{j}\alpha_i k(x_j, x_i)$$

In general usage, the single class SVM is used when only one group of labeled data is available. The other classes, which are usually anomalous classes, are either completely unavailable or not labeled reliably enough for training a classifier. Khan and Madden [35], in a survey of the usage of one class learning, classifies them under three broad categories:

1. Learning with positive examples only

2. Learning with positive examples and some amount of poorly distributed negative examples

3. Learning with positive and unlabeled data

Scholkopf et al [62] demonstrate the efficacy of the single class SVM by using it for classification of handwritten digits. The data is available from the US Postal Services and contains a set of images with handwritten digits; the goal of classification is to differentiate between the relatively legible handwriting and the illegible writing. Manevitz and Yousef [44] formulate a different one class classification method by designating the outliers as a single class; the classifier is trained to learn the outliers. However, the results presented seem to indicate that such an approach performs poorly compared to learning the non-outlier class. Yu et al [88] use an SVM based method to classify Web pages with positive and unlabeled pages, using an iterative scheme of SVMs. However, this method is sensitive to the number of normals and requires large number of samples to learn well. Since availability of labeled negative data is rare, the usage of single class SVMs for classification of anomalies as presented here is a relatively novel approach.

Chapter 5: An Anomaly Support Vector Machine

5.1 Introduction

The task of anomaly, or outlier, detection is to identify data samples that deviate significantly from a class for which training samples are available. We explore anomaly classification as an extension to this setting, where the goal is to distinguish data samples from a number of anomalous and heterogeneous classes based on their pattern of deviation from a normal stable class. Specifically, presented with samples from a normal class, along with samples from 2 or more anomalous classes, we want to train a classifier to distinguish samples from the anomalous classes. Since the anomalous classes are heterogeneous using deviation from the normal class as the basis of classification instead of building a classifier for the anomalous classes that ignores samples from the normal class may lead to classifiers and results that are more stable and reproducible.

Anomaly classification remains a largely unfocused area of research. Sometimes an ad-hoc approach may be taken resulting in solutions specific to the problem domain that are not generalized to multiple domains. These approaches usually consist of either applying a standard classification method, e.g., Support Vector Machines, to directly distinguish the anomalous classes, or applying an anomaly de-

tection technique [13]. The former approach does not integrate relevant information that could be obtained from the normal class while the latter approach makes it difficult to incorporate differences between the anomalous classes themselves. Attempts to combine both approaches can result in bootstrapping; while this might provide good solutions for the problem at hand, it cannot be used as a general approach to anomaly classification [34].

In this chapter we present an approach that provides an alternative to the above methods. We first formalize the anomaly classification problem as a general learning setting. We next characterize it statistically and computationally using a similarity measurement, and further demonstrate how to use kernel learning methods for this task. Since kernel functions can be interpreted as pairwise similarity functions (satisfying certain conditions), they serve as a natural framework to measure within normal class similarity and anomalous deviation from this class. We also develop kernel-based classification approaches specifically designed for the anomaly classification task and show that they produce accurate classifiers with higher stability than standard classification methods.

The main inspiration behind this approach comes from a widely studied problem in cancer genomics: the development of microarray-based methods for diagnosis and prognosis of cancer [37]. Genomic data obtained from tumor samples from patients with cancer demonstrate that cancer makes genomic markers highly unstable, which has come to be the main barrier to developing stable method of developing statistical methods for screening cancer. Apart from a few exceptions, many gene signatures developed have provided poor results when tested on independently ob-

tained data, indicating that the signature is not adequately robust to be deployed in a clinical setting.

However, recent work by Corrada-Bravo et al [11] demonstrate that since this instability is not present in healthy people, it is possible to take advantage of the increased gene expression variability in cancer to develop a statistical model which provides a stable and robust universal predictor of cancer prognosis. Further research based on this principle has shown that deviation from a stable, healthy population can be used as a stable marker of cancer prognosis [cite].

These insights form the principle behind the anomaly classification approach presented here: that by primarily modeling the solution based on a stable, cohesive non-anomalous class while incorporating the deviations of the anomaly classes for differentiation results in a more robust and stable classifier. To demonstrate this, we present the anomaly SVM (aSVM), a kernel method based on the same principles as the anti-profile method for classification of tumor prognosis. We compare this method with other SVM methods, the regular SVM and the one-class SVM (OC-SVM) which is generally used for anomaly or novelty detection. Applicaiton of these SVM methods to mutiple datasets including genomic data for cancer prognosis demonstrate that the anomaly SVM can yield more robust and stable classifiers for the classification of anomalies. We further use the same kernel method for performing anomaly regression via the anomaly SVM as well.

5.2 Anomaly Classification

As a general formulation of the anomaly classification problem, we develop the case for two anomalous classes as follows: assume we are given training samples in $\mathbb{R}^p$ from three classes: m data points from a normal class Z, and n training data points $< x_1, y_1 >, ..., < x_n, y_n >$ where $x_i \in \mathbb{R}^p$ are the training data vectors and $y_i \in \{-1, 1\}$ are the labels indicating the respective anomalous classes A^- and A^+. Furthermore, we assume that samples x_i from the anomalous classes are *heterogeneous* relative to the normal class Z under some measure of deviation s such that $E[s(z_i, z_j)] << E[s(x_i, x_j)]$ for $z_i, z_j \in Z$ and x_i, x_j in A^-, and similarly for A^+.

For the standard binary classification setting, a training set X of x_i data points and the y_i class labels are assumed to obtain a set of parameters or weights w that parameterize a classification function $f(x; w)$ to distinguish classes A^- and A^+. Parameters are estimated as the optimal weights w^* that minimize regularized loss under some loss function ε that measures generalization performance:

$$\varepsilon(f, w, X) = \frac{1}{n} \sum_{i=1}^{n} \varepsilon(f(x_i, w), y_i)$$

$$w^* = argmin_w \varepsilon(f, w, X)$$

Once the optimal choice of weights have been identified, for a new given data point $\hat{x}$ we obtain the value of the decision function for the new point, $f(\hat{x}, w^*)$, Usually the decision function includes a bias term which can move the cutoff between the two classes to zero. Then the predicted class label for the new point becomes

$sign(f(\hat{x}, w^*))$.

The aim in standard anomaly detection is to determine if data points $\hat{x}$ do not belong to some given class Z, in which case we say that the data points $\hat{x}$ are anomalous with respect to class Z. In anomaly detection, anomalies are not assumed to belong to distinct classes, nor are they assumed to be available during training. Therefore, a function f is learned that can decide whether a given data point belongs in class Z or not. The standard anomaly detection setting does not apply directly here since we consider situations where samples are obtained for anomalies from multiple classes and the class with respect to which they become anomalies. In this case, optimization should be over $\varepsilon(f, w, X, Z)$, where X is the matrix of anomalies from multiple classes.

Another option is to label samples using $y_i = \{-1, 0, 1\}$ and incorporate the normal(0) class along with the anomalous classes$(-1, 1)$ in a multi-class classification problem. However, in the applications we are targeting, anomalous behavior is defined with respect to the normal class and a standard multi-class classification method does not incorporate this knowledge. In principle, normal samples (Z) should be treated as an additional parameter for choosing optimal w:

$$w^* = argmin_w \varepsilon(f, w, X, Z)$$

In summary, standard methods for anomaly detection and multi-class classification are insufficient in the anomaly classification setting: anomaly detection methods are not capable of modeling distinct classes of anomalies, while standard classification methods do not incorporate the underlying anomalous and heteroge-

neous assumptions in this setting. Classifiers suitable for this task should define a boundary between the two anomalous classes based on measurements of the anomalous behavior of the anomalous samples with respect to the normal samples. In this thesis we use kernel methods to incorporate similarity measures that capture anomalous behavior. Kernel functions have the advantage that in addition to being a measure of similarity, it can be applied to obtain class separators in high dimensional spaces. Among the many different methods using kernel functions for learning, here we focus on support vector machines(SVM).

5.3 The anomaly Support Vector Machine

This approach to anomaly detection begins with the formulation of the anomaly detection problem based on similarity measurements. To incorporate this approach into binary classification, we modify the binary SVM classifier to derive the anomaly SVM (aSVM) which we believe would function as a more stable anomaly classifier. While the regular SVM involves measuring the similarity between the samples, the aSVM measures these similarities through comparing them to a third group which we call the normal class. The anomalous behavior of the samples are defined with respect to this third group of data(i.e. the normal class), and when the normal class is more cohesive and stable across experiments and across samples, the measurement of similarity between anomalous samples(i.e. tumors) by measuring their deviation from the normal class will be stable as well.

A binary classification problem can be seen as a problem of measuring a novel

point against two existing classes to decide to which class the point shows more similarity towards. Then, the anomaly detection problem can be stated as a problem of measuring the dissimilarity of a novel point against a given normal class. We can then further extend this approach to anomaly classification via scoring the dissimilarity and deciding the appropriate anomalous class based on the score.

Let $k(x_i, x_j)$ be a similarity function which compares the two points x_i and $x_j \in \mathbb{R}^p$. Weighing the similarity of a new observation against each existing observation, the difference of the sum of weighted similarities for the two groups will be a cumulative measure of the similarity difference against the two groups, and the sign of this sum can hence be taken as a class indicator for classification and prediction.

With the added condition that the similarity function be positive definite, the above formulation becomes analogous to the decision function of an SVM with the similarity function being the kernel function. The decision function can be written as

$$g(x) = sign\left(\sum_{i=1}^{n} c_i k(x_i, x_j) + d\right)$$

where $c_i \geq 0 \ \forall i$ is the weight associated with each point, and d is a bias term.

For this anomaly detection method, while approaching the learning problem using a similarity based method will be useful to understand the principles involved, the main focus will be on kernel functions and kernel based learning techniques.

Here we review SVMs from a function approximation perspective: consider a set of n observations, each observation being drawn from $X \times Y$, where $X \in \mathbb{R}^p$, and $Y \in \{-1, 1\}$. Here p is the number of features in each observation, or the

dimensionality of the feature space. Thus each observation consists of a pair $\langle x_i, y_i \rangle$, $x_i \in \mathbb{R}^p$ and $y_i \in \{-1, 1\}$, for $i = 1..n$; here y_i indicates which of the two classes the observation belongs to. If we introduce a new observation x' which needs to be classified, then the classification problem amounts to comparing x' to the existing set of points and combining the comparisons to make a decision.

Let $k(x_i, x_j)$ be a positive-definite similarity function. Usually in SVMs function k is further assumed to have the reproducing property in a Reproducing Kernel Hilbert Space $\mathcal{H}$ associated with k: $\langle f, k(x, \cdot) \rangle_{\mathcal{H}} = f(x)$ for all $f \in \mathcal{H}$, and in particular $\langle k(x, \cdot), k(y, \cdot) \rangle_{\mathcal{H}} = k(x, y)$. In this case, the basis functions in the classifier correspond to representers $k(x, \cdot)$. In the standard SVM, the representers of all training points are potentially used as basis functions in the classifier, but effectively only a small number of representers are used as basis functions, namely the Support Vectors. However, for a given problem, we may choose a different set of points for the derivation of the set of basis functions; the basis functions determine how the similarities are measured for a new point.

The core idea in the anomaly SVM (aSVM) is to make use of this characterization of the Support Vector Machine as a linear expansion of basis functions defined by representers of training samples. In order to address the heterogeneity assumption underlying the anomaly classification problem we define basis functions only using samples from the stable normal class.

Formally, we restrict the set of functions available to define the subspace of $\mathcal{H}$ spanned by the representers of samples $z_1, \ldots, z_m$ from normal class Z: $f(x) = d + \sum_{i=1}^{m} c_i k(z_i, x)$. To estimate coefficients c_i in the basis expansion we apply the

usual regularized risk functional based on hinge loss

$$R_{\text{reg}}(f) = \frac{1}{n}\sum_{j=1}^{n}(1 - y_i f(x_i))_+ + \frac{\lambda}{2}\|h\|^2_{\mathcal{H}},$$

where $(\cdot)_+ = \max(0, \cdot)$, $f(x)$ is defined as $f(x) = d + h(x)$, and $\lambda > 0$ is a regularization parameter. By the reproducing kernel property, we have in this case that $\|h\|^2_{\mathcal{H}} = c' K_n c$ where K_n is the kernel matrix defined on the m normal samples.

The minimizer of the empirical risk functional is given by the solution of a quadratic optimization problem, similar to the standard SVM, but with two kernel matrices used: K_n, defined in the previous paragraph, and K_s, which contains the evaluation of kernel function k between anomalous samples $x_1, \ldots, x_n$ and normal samples $z_1, \ldots, z_m$:

$$\begin{aligned} \min_{d,c,\xi} \quad & e^T\xi + \tfrac{n\lambda}{2} c^T K_n c \\ \text{s.t.} \quad & Y(K_s c + de) + \xi \geq e, \xi \geq 0 \end{aligned} \tag{5.1}$$

Here we use slack variables $\xi = (\xi_1, \xi_2, ..., \xi_n)'$, denote the unit vector of size n as e (i.e. $e_{1\times n} = [1, 1, ..., 1]^T$), and define matrix Y as the diagonal matrix such that $Y_{ii} = y_i$.

The Lagrangian of problem 5.1 is given by

$$L(c, d, \xi, \alpha, \beta) = e^T\xi + \frac{n\lambda}{2} c^T K_n c - \alpha^T \left[Y(K_s c + de) + \xi - e\right] - \beta^T \xi$$

where $\alpha_{n\times 1} = (\alpha_1, ..., \alpha_n)^T$ and $\beta_{n\times 1} = (\beta_1, ..., \beta_s)^T$ are the Lagrangian multipliers.

Minimizing with respect to z, c and d, we obtain the Wolfe dual as follows:

$$\frac{\partial L_p}{\partial z} = 0 \Rightarrow e^T - \alpha^T - \mu^T = 0$$

$$\frac{\partial L_p}{\partial d} = 0 \Rightarrow \alpha^T Y = 0$$

$$\frac{\partial L_p}{\partial c} = 0 \Rightarrow \frac{\lambda}{2} 2c^T K_n - \alpha^T Y K_s^T = 0$$

$$c^T = \frac{1}{\lambda}\left(\alpha^T Y K_s^T K_n^{-1}\right)$$

$$c = \frac{1}{\lambda}\left[(K_n^{-1})^T (Y K_s^T)^T \alpha\right]$$

$$c = \frac{1}{\lambda}\left[K_n^{-1} K_s^T Y \alpha\right]$$

Applying these results to L_p, we get

$$L_d = e^T\alpha + \frac{\lambda}{2}.\frac{\alpha^T Y K_s^T K_n^{-1}}{\lambda}.K_n.\frac{K_n^{-1}K_s Y\alpha}{\lambda} - \alpha^T Y K_s^T \frac{K_n^{-1}K_s Y\alpha}{\lambda}$$

$$L_d = e^T\alpha - \frac{\alpha^T Y K_s^T K_n^{-1} K_s Y\alpha}{2\lambda}$$

Thus the dual problem is:

$$\begin{aligned} \max_{\alpha} \quad & e^T\alpha - \tfrac{1}{2n\lambda}\alpha^T Y \tilde{K} Y \alpha \\ \text{s.t.} \quad & 0 \le \alpha \le e, e^T Y \alpha = 0 \end{aligned} \tag{5.2}$$

where $\tilde{K} = K_s K_n^{-1} K_s^T$. Here we assume K_n^{-1} represents a pseudo-inverse in the case where K_n is not positive definite.

For a standard SVM, the objective of the Wolfe dual is $e^T\alpha - \frac{1}{2n\lambda}\alpha^T Y K Y \alpha$, with K the kernel matrix the training datapoints. Thus the dual problem of the aSVM has the same form as the standard SVM dual problem with the exception that kernel matrix K is replaced by induced kernel matrix $\tilde{K}$ in the aSVM. Kernel matrix

$\tilde{K}$ essentially represents an indirect kernel between anomalous samples induced by the set of basis functions determined by the samples from the normal class. Since the essential form of the SVM solution is unchanged by the modification, this provides the additional advantage that the modified SVM can be solved by the same tools that solve a regular SVM, but with a different kernel matrix provided. For our particular problem domain, we use the indirect kernel to represent deviation from the profile of normal samples, and thus refer to this classifier as the anti-profile SVM.

We see above that the aSVM can be solved as a standard SVM with induced kernel $\tilde{K} = K_s K_n^{-1} K_s^T$. In this section we characterize this indirect kernel, and state a general result that elucidates how the aSVM can produce classifiers that are more robust and reproducible that a standard SVM in this setting.

Let the matrix Z contain the normal samples $z_1, ..., z_m$. Then for a given anomalous sample x_i we can obtain the projection of x_i to the space spanned by Z by obtaining a vector β that provides the necessary projection via minimizing the residual sum of squares(RSS).

$$RSS(\beta) = (x_i - Z\beta)^T (x_i - Z\beta)$$

$$\frac{\partial RSS}{\partial \beta} = -2Z^T (x_i - Z\beta)$$

$\frac{\partial RSS}{\partial \beta} = 0$ gives

$$Z^T x_i = Z^T Z\hat{\beta}$$

$$\hat{\beta} = (Z^T Z)^{-1} Z^T x_i$$

Then the required projection is

$$\hat{x}_i = Z(Z^T Z)^{-1} Z^T x_i$$

If we obtain the projection for another anomalous sample x_j, we can calculate the product

$$< \hat{x}_i, \hat{x}_j >= \left[Z(Z^T Z)^{-1} Z^T x_i\right]^T \left[Z(Z^T Z)^{-1} Z^T x_j\right]$$

$$< \hat{x}_i, \hat{x}_j >= x_i^T Z(Z^T Z)^{-1^T} Z^T Z(Z^T Z)^{-1} Z^T x_j$$

$$< \hat{x}_i, \hat{x}_j >= x_i^T Z(Z^T Z)^{-1} Z^T x_j$$

To transform this using a kernel k, we simply replace x_i and z_i with the representers $k(., x_i)$ and $k(., z_i)$ respectively. Then $x_i^T Z = K_{x_i,z}$, $Z^T x_j = K_{x_j,z}^T$ and $Z^T Z = K_n$.

$$< \hat{x}_i, \hat{x}_j >= K_{x_i,z} {K_n}^{-1} K_{x_j,z}^T$$

This result states that the indirect kernel is the inner product in Reproducing Kernel Hilbert Space $\mathscr{H}$ between the representers of anomalous samples projected to the space spanned by the representers of normal samples. Given our assumptions regarding the heterogeneity among anomalies, the space spanned by any subset of anomalous samples will be smaller after the projection onto the normal samples, which are less heterogenous. In particular, the smallest sphere enclosing the projected representers will be smaller, and from results such as the Vapnik-Chapelle support vector span rule [82], classifiers built from this projection will be more robust and stable.

5.4 Results

5.4.1 Classification with cardiotocogram data

We obtained a popular machine learning benchmarking dataset from the UCI machine learning repository. This is a portugese cohort of 2126 fetal cardiotocograms (CTGs) that were automatically processed and the respective diagnostic features measured. The CTGs were also classified by three expert obstetricians and a consensus classification label assigned to each of them. There are 23 features, out of which two features can be used as target classes: fetal state and fetal heart rate pattern code. The former takes three possible values: normal, suspect and pathologic, which makes it a suitable target for anomaly classification. The latter feature takes an interget value from 1 to 10, and we use this to demonstrate the utility of the aSVM for support vector based regression (see 'Regression with cardiotocogram data' section below). The remaining 21 features are used for training the SVM.

For classification using the fetal state, we selected the 'normal' category as the normal class ($n = 1655$) and selected the 'suspect' ($n = 295$) and 'pathologic' ($n = 176$) as the anomaly classes.

We used two support vector based classifiers for comparison with the aSVM: the regular SVM classifier (used for binary classification) which does not use normal data, and the one-class SVM, which uses only normal data for training the SVM. The linear kernel was used for all three SVM methods. With each method we performed five-fold cross validation over the anomalous data and calculated the area of the

receiver operating characteristic curve (AUC). This process was repeated five times, where for each iteration the data was randomly divided into five folds. The AUC values obtained by each classifier for each iteration and each fold are shown in figure 5.1. While the regular binary SVM classifier provides a higher AUC than the one-class SVM for almost all of the tests, for a majority of test cases the aSVM provides an even higher AUC than the regular SVM.

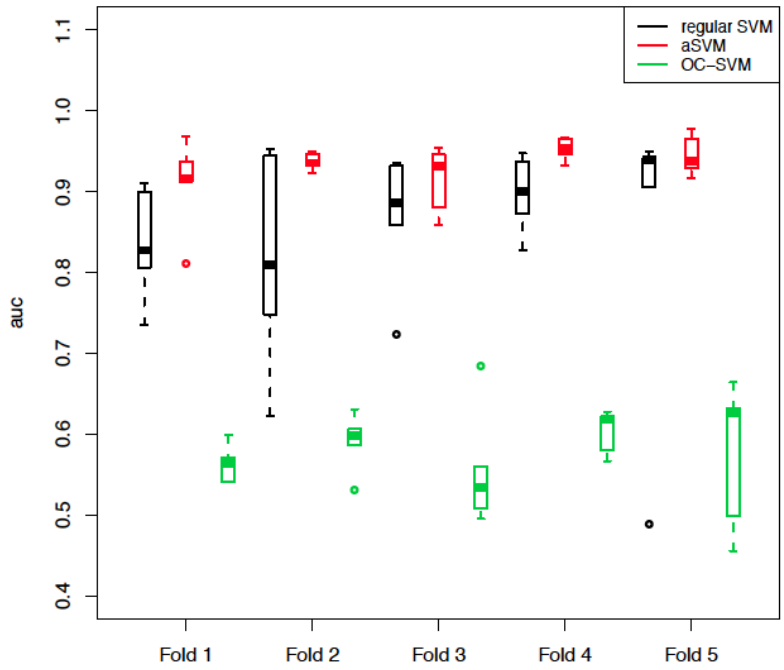

Figure 5.1: **SVM classification of cardiotocogram data.**
Distribution of AUC values from five randomized five fold cross-validations of the cardiotocogram data.

5.4.2 Classification with connect-4 data

We also obtained another dataset from the UCI machine learning repository, the connect-4 game dataset. The samples in this dataset represent states of a game board with 42 squares, with each position being either blank or being held by one of

two players, thus making the state of each square a categorical variable with three possible values. The target value for each board state indicates the winning state for one of the players: i.e. the targets are 'win', 'loss' and 'draw'. Thus we can use the 'draw' ($n = 6449$) state as the normal class and apply the anti-profile method to perform classification between the 'win' ($n = 44473$) and 'loss' ($n = 16635$) states.

For testing the svm methods, we randomly sampled training and testing subsets with resampling from both the normals and the anomalous classes. From each class, we sampled subsets of size 500 and 1000. This process was repeated for a 100 samplings of training and testing subsets. For each sampling, the regular SVM, the aSVM and the one-class SVM was fitted to the training data and the AUC values were calculated from the decision values obtained for the testing data. The results, show in Figure 5.2 indicate that the aSVM performs better than the regular SVM and the one-class SVM performs much poorly in comparison with the other two methods. The results hold well irrespective of the number of samples drawn from each class.

5.4.3 Regression with cardiotocogram data

The fetal cardiotocogram dataset mentioned above also contains a fetal heart rate attribute, ranked from 1 to 10. This attribute can also be considered an anomaly indicator, and therefore we select samples with fetal heart rate code 1 as the normal class ($n = 384$) while the remaining samples (codes 2 to 10, $n = 1742$) are selected as anomalies.

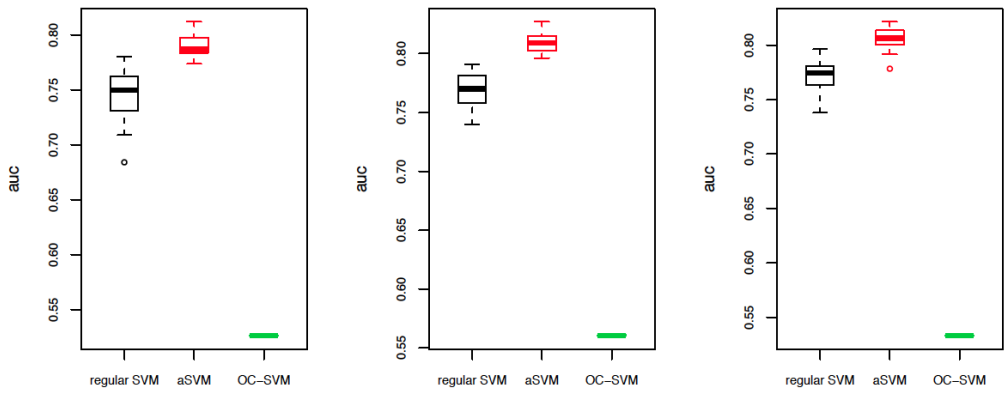

Figure 5.2: **SVM classification of Connect-4 game data.** Distribution of auc values for 100 test cases from the regular SVM, the apSVM, and the one class SVM. Number of samples randomly selected for a test case are: (Left)500 from each anommaly class, 500 normals. (Middle)1000 from each anomaly class, 500 normals. (Right)1000 from each anomaly class, 1000 normals.

Further performed SVM based regression with the regular SVM and the aSVM with the fetal heart rate code as the target variable. The radial basis kernel was used for these experiments, and 20% of the anomalous data was used for selecting the kernel parameter and the cost parameter for each SVM method via cross-validation. From the remaining anomalous data, 100 training and testing subsets were selected with resampling, with 60% of the data being used for training and the remaining 40% for testing. Figure 5.3 shows the distribution of the root mean squared error (RMSE) of the decision values obtained for the testing data from each method. The aSVM demonstrates better performance than the regular SVM, with the former producing a lower RMSE for all test cases.

5.4.3.1 Lung cancer survival data

We applied the aSVM to analyze lung cancer survival. Here we tested the universal anti-profile from [11] with a microarray lung cancer dataset containing

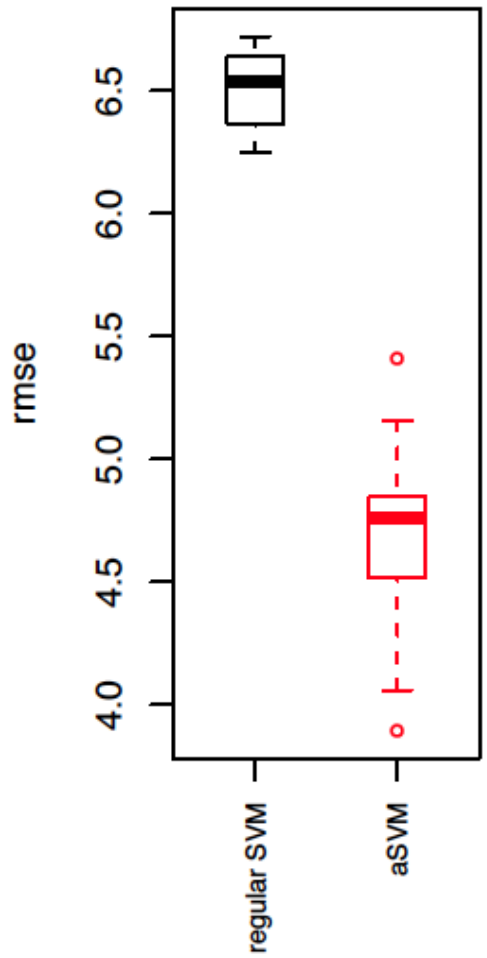

Figure 5.3: **SVM Regression analysis with cardiotocogram data.** Distribution of root mean squared error from SVM based regression performed on cardiotocogram data.

patient survival information based on patient relapse; the primary dataset containing both normal and tumor samples [55].

We stratified the samples into high risk and low risk based on patient relapse within 5 years. We used the 100 most significant genes obtained from the universal anti profile with a radial basis kernel for the aSVM. The kernel function used by the aSVM uses the normal samples along with a randomly selected set of high risk and low risk tumor samples as training data. The training and testing process was run

100 times and the results were aggregated.

With the different SVM formulations, we ran experiments to compare it with the results obtained from the anti profile scoring method. For each SVM, we used the decision values provided for the tumors of the testing set to rank the tumors and divide them to two equal sized groups, with one group containing the upper half of the decision values. Using the survival information available, we used a log-rank test to calculate the survival difference between the two groups based on patient relapse. The aggregated results are shown in Figure 5.4. The single class SVM and the aSVM provide higher survival differences than the regular SVM, indicating that the incorporation of normal samples leads to results that are more relevant biologically. The highest log-rank statistic is provided by the anti profile scoring method, indicating that the methodology provided by the anti profile scoring is of high biological relevancy.

To evaluate the generalization and robustness of the classifiers, we used the SVM decision values and the true classifications to obtain an AUC(Area Under ROC Curve) value for all SVMs (Figure 5.4). Given the unbalanced class sizes, and considering that the sensitivity of the classifier is more important than accuracy, the AUC provides a better criteria for evaluating the different classifiers. In addition, the ROC curve is an evaluation of the robustness of the generalization ability of a classifier as well, and therefore the AUC functions as an evaluation of the robustness of the SVM as well.

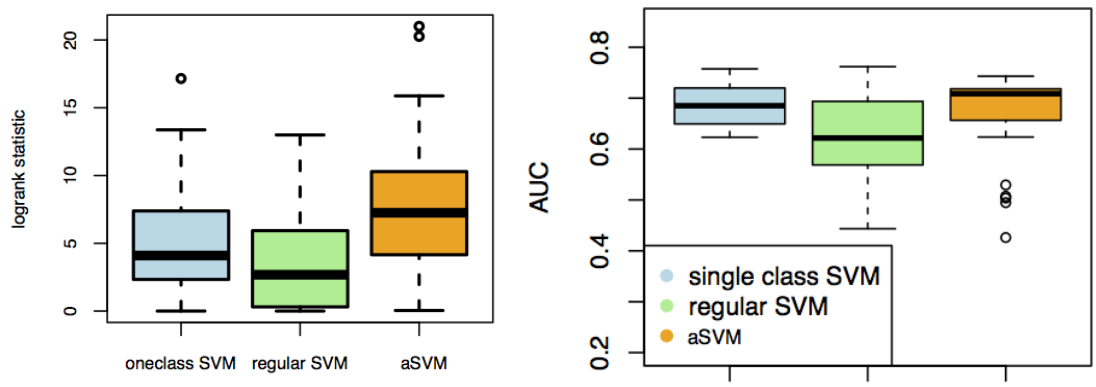

Figure 5.4: **Binary classification between high risk and low risk tumors using three SVM formulations.**

Results for a 100 test cases. (Left) Distribution of logrank scores. (Right) Distribution of AUC values.

5.4.4 Thyroid methylation data

We applied the aSVM for DNA methylation data obtained by Hansen et al [30]. The experiment was performed using a custom nucleotide-specific bead array on the Illumina GoldenGate platform to analyze 151 colon cancer-specific differentially DNA-methylated regions (cDMRs). The data contains DNA methylation levels for 384 probesets and 36 cancer, 28 adenoma and 15 normal thyroid tissue samples. Plotting the two primary principal components show the expected dispersion pattern of the cancers and the adenomas(figure 5.5).

We used F-tests and t-tests to select probesets; for the regular SVM we calculated t-statistics between cancer and adenoma groups, and for the single class SVM and aSVM we used F-statistics between normals and tumors to select probesets. We ran each SVM multiple times, each time dividing the tumors to two groups as

the training and testing set. Due to the small class sizes, the hyper parameters for the SVMs were chosen by using a 2-fold cross validation over the training set to maximize AUC. After training, we tested the predictive ability of the classifier by calculating the AUC over the testing set.

The aSVM was able to provide a better fit in terms of both accuracy and AUC in comparison with the regular SVM, and in particular the use of F-tests between tumors and normals proved to be a better method of probeset selection rather than the use of t-tests between the adenoma and cancer tumor classes. In addition, similar to previous experiments, the aSVM used a smaller number of support vectors in comparison with the regular SVM, indicating that the projection of the tumor samples onto the space of normals provides a more stable boundary that can be defined by a smaller number of tumors.

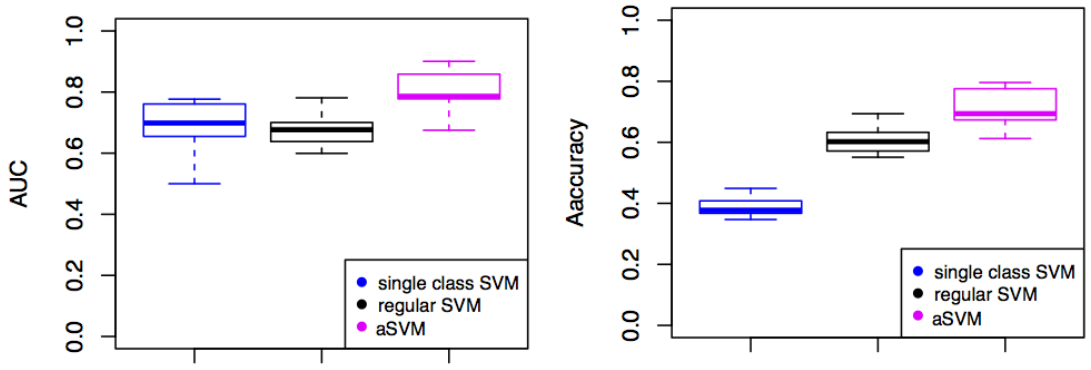

Figure 5.5: **Thyroid methylation data classification.**

Results obtained from a 100 test cases.

5.5 Methods

We used the R packages svmpath and kernlab for running experiments. For single class SVM experiments, the kernlab packaged was used where the ksvm method was used for training the SVM with the type parameter set to "one-class" for single class SVM fititng. For selecting the cost parameter and the gaussian kernel parameter(sigma), we used a grid search: for a large number of pre-selected cost and sigma values the single class SVM was trained, while a small sample of randomly selected normals was set aside without being used for fitting the SVM. For a trained SVM for a particular combination of parameters, the sum of the decision values obtained for the normals set aside was calculated, and the parameter combination that produced the largest sum was selected. For initial comparisons with the anti profile scoring method, we did not use tumor samples in the process of fitting the single class SVM. This allowed us to make a full comparison between the results obtained for all the tumor samples from the single class SVM. In the subsequent experiments, for comparisons between the single class SVM, regular SVM and the aSVM, we used a separate, randomly selected set of low and high risk tumors were used to calculate a cutoff decision value using this fitted SVM - i.e. once the single class SVM was fitted, the decision values for these tumor samples were obtained, and the mean decision value for each of the high and low risk groups were calculated, and the mid point between those two values was used as the cutoff decision value between the high risk and low risk groups.

For the comparisons between the different SVM formulations, we used different

methods for probeset selection. While the aSVM used the universal anti profile probeset, the single class SVM used a t-test between the tumor and normal samples used for training and tuning, while the regular SVM used a t-test between the two tumor classes. All methods selected the 100 highest ranked probesets.

The svmpath tool provides a fitting for the entire path of the SVM solution to a model at little additional computational cost. For the regular SVM and aSVM experiments, the svmpath packaged was used, with a new kernel function coded for the aSVM. The training data was divided to two sets, one of which was used for fitting the SVM and the other used for tuning the cost and kernel parameters. Using a grid search, the SVM was fitted for a number of parameter combinations, and the fitted SVM was tested on the tuning set. The parameters that produced the largest AUC value for the tuning set was selected, and this fitting was subsequently used on the testing set. In addition, the percentage of the training data which become the support vectors was obtained as well.

5.6 Conclusion

We have introduced the anomaly SVM as a novel algorithm to address the anomaly classification problem. We have shown that under the assumption that the classes we are trying to distinguish with a classifier are heterogeneous with respect to a third stable class, we can define a Support Vector Machine based on an indirect kernel using the stable class. We have shown that the dual of the aSVM optimization problem is equivalent to that of the standard SVM with the addition of

an indirect kernel that measures similarity of anomalous samples through similarity to the stable normal class. Furthermore, we have characterized this indirect kernel as the inner product in a Reproducing Kernel Hilbert Space between representers that are projected to the subspace spanned by the representers of the normal samples. This led to the result that the aSVM will learn classifiers that are more robust and stable than a standard SVM in this learning setting. We have shown by simulation and application to cancer genomics datasets that the anomaly SVM does in fact produce classifiers that are more accurate and stable than the standard SVM in this setting.

While the motivation and examples provided here are based on cancer genomics we expect that the anomaly classification setting is applicable to other areas.

The characterization of the indirect kernel through projection to the normal subspace also suggests other possible classifiers suitable to this task. For instance, by defining a margin based on the projection distance directly. Furthermore, connections to kernel methods for quantile estimation will be interesting to explore.

Chapter 6: Anomaly Based SVM Regression for Survival Analysis

6.1 Introduction

Clinical trials are an essential aspect of medical research, and therefore the analysis of the data that is obtained from such data forms an integral part of medical and healthcare research [48]. A core feature of many clinical trials is the analysis of disease related characteristics of a patient during a certain time period following the enrollment in the trial by the patient. In trials studying the effect of a drug aimed at life-threatening diseases this may involve recording the time period each patient survives while being treated with a certain drug (or alternatively a placebo). This is a common feature of oncology research trials, and for diseases such as cancer, apart from this information trials also typically record whether the patient exhibited relapse of the disease after surgery and medication.

We refer this type of data as time-to-event data, where the event usually is taken to be either death of the patient or relapse of the disease after initial treatment. However, given the multitude of logistical issues that are an inherent part of such research, it is almost always the case that the recorded data will be partial - i.e. it can be expected that there will be many instances of incomplete records. It is quite common for a patient to terminate his or her involvement with the trial at any point

during its course, and therefore in lieu of a record indicating the death or disease recurrence of the patient there may simply be a record that the patient is no longer under observation. Such incomplete data is referred to as censored data, where the censored information is the the survival or disease status of the patient. It may be that the patient exhibited recurrence soon after withdrawing from the trial, or it may be that the patient remained healthy for a significantly long time period after the withdrawal - any number of such scenarios are possible, but we cannot know this information. Since the censoring of information has taken place at the end of our observation period, this is more specifically referred to as right-censored data.

It is obvious that the analysis of such data cannot be performed using the same methods and techniques that are commonly used to analyze complete (i.e. non-censored) data. Given the unique challenges presented, the development of specific methods for analyzing such data has been an important part of Biostatistics and Computational Biology research. Here we take up one such method, the survival SVM proposed by Evers and Messow [18], and extend it to employ support vector machine based anomaly classification methods we have introduced previously. We demonstrate that by utilizing normal data in addition to patient data, it is possible to produce more robust predictive tools for the analysis of right censored data.

6.2 Background

Right censored data is typically available in the form of a number of clinical covariates that characterize the health condition of the patient with regard to the

disease in question, and in addition there are two variables that indicate the time-to-event information. Whereas in an uncensored information setting this information may be indicated with a single real value (e.g. time to death) or a binary indicator (e.g. diseased or not) with right-censored data this information is indicated with a non-negative real *time* variable and a binary *event* variable. The former indicates the time taken to observe the event mentioned, and the latter indicates the censored or uncensored status of said event. For example if the event in question is relapse of cancer, then $event = 1$ indicates that the patient relapsed at the time indicated by the corresponding *time* variable, and $event = 0$ indicates that the patient was no longer under observation after this time.

The simplest method for analyzing such data is to stratify the data in such a manner that it conforms to the same format as uncensored data, an approach we have employed in past chapters. This allows the data to be analyzed with any regular statistical and computational tool available following this transformation. Usually, a certain cutoff time value is decided upon, and the data is stratified as either high risk level or low risk level based on the cutoff. Suppose the event considered is an adverse condition such as recurrence of cancer or death of the patient. If we select 5 years as our cutoff time period, patients that survived beyond 5 years without exhibiting said event are considered to be in the 'low risk' category, while patients that exhibited relapse, or patient that died within 5 years considered to be in the 'high risk' category. Thus the trial data have now been transformed into a regular binary classification setting where each patient is simply described as high risk or low risk. If the event observed is a desirable one, such as a patient returning to fully

healthy condition or the complete eradication of disease indicators, then in a similar scenario patients that are fully cured within 5 years would be considered low risk while patients who are not cured within 5 years would be considered high risk.

It is obvious that there are a number of drawbacks to such a stratification, which are a result of the significant loss of information involved in such a transformation. The stratification does not assign one of the risk level labels to all patients. Patients that were censored within the cutoff time period cannot be assigned either the low risk label or the high risk label, and therefore have to be removed from further consideration in the analysis. In addition, the stratification treats patients that exhibit different disease characteristics as being similar. For example, considering relapse of cancer under a 5 year cutoff would mean that a patient that relapsed at 4.9 years and a patient that relapsed at 5.1 years would be considered as being in different categories, while a patient that relapsed at 5.1 years and a patient that was censored at 12 years would be considered to be in the same category. If the cutoff is chosen at a late time period within the timeline of the clinical trial, more observations can be included in the final labeling from the stratification but would include a large number of patients in one category, resulting in a highly imbalanced categorization. Thus the choosing of the cutoff period entails a number of considerations as well. Overall, such a stratification is only reasonable if a large number of patients that exhibited the event did so within the cutoff while a a large number of patients that were censored are located after the cutoff. These are unrealistic assumptions for most clinical trials. It is for these reasons that methods specifically designed for analyzing right censored data need to to be developed.

Support vector machines (SVMs) have been introduced in the previous chapters of this dissertation. As a short recapitulation, we note that SVMs are one of the most common off-the-shelf machine learning tools employed at present. Usually employed as a binary classifier, SVMs are considered maximum margin classifiers since they are designed to compute the hyperplane which separates two classes in the best possible manner. However SVMs present two advantages that are not typically found in other classifiers: the ability to kernelize, and the ability to produce a sparse solution to the learning problem. The former allows the projection of the data to a higher dimensionality before computing the separating hyperplane with no additional cost via utilizing kernel functions, this producing more efficient classifiers in comparison with many other learning methods that cannot be kernelized are thus limited to be used in the feature space of the problem domain. The latter feature is derived from the fact that usually the separating hyperplane is dependent on a few data points (which are referred to as the support vectors).

While the most common form of the SVM is aimed at binary classification, other forms of SVMs have been devised for regression and one-class detection. We have presented the derivation of these SVMs in previous chapters. As mentioned previously, SVMs have been used successfully in many domains within Computational Biology. However, similar to many other off-the-shelf classifiers it cannot be used to directly analyze right-censored survival information and it is for this reason that a SVM directly aimed at performing survival regression of such data, the survival SVM has been developed. In the next section we present the survival SVM and how to extend it to function as an anomaly based survival regression tool according

to the principles discussed in previous chapters.

6.3 Derivation

We briefly re-state the derivation of the binary SVM classifier as follows: consider a data set consisting of N pairs $(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)$ with $x_i \in \mathbb{R}^p$ and $y_i \in \{-1, 1\}$. If the given two classes are class A and class B, then for any data point i, $y_i = -1$ indicates that the point belongs to class A and $y_i = 1$ indicates that the point belongs to class B. Then the primal optimization function for computing the maximum margin classifier can be written as:

$$\begin{aligned} \underset{\beta,\beta_0}{\text{minimize}} \quad & \frac{1}{2}||\beta||^2 + C\sum_{i=1}^{N}\xi_i \\ \text{subject to} \quad & y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i,\ \xi_i \geq 0\ i = 1..N. \end{aligned}$$

Here ξ_i are the slack variables added to handle non-separability and C is the penalty imposed to limit the amount of error allowed via the slack variables. The hyperplane derived from this optimization function can be written as $f(x) = x^T\beta + \beta_0 = 0$.

The dual of this optimization problem is

$$\begin{aligned} \underset{\alpha_i}{\text{maximize}} \quad & \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j < x_i.x_j > \\ \text{subject to} \quad & 0 \leq \alpha_i \leq C\ \ i = 1..N. \\ & \sum_{i=1}^{N}\alpha_i y_i = 0 \end{aligned}$$

It is obvious that this formulation in its current form cannot be used for censored survival data. In order to derive an SVM for such data, rather than use

individual data points (i.e. patients), Evers and Messow [18] transform the data to use pairs of samples between which a survival based relationship can be established. Following the same principles used in the Cox proportional hazards model, we choose a patient who exhibits the observed event at a given time t (say patient p_1) with a patient who survives beyond that time (say patient p_2) to form a pair. It is now possible to obtain survival relationship between these two patients: $p_1 < p_2$. That is, patient p_1 is at a higher risk level than patient p_2.

To form all such pairs possible from a survival dataset, we can choose all patients who exhibited the event as p_1. If the corresponding survival time for p_1 is t_1, then for each such patient, all other patients who either exhibited the event or were censored at a time $t > t_1$ can be chosen for p_2. We can alternatively state this as: let D be the set of all patients who have died (or relapsed). For all $i \in D$, we can derive a set $R_{t_i}^+$ comprising of individuals who are at risk at time of death of patient i (time t_i).

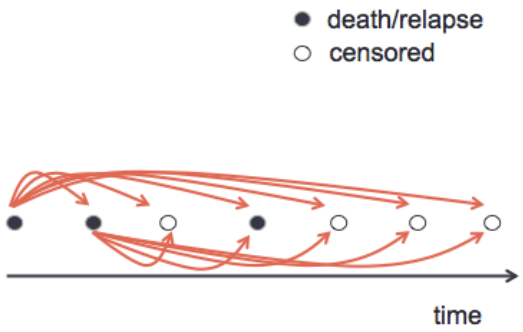

Figure 6.1: **Survival regression formulation.**
Selection of sample pairs between which a survival based relationship can be established.

For every (i, j) pair constructed from $J = (i, j) : i \in D \wedge j \in R_{t_i}^+$, x_i is at

higher risk and x_j (Figure 6.1). If the decision function we aim to obtain from our SVM is $f(x) = x^T\beta + \beta_0 = 0$, we require $f(x_i) \geq f(x_j)$ - i.e. $x_i^T\beta - x_j^T\beta \geq 0$. Therefore for each (i, j) pair we can add a constraint

$$x_i^T\beta - x_j^T\beta \geq 1 \;\; \forall (i,j) : i \in D \wedge j \in R_{t_i}^+$$

for our primal optimization problem. With the addition of slack variables this becomes:

$$x_i^T\beta - x_j^T\beta \geq 1 - \xi_{i,j} \;\; \xi_{i,j} \geq 0$$

Then our main optimization problem is:

$$\underset{\beta,\beta_0}{\text{minimize}} \quad \frac{1}{2}||\beta||^2 + C\sum_{i\in D}\sum_{j\in R_{t_i}^+}\xi_{i,j}$$

$$\text{subject to} \quad \beta^T x_i - \beta^T x_j \geq 1 - \xi_{i,j}, \;\; \xi_i \geq 0 \;\; i = 1..N, \;\; j \in R_{t_i}^+.$$

Similar to the regular SVM derivation (see chapter 3), we can derive the dual of this problem:

$$\underset{\alpha_i}{\text{maximize}} \quad \sum_{(i,j)\in J}\alpha_{ij} - \frac{1}{2}\sum_{(i,j)\in J}\sum_{(k,l)\in J}\alpha_{ij}\alpha_{kl}K_{ij,kl}$$

$$\text{subject to} \quad 0 \leq \alpha_{ij} \leq C \;\; (i,j) \in J$$

Here $K_{ij,kl}$ is the kernel function evaluated between the sample pair (i, j) and the pair (k, l), and the α_{ij}, α_{kl} are the Lagrangian multipliers introduced similar to the regular SVM formulation. Since $K_{ij,kl}$ is obtained through the product of (i, j) and (k, l) pairs, we can write the linear kernel expression of $K_{ij,kl}$ as follows:

$K_{ij,kl} = x_i^T x_k - x_i^T x_l - x_j^T x_k + x_j^T x_l$. Thus $K_{ij,kl}$ can be computed from the kernel matrix of the individual sample data, and thus no additional step is necessary for kernelization.

Since the survival SVM (SSVM) can be computed using any kernel matrix used for the regular SVM, we can also utilize the anomaly kernel previously derived for this purpose, where $K_{ij,kl}$ is simply calculated from the gram matrix computer according to the anomaly kernel. We refer to such a formulation as the anomaly survival SVM (A-SSVM).

It is also possible to formulate a survival SVM for anomaly regression by including normal samples directly in the data. Considering the constraint formulation of the survival SVM, since forming a pairwise constraint of the form $x_i^T \beta - x_j^T \beta \geq 1 - \xi_{i,j}$ where a risk level ordering can be established between samples i and j, normal samples can also be included since it is possible to formulate a risk level ordering between any normal sample and any non-normal (i.e. censored or un-censored survival data) sample (Figure 6.2). Since the (i, j) pair consists of a sample i which is of higher risk than sample j, if we consider a tumor sample i and a normal sample k the relevant constraint would be as follows:

$$x_i^T \beta - x_k^T \beta \geq 1 - \xi_{i,k} \;\; \xi_{i,k} \geq 0$$

Here i could either be censored or uncensored. We refer to inclusion of both normals and anomalies in the survival SVM in this way as the ordinal anomaly SVM (OA-SSVM).

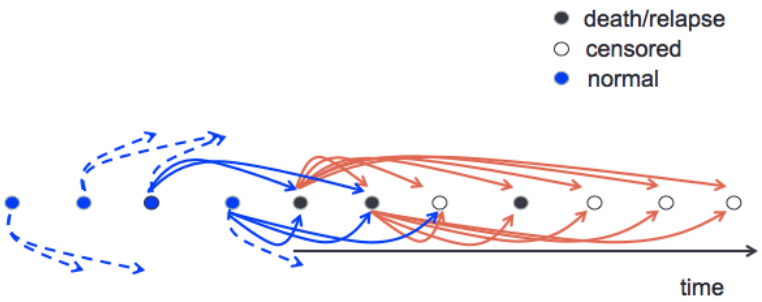

Figure 6.2: **Survival regression formulation with normals included.** Inclusion of normals in the selection of sample pairs between which a survival based relationship can be established.

6.4 Results

6.4.1 Application to microarray data

We applied the different survival SVM methods to lung cancer and breast cancer microarray data utilized previously (identification numbers GSE31210 and GSE1456 respectively) [55, 58]. The samples were selected from the same curated datasets that were used in the anti-profile experiments discussed in chapter 3. Along with the microarray samples, this data also contained survival data: the event considered was relapse for lung cancer data and patient death for breast cancer data. In addition we also utilized the universal anti-profile signature used with the anti-profile experiments; we selected the 100 most hypervariable probesets from the universal anti-profile signature corresponding to the two microarray platforms from which the lung and breast cancer data were obtained from (GPL570 and GPL96 respectively). In the following experiments, the feature space for our genomic data consisted of these probesets unless otherwise mentioned. To assess the performance of the SVM methods, we utilize the Concordance index (C-index) [6] and the logrank

test [45], which are standard tools for analyzing survival data.

6.4.2 Selection of samples

Given that the optimization contains a kernel matrix which can become quite large from the pairwise inclusion of samples, we experimented with multiple methods of selecting subsets of data that may yield comparatively good performance levels. To select such data, we arrange our samples in increasing order of survival time t. We refer to samples with event $= 0$ as the censored group and samples with event $= 1$ as the un-censored group, and select n samples from each. For a given n where $n < N$, N being the total number of samples available, we chose samples in the following ways:

1. **Method 1:** Select n samples at highest risk (i.e. n un-censored samples with the lowest survival times) and n samples with lowest risk (i.e. n censored samples with the highest survival times).

2. **Method 2:** Select n un-censored samples and n censored samples randomly.

3. **Method 3:** From un-censored samples, select $\frac{n}{2}$ samples with the highest survival times and $\frac{n}{2}$ samples with the lowest survival times. Select n samples from the censored group in the same manner.

4. **Method 4:** Select n samples from each group that are evenly distributed when ranked along survival times.

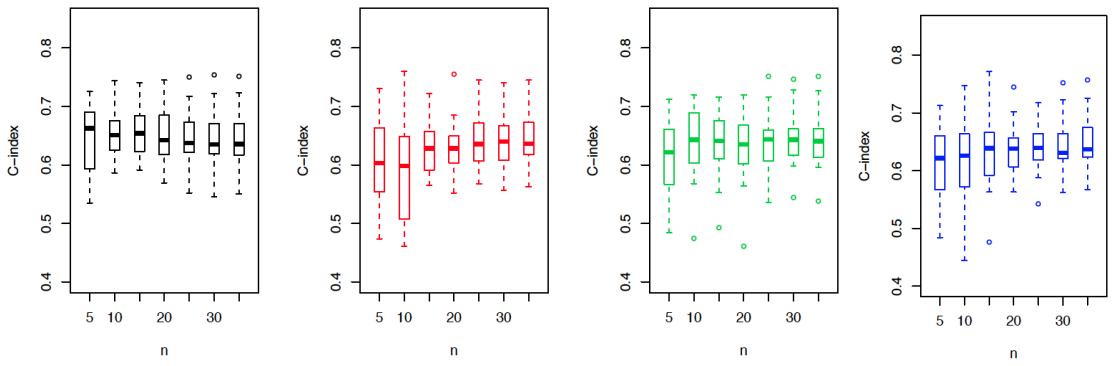

Figure 6.3: **Sample selection with lung cancer microarray data.** Effect of selection of samples on the survival SVM performance, Methods 1-4 (left to right).

We chose lung cancer and breast cancer microarray datasets to examine the behavior of the survival SVM with respect to sample selection. At this stage of our analysis we limited our experiments to the regular survival SVM and did not include normal samples. We generated a set of training and testing subsets from each dataset with resampling, where the training vs. testing sample distribution was 75%-25% from the total dataset. For each sample selection method we fitted a regular survival SVM for increasing n where $n \in \{5, 10, 15, 20, 25, 30, 35\}$. Figure 6.3 shows the results obtained, where the C-index was measured for the decision values obtained for the testing data from each fitted SVM. We saw that all sample selection methods except for method 1 perform increasingly well with n, though beyond the $n = 20$ boundary increasing the number of samples selected did not yield any increase in performance. Random sampling produces a slightly higher variability in performance with low values of n, which is expected, though with increasing n the performance behaves similarly to the other methods. Method 1 performs quite efficiently even with low levels of n.

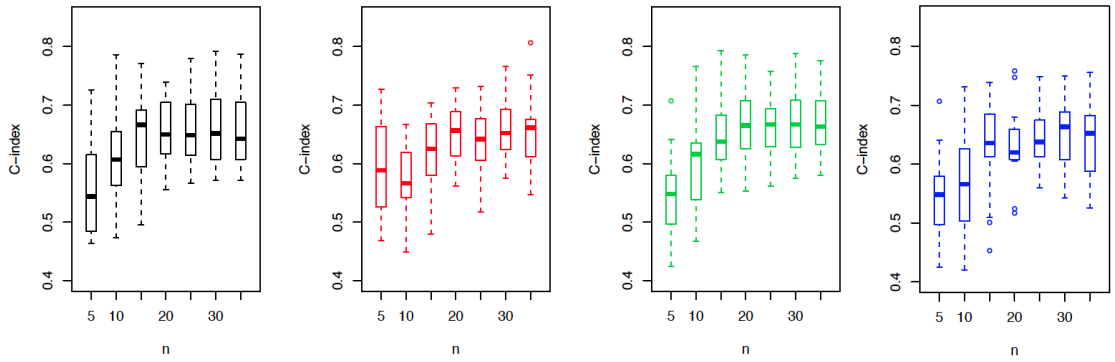

Figure 6.4: **Sample selection with breast cancer microarray data.** Effect of selection of samples on the survival SVM performance, Methods 1-4 (left to right).

A similar analysis was conducted with the breast cancer data. The results obtained (Figure 6.4) show a similar characterization of performance with respect to sample selection method and size. The effect of low values of n is more pronounced in this dataset, as all methods show increasing performance with increasing n. However, similar to our observations with the lung cancer data, we see that beyond the $n = 20$ boundary there is little variation in performance, both in terms of varying n and varying method. The first method seems to perform most consistently.

These observations suggest that the survival profile we seek from the survival SVM optimization is largely determined by samples exhibiting the highest and the lowest risk levels, as these samples are selected from start by Method 1, and we have observed the most consistent performance from Method 1 with respect to n. This particularly seems to be the case for the lung cancer data, whereas for the breast cancer data additional samples are necessary for fully characterizing this survival regression function. Based on these observations, we limited further analysis to Method 1 in terms of sampling selection.

6.4.3 Survival analysis with normals included

In the next stage of our analysis, we compared the performance of the regular survival SVM (SSVM) with the two anomaly classification survival SVMs we derived in the previous section. We refer to the survival SVM with the anomaly kernel included as the anomaly survival SVM (A-SSVM) and we refer to the derivation of the survival SVM provided in the previous chapter with normal-anomaly pairs included as the ordinal anomaly survival SVM (OA-SSVM).

We compared the performance of these three SVMs based on the lung cancer and breast cancer microarray datasets mentioned in the previous section. We applied the different SVMs to the lung cancer microarray data referred to above. We formed 100 randomly selected subsets of training and testing data with resampling, using a 60% - 40% division of training and testing subsets respectively. A similar experiment was carried out with the aforementioned breast cancer dataset as well.

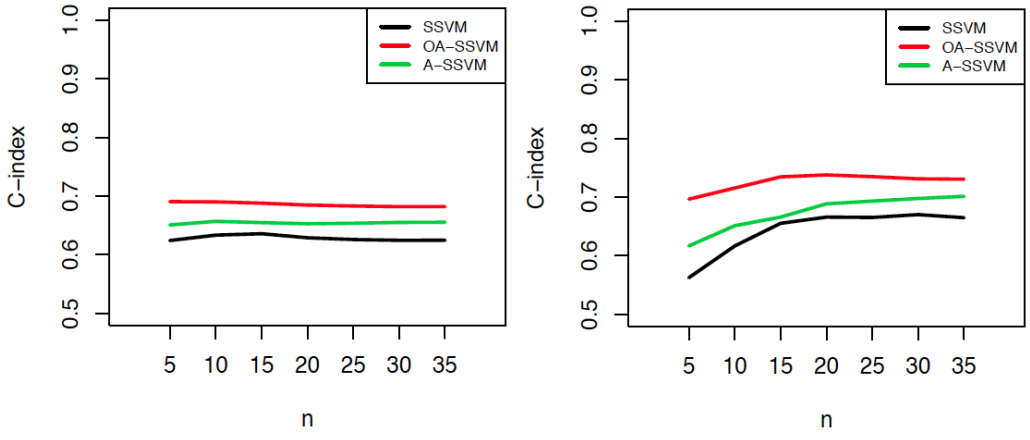

Figure 6.5: **Cancer microarray data analysis.**
Comparison of performance of different survival SVM methods on (Left)Lung and (Right)Breast gene expression data.

Figure 6.5 shows the mean C-index obtained over the training and testing

subset pairs curated from the main dataset. From all three SVM formulations we see that the C-index yields better predictive performance with increasing n for the breast cancer dataset while for the lung cancer dataset even a small number of samples is adequate. However, even for the breast cancer data, this behavior is less pronounced for the anomaly SVM formulations; in particular the OA-SSVM is able to provide a high level of C-index even with very small n. In addition we see that both anomaly SVMs provide a higher C-index than the regular survival SVM, regardless of the choice of n. Between the two anomaly SVMs, we see that the OA-SSVM in particular performs better than both other SVMs. These characterizations hold for both lung cancer and breast cancer data, and it demonstrate that the inclusion of normals can yield more efficient survival regression machines.

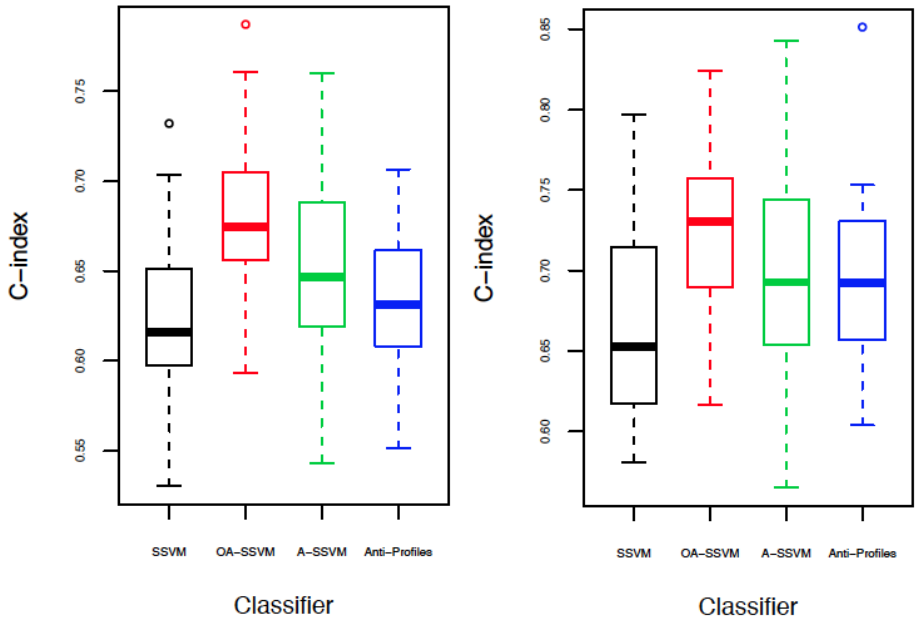

Figure 6.6: **Comparison of SVM performance with the anti-profile method.** Comparison of performance of different survival SVM methods on (Left)Lung and (Right)Breast gene expression data with the anti-profile scoring method.

In addition to comparing the regular survival SVM with the anomaly survival

SVMs we also compared the results obtained with the anti-profile scores calculated previously for this data. Treating the anti-profile scores in the same manner as we would the decision values obtained from an SVM, we calculated the C-index for each training and testing subset pair. A comparison of these results against the SVMs are shown in Figure 6.6 for the lung cancer data and for the breast cancer data. For both the lung cancer data and the breast cancer data we selected $n = 30$ sample size for all the SVM formulations.

For both datasets, while the SSVM slightly underperforms in comparison with the anti-profile method, the anomaly SVMs perform better with respect to the C-index. In the breast cancer data, the A-SSVM has a similar level of performance in comparison with the anti-profile scores, but for both datasets the OA-SSVM performs well above the other methods (Lung cancer data: A-SSVM against Anti-Profiles, rank-sum test $p = 0.18$, OA-SSVM against Anti-Profiles, $p = 0.001$; Breast cancer OA-SSVM against Anti-Profiles, $p = 0.05$). This further validates our observations from the one-class SVM in previous chapters. In addition to the C-index we caluclated the logrank statistic and associated p-value as well. These results can be seen in Figure 6.7 and Figure 6.8.

6.5 Conclusion

In this chapter we have presented how anomaly classification and regression can be utilized for analysis of survival data. In particular, we have developed two anomaly regression SVMs for censored survival regression by extending the survival

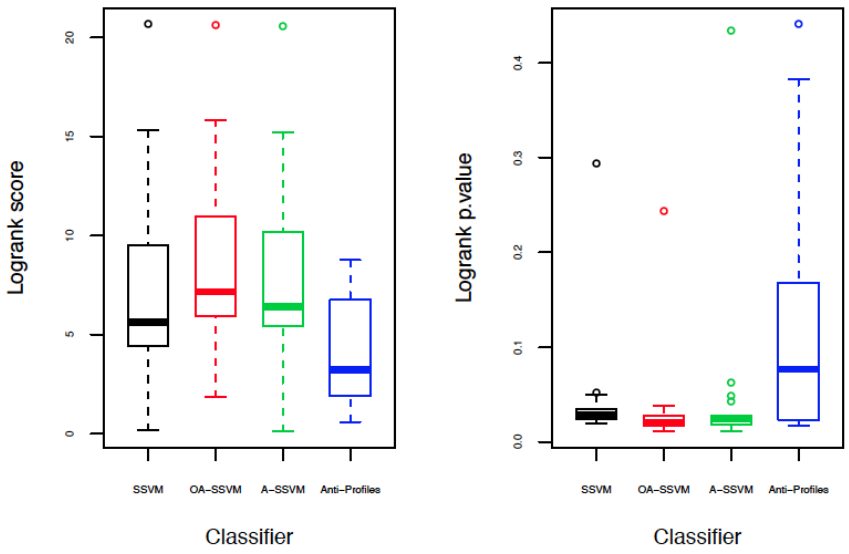

Figure 6.7: **Lung cancer microarray data: performance comparison.** Comparison of SVM performance and anti-profile scores using the logrank statistic for lung cancer data.

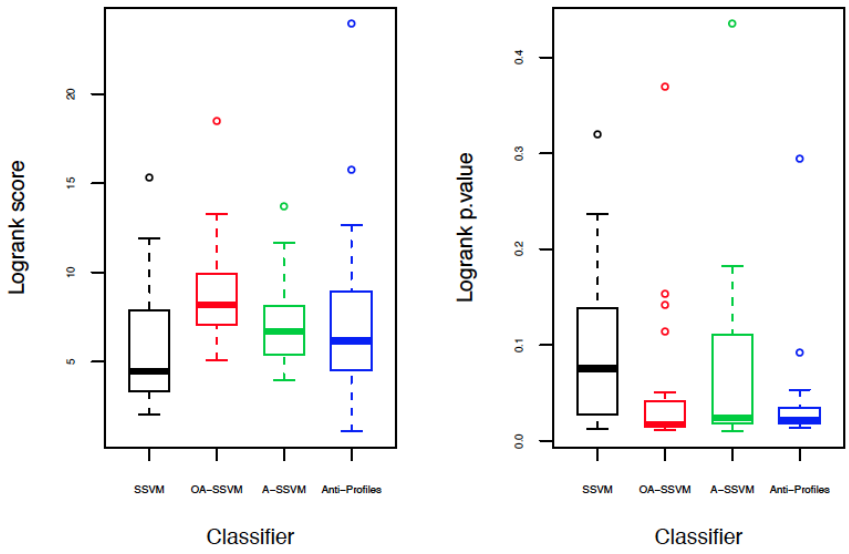

Figure 6.8: **Breast cancer microarray data: performance comparison.** Comparison of SVM performance and anti-profile scores using the logrank statistic for breast cancer data.

SVM. We have compared the predictive performance of these SVMs with lung and breast cancer microarray data.

The results presented here suggest that the two anomaly regression based

survival SVMs derived in this chapter can be an important tool for analyzing survival data. Our comparison of the SVM results with the previously analyzed anti-profile scoring method demonstrate that the anomaly survival SVMs have higher predictive potential than the anti-profile method.

Chapter 7: Conclusion

In this dissertation we have provided a comprehensive overview of how anomaly classificationa and regression can be utilized to address classification and regression problems arising in gene expression and methylation data, as well as censored survival data arising from clinical trials.

In the first chapter we carried out a thorough survey of recent advances made in statistical and computational analysis of gene expression and methylation variability analysis. We explored the methods used in a number of studies and discussed the biologically relevant novel results obtained by carrying out expression variability analysis. In particular we states how stochastic variability observed in gene expression and methylation data in cancer can be considered a stable characteristics, and how this has been utilized to develop a statistical cancer screening method from gene expression data: the anti-profile method.

Next, we extended the anti-profile approach devised from these principles to cancer prognosis and progression prediction. We demonstrated how the anti-profile scoring method can efficient classify between stages of tumor progression from gene expression data. By application of this methodology to data of multiple cancer types, we showed how the anti-profile method can be used for predicting the risk

levels associated with cancer prognosis. We similarly demonstrated the applicability of this technique to methylation data as well.

Following this we demonstrated how more complex machine learning methods can be adapted to perform anomaly classification and regression by focusing on support vector machines (SVMs). A novel SVM method was introduced which utilizes the same principles observed with regard to the anti-profile method. The application of this SVM method showed that by considering the deviation from a stable normal class, efficient and stable classifiers and regressors can be built for classification and regression of relatively more unstable anomaly groups.

In the final section we further advanced these techniques by application to censored survival data. Since such data is common in studies of statistical analysis of cancer, we derived two novel SVM methods for modeling survival regression with kernels as an anomaly regression problem. The utility of these methods were demonstrated by application to lung and breast cancer gene expression data.

Thus we can summarize the contributions made in this dissertation as follows:

1. Carry out a survey of recent developments in gene expression variability analysis.

2. Extend the anti-profile method for cancer progression and prognosis prediction with application to survival data.

3. Derive a novel SVM technique for classification and regression of anomalies.

4. Apply anomaly classification and regression models to derive an SVM for analyzing censored survival data.

It it our aim that anomaly classification and regression methods become more widely used in computational biology research. We hope that the ideas and methods presented here will be of assistance to further this aim.

Bibliography

[1] B Afsari, D Geman, and E J Fertig. Learning dysregulated pathways in cancers from differential variability analysis. *Cancer informatics*, 2014.

[2] Bahman Afsari, Donald Geman, and Elana J Fertig. Learning dysregulated pathways in cancers from differential variability analysis. *Cancer informatics*, 13(Suppl 5):61, 2014.

[3] Elfalem Y Alemu, Joseph W Carl, Hector Corrada Bravo, and Sridhar Hannenhalli. Determinants of expression variability. *Nucleic acids research*, 42(6):3503–3514, 2014.

[4] Gerald L Andriole, E David Crawford, Robert L Grubb III, Saundra S Buys, David Chia, Timothy R Church, Mona N Fouad, Edward P Gelmann, Paul A Kvale, Douglas J Reding, et al. Mortality results from a randomized prostate-cancer screening trial. *New England Journal of Medicine*, 360(13):1310–1319, 2009.

[5] Martin J Aryee, Andrew E Jaffe, Hector Corrada-Bravo, Christine Ladd-Acosta, Andrew P Feinberg, Kasper D Hansen, and Rafael A Irizarry. Minfi: a flexible and comprehensive bioconductor package for the analysis of infinium dna methylation microarrays. *Bioinformatics*, 30(10):1363–1369, 2014.

[6] Donald Bamber. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of mathematical psychology*, 12(4):387–415, 1975.

[7] Artem Barski, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–837, 2007.

[8] Elizaveta V Benevolenskaya. Histone h3k4 demethylases are essential in development and differentiation this paper is one of a selection of papers published

in this special issue, entitled 28th international west coast chromatin and chromosome conference, and has undergone the journal's usual peer review process. *Biochemistry and cell biology*, 85(4):435–443, 2007.

[9] Johan Botling, Karolina Edlund, Miriam Lohr, Birte Hellwig, Lars Holmberg, Mats Lambe, Anders Berglund, Simon Ekman, Michael Bergqvist, Fredrik Pontén, et al. Biomarker discovery in non–small cell lung cancer: Integrating gene expression profiling, meta-analysis, and tissue microarray validation. *Clinical Cancer Research*, 19(1):194–204, 2013.

[10] Hector Corrada Bravo, Rafael A. Irizarry, and Jeffrey T. Leek. *antiProfiles: Implementation of gene expression anti-profiles*. R package version 1.6.0.

[11] Héctor Corrada Bravo, Vasyl Pihur, Matthew McCall, Rafael Irizarry, and Jeffrey Leek. Gene expression anti-profiles as a basis for accurate universal cancer signatures. *BMC bioinformatics*, 13(1):272, 2012.

[12] Nyasha Chambwe, Matthias Kormaksson, Huimin Geng, Subhajyoti De, Franziska Michor, Nathalie A Johnson, Ryan D Morin, David W Scott, Lucy A Godley, Randy D Gascoyne, Ari Melnick, Fabien Campagne, and Rita Shaknovich. Variability in DNA methylation defines novel epigenetic subgroups of DLBCL associated with different clinical outcomes. *Blood*, January 2014.

[13] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3):15, 2009.

[14] W Dinalankara and H C Bravo. Anomaly Classification with the Anti-Profile Support Vector Machine. *arXiv.org*, 2013.

[15] Wikum Dinalankara and Hctor Corrada Bravo. Gene expression signatures based on variability can robustly predict tumor progression and prognosis. *Cancer Informatics*, 14:71–81, 06 2015.

[16] James A Eddy, Leroy Hood, Nathan D Price, and Donald Geman. Identifying tightly regulated and variably expressed networks by differential rank conservation (dirac). *PLoS computational biology*, 6(5):e1000792, 2010.

[17] Lyndsey A Emery, Anusri Tripathi, Chialin King, Maureen Kavanah, Jane Mendez, Michael D Stone, Antonio De Las Morenas, Paola Sebastiani, and Carol L Rosenberg. Early dysregulation of cell adhesion and extracellular matrix pathways in breast cancer progression. *The American journal of pathology*, 175(3):1292–1302, 2009.

[18] Ludger Evers and Claudia-Martina Messow. Sparse kernel methods for high-dimensional survival data. *Bioinformatics*, 24(14):1632–1638, 2008.

[19] Andrew P Feinberg and Rafael A Irizarry. Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proceedings of the National Academy of Sciences*, 107(suppl 1):1757–1764, 2010.

[20] Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):R80, 2004.

[21] Debashis Ghosh and Arul M Chinnaiyan. Genomic outlier profile analysis: mixture models, null hypotheses, and nonparametric estimation. *Biostatistics*, 10(1):60–69, 2009.

[22] Debashis Ghosh and Song Li. Unsupervised outlier profile analysis. *Cancer informatics*, 13(Suppl 4):25, 2014.

[23] Debashis Ghosh and Song Li. Unsupervised outlier profile analysis. *Cancer informatics*, 13(Suppl 4):25–33, 2014.

[24] Luca Gianni, Milvia Zambetti, Kim Clark, Joffre Baker, Maureen Cronin, Jenny Wu, Gabriella Mariani, Jaime Rodriguez, Marialuisa Carcangiu, Drew Watson, et al. Gene expression profiles in paraffin-embedded core biopsy tissue predict response to chemotherapy in women with locally advanced breast cancer. *Journal of clinical oncology*, 23(29):7265–7277, 2005.

[25] Thomas J Giordano, Rork Kuick, Tobias Else, Paul G Gauger, Michelle Vinco, Juliane Bauersfeld, Donita Sanders, Dafydd G Thomas, Gerard Doherty, and Gary Hammer. Molecular classification and prognostication of adrenocortical tumors by transcriptome profiling. *Clinical Cancer Research*, 15(2):668–676, 2009.

[26] Thomas J Giordano, Rork Kuick, Dafydd G Thomas, David E Misek, Michelle Vinco, Donita Sanders, Zhaowen Zhu, Raffaele Ciampi, Michael Roh, Kerby Shedden, et al. Molecular classification of papillary thyroid carcinoma: distinct braf, ras, and ret/ptc mutation-specific gene expression profiles discovered by dna microarray analysis. *Oncogene*, 24(44):6646–6656, 2005.

[27] K Graham, A de Las Morenas, A Tripathi, C King, M Kavanah, J Mendez, M Stone, J Slama, M Miller, G Antoine, et al. Gene expression in histologically normal epithelium from breast cancer patients and from cancer-free prophylactic mastectomy patients shares a similar profile. *British journal of cancer*, 102(8):1284–1293, 2010.

[28] Kelly Graham, Xijin Ge, Antonio de las Morenas, Anusri Tripathi, and Carol L Rosenberg. Gene expression profiles of estrogen receptor–positive and estrogen receptor–negative breast cancers are detectable in histologically normal breast epithelium. *Clinical Cancer Research*, 17(2):236–246, 2011.

[29] Balazs Gyorffy, Bela Molnar, Hermann Lage, Zoltan Szallasi, and Aron C Eklund. Evaluation of microarray preprocessing algorithms based on concordance with rt-pcr in clinical samples. *PloS one*, 4(5):e5645, 2009.

[30] Kasper Daniel Hansen, Winston Timp, Héctor Corrada Bravo, Sarven Sabunciyan, Benjamin Langmead, Oliver G McDonald, Bo Wen, Hao Wu, Yun Liu, Dinh Diep, et al. Increased methylation variation in epigenetic domains across cancer types. *Nature genetics*, 43(8):768–775, 2011.

[31] Gavin C Harewood, Maurits J Wiersema, and L Joseph Melton III. A prospective, controlled assessment of factors influencing acceptance of screening colonoscopy. *The American journal of gastroenterology*, 97(12):3186–3194, 2002.

[32] Trevor. Hastie, Robert. Tibshirani, and J Jerome H Friedman. *The elements of statistical learning*, volume 1. Springer New York, 2001.

[33] Joshua WK Ho, Maurizio Stefani, Cristobal G dos Remedios, and Michael A Charleston. Differential variability analysis of gene expression and its application to human diseases. *Bioinformatics*, 24(13):i390–i398, 2008.

[34] A. Khan and S. Khan. Two level anomaly detection classifier. In *Computer and Electrical Engineering, 2008. ICCEE 2008. International Conference on*, pages 65–69, 2008.

[35] Shehroz S. Khan and Michael G. Madden. A survey of recent trends in one class classification. In Lorcan Coyle and Jill Freyne, editors, *AICS*, volume 6206 of *Lecture Notes in Computer Science*, pages 188–197. Springer, 2009.

[36] Carrie N Klabunde, Sally W Vernon, Marion R Nadel, Nancy Breen, Laura C Seeff, and Martin L Brown. Barriers to colorectal cancer screening: a comparison of reports from primary care physicians and average-risk adults. *Medical care*, 43(9):939–944, 2005.

[37] Serge Koscielny. Why most gene expression signatures of tumors have not been useful in the clinic. *Science Translational Medicine*, 2(14):14ps2–14ps2, 2010.

[38] Caryn Lerman, Barbara Rimer, Bruce Trock, Andrew Balshem, and Paul F Engstrom. Factors associated with repeat adherence to breast cancer screening. *Preventive medicine*, 19(3):279–290, 1990.

[39] Jingjing Li, Yu Liu, TaeHyung Kim, Renqiang Min, and Zhaolei Zhang. Gene expression variability within and between human populations and implications toward disease susceptibility. *PLoS computational biology*, 6(8):e1000910, 2010.

[40] Yingrui Li, Jingde Zhu, Geng Tian, Ning Li, Qibin Li, Mingzhi Ye, Hancheng Zheng, Jian Yu, Honglong Wu, Jihua Sun, et al. The dna methylome of human peripheral blood mononuclear cells. *PLoS biology*, 8(11):e1000533, 2010.

[41] Yu Liu, Mehmet Koyutürk, Jill S Barnholtz-Sloan, and Mark R Chance. Gene interaction enrichment and network analysis to identify dysregulated pathways and their interactions in complex diseases. *BMC systems biology*, 6(1):65, 2012.

[42] James W MacDonald and Debashis Ghosh. COPA–cancer outlier profile analysis. *Bioinformatics (Oxford, England)*, 22(23):2950–2951, December 2006.

[43] Jack S Mandel, John H Bond, Timothy R Church, Dale C Snover, G Mary Bradley, Leonard M Schuman, and Fred Ederer. Reducing mortality from colorectal cancer by screening for fecal occult blood. *New England Journal of Medicine*, 328(19):1365–1371, 1993.

[44] Larry M. Manevitz and Malik Yousef. One-class svms for document classification. *Journal of Machine Learning Research*, 2:139–154, 2001.

[45] Nathan Mantel. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer chemotherapy reports. Part 1*, 50(3):163–170, 1966.

[46] Jessica C Mar, Nicholas A Matigian, Alan Mackay-Sim, George D Mellick, Carolyn M Sue, Peter A Silburn, John J McGrath, John Quackenbush, and Christine A Wells. Variance of gene expression identifies altered network constraints in neurological disease. *PLoS genetics*, 7(8):e1002207, 2011.

[47] Laetitia Marisa, Aurélien de Reyniès, Alex Duval, Janick Selves, Marie Pierre Gaub, Laure Vescovo, Marie-Christine Etienne-Grimaldi, Renaud Schiappa, Dominique Guenot, Mira Ayadi, et al. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. *PLoS medicine*, 10(5):e1001453, 2013.

[48] Ettore Marubini and Maria Grazia Valsecchi. *Analysing survival data from clinical trials and observational studies*, volume 15. John Wiley & Sons, 2004.

[49] Matthew N. McCall, Harris A. Jaffee, and Rafael A. Irizarry. frma st: frozen robust multiarray analysis for affymetrix exon and gene st arrays. *Bioinformatics*, 28(23):3153–3154, 2012.

[50] Matthew N. McCall, Karan Uppal, Harris A. Jaffee, Michael J. Zilliox, and Rafael A. Irizarry. The gene expression barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. *Nucleic Acids Research*, 39(Database-Issue):1011–1015, 2011.

[51] Lance D Miller, Johanna Smeds, Joshy George, Vinsensius B Vega, Liza Vergara, Alexander Ploner, Yudi Pawitan, Per Hall, Sigrid Klaar, Edison T Liu, et al. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proceedings of the National Academy of Sciences of the United States of America*, 102(38):13550–13555, 2005.

[52] Polly A Newcomb, Robert G Norfleet, Barry E Storer, Tanya S Surawicz, and Pam M Marcus. Screening sigmoidoscopy and colorectal cancer mortality. *Journal of the National Cancer Institute*, 84(20):1572–1575, 1992.

[53] Jerzy Neyman and Elizabeth Scott. On the use of c (alpha) optimal tests of composite hypotheses. *Bulletin of the International Statistical Institute*, 41(1):477–497, 1965.

[54] Michael F Ochs, Jason E Farrar, Michael Considine, Yingying Wei, Soheil Meshinchi, and Robert J Arceci. Outlier analysis and top scoring pair for integrated data analysis and biomarker discovery. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 11(3):520–532, 2014.

[55] Hirokazu Okayama, Takashi Kohno, Yuko Ishii, Yoko Shimada, Kouya Shiraishi, Reika Iwakawa, Koh Furuta, Koji Tsuta, Tatsuhiro Shibata, Seiichiro Yamamoto, et al. Identification of genes upregulated in alk-positive and egfr/kras/alk-negative lung adenocarcinomas. *Cancer research*, 72(1):100–111, 2012.

[56] Alicia Oshlack, Matthew J Wakefield, et al. Transcript length bias in rna-seq data confounds systems biology. *Biol Direct*, 4(1):14, 2009.

[57] Ivyna Bong Pau Ni, Zubaidah Zakaria, Rohaizak Muhammad, Norlia Abdullah, Naqiyah Ibrahim, Nor Aina Emran, Nor Hisham Abdullah, and Sharifah Noor Akmal Syed Hussain. Gene expression patterns distinguish breast carcinomas from normal breast tissues: the malaysian context. *Pathology-Research and Practice*, 206(4):223–228, 2010.

[58] Yudi Pawitan, Judith Bjöhle, Lukas Amler, Anna-Lena Borg, Suzanne Egyhazi, Per Hall, Xia Han, Lars Holmberg, Fei Huang, Sigrid Klaar, et al. Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Research*, 7(6):R953, 2005.

[59] Daisy Phillips, Debashis Ghosh, et al. Testing the disjunction hypothesis using voronoi diagrams with applications to genetics. *The Annals of Applied Statistics*, 8(2):801–823, 2014.

[60] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[61] Jacob Sabates-Bellver, Laurens G Van der Flier, Mariagrazia de Palo, Elisa Cattaneo, Caroline Maake, Hubert Rehrauer, Endre Laczko, Michal A Kurowski, Janusz M Bujnicki, Mirco Menigatti, et al. Transcriptome profile of human colorectal adenomas. *Molecular Cancer Research*, 5(12):1263–1275, 2007.

[62] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, 2002.

[63] Alex K Shalek, Rahul Satija, Xian Adiconis, Rona S Gertner, Jellert T Gaublomme, Raktima Raychowdhury, Schraga Schwartz, Nir Yosef, Christine Malboeuf, Diana Lu, et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, 498(7453):236–240, 2013.

[64] Tobias Sing, Oliver Sander, Niko Beerenwinkel, and Thomas Lengauer. Rocr: visualizing classifier performance in r. *Bioinformatics*, 21(20):3940–3941, 2005.

[65] Magdalena Skrzypczak, Krzysztof Goryca, Tymon Rubel, Agnieszka Paziewska, Michal Mikula, Dorota Jarosz, Jacek Pachlewski, Janusz Oledzki, and Jerzy Ostrowsk. Modeling oncogenic signaling in colon tumors by multidirectional analyses of microarray data directed for maximization of analytical reliability. *PLoS One*, 5(10):e13091, 2010.

[66] Christos Sotiriou, Pratyaksha Wirapati, Sherene Loi, Adrian Harris, Steve Fox, Johanna Smeds, Hans Nordgren, Pierre Farmer, Viviane Praz, Benjamin Haibe-Kains, et al. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *Journal of the National Cancer Institute*, 98(4):262–272, 2006.

[67] Spotswood L Spruance, Julia E Reid, Michael Grace, and Matthew Samore. Hazard ratio in clinical trials. *Antimicrobial agents and chemotherapy*, 48(8):2787–2792, 2004.

[68] John D Storey, Jennifer Madeoy, Jeanna L Strout, Mark Wurfel, James Ronald, and Joshua M Akey. Gene-expression variation within and among human populations. *The American Journal of Human Genetics*, 80(3):502–509, 2007.

[69] Jyothi Subramanian and Richard Simon. Gene expression–based prognostic signatures in lung cancer: Ready for clinical use? *Journal of the National Cancer Institute*, 102(7):464–474, 2010.

[70] Duxin Sun, Hans Lennernas, Lynda S Welage, Jeffery L Barnett, Christopher P Landowski, David Foster, David Fleisher, Kyung-Dall Lee, and Gordon L Amidon. Comparison of human duodenum and caco-2 gene expression profiles for 12,000 gene sequences tags and correlation with permeability of 26 drugs. *Pharmaceutical research*, 19(10):1400–1416, 2002.

[71] Miho M Suzuki and Adrian Bird. Dna methylation landscapes: provocative insights from epigenomics. *Nature Reviews Genetics*, 9(6):465–476, 2008.

[72] Lazio Tabar, A Gad, LH Holmberg, U Ljungquist, CJG Fagerberg, L Baldetorp, O Gröntoft, B Lundström, JC Månson, G Eklund, et al. Reduction in mortality from breast cancer after mass screening with mammography: randomised trial from the breast cancer screening working group of the swedish national board of health and welfare. *The Lancet*, 325(8433):829–832, 1985.

[73] Andrew E Teschendorff, Allison Jones, Heidi Fiegl, Alexandra Sargent, Joanna J Zhuang, Henry C Kitchener, and Martin Widschwendter. Epigenetic variability in cells of normal cytology is associated with the risk of future morphological transformation. *Genome medicine*, 4(3):24, 2012.

[74] Terry M Therneau. *A Package for Survival Analysis in S*, 2014. R package version 2.37-7.

[75] Robert Tibshirani and Trevor Hastie. Outlier sums for differential gene expression analysis. *Biostatistics*, 8(1):2–8, 2007.

[76] Robert Tibshirani and Trevor Hastie. Outlier sums for differential gene expression analysis. *Biostatistics (Oxford, England)*, 8(1):2–8, January 2007.

[77] Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10):6567–6572, 2002.

[78] Winston Timp, Hector Corrada Bravo, Oliver G McDonald, Michael Goggins, Chris Umbricht, Martha Zeiger, Andrew P Feinberg, and Rafael A Irizarry. Large hypomethylated blocks as a universal defining epigenetic alteration in human solid tumors. *Genome medicine*, 6(8):61, 2014.

[79] Scott A Tomlins, Daniel R Rhodes, Sven Perner, Saravana M Dhanasekaran, Rohit Mehra, Xiao-Wei Sun, Sooryanarayana Varambally, Xuhong Cao, Joelle Tchinda, Rainer Kuefer, et al. Recurrent fusion of tmprss2 and ets transcription factor genes in prostate cancer. *Science*, 310(5748):644–648, 2005.

[80] Laura J van't Veer, Hongyue Dai, Marc J Van De Vijver, Yudong D He, Augustinus AM Hart, Mao Mao, Hans L Peterse, Karin van der Kooy, Matthew J Marton, Anke T Witteveen, et al. Gene expression profiling predicts clinical outcome of breast cancer. *nature*, 415(6871):530–536, 2002.

[81] Vladimir Vapnik. *The nature of statistical learning theory*. springer, 2000.

[82] Vladimir Vapnik and Olivier Chapelle. Bounds on error expectation for support vector machines. *Neural computation*, 12(9):2013–2036, 2000.

[83] Grace Wahba et al. Support vector machines, reproducing kernel hilbert spaces and the randomized gacv. *Advances in Kernel Methods-Support Vector Learning*, 6:69–87, 1999.

[84] Judith ME Walsh and Jonathan P Terdiman. Colorectal cancer screening: scientific review. *Jama*, 289(10):1288–1296, 2003.

[85] Chenwei Wang, Alperen Taciroglu, Stefan R Maetschke, Colleen C Nelson, Mark A Ragan, and Melissa J Davis. mCOPA: analysis of heterogeneous features in cancer expression data. *Journal of clinical bioinformatics*, 2(1):22, 2012.

[86] Y Wang, J G Klijn, Y Zhang, A M Sieuwerts, M P Look, F Yang, D Talantov, M Timmermans, M E Meijer-van Gelder, J Yu, T Jatkoe, E M Berns,

D Atkins, and J A Foekens. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, 365(9460):671–679, February 2005.

[87] John Watkinson, Xiaodong Wang, Tian Zheng, and Dimitris Anastassiou. Identification of gene interactions associated with disease from gene expression data using synergy networks. *BMC systems biology*, 2(1):10, 2008.

[88] Hwangjo Yu, Jiawei Han, and Kevin Chen-Chuan Chang. PEBL: Web page classification without negative examples. *IEEE Transactions on Knowledge adn data engineering*, 16(1), January 2004.

[89] Jigang Zhang, Jian Li, and Hong-Wen Deng. Identifying gene interaction enrichment for gene expression data. *PloS one*, 4(11):e8064, 2009.