12-2014

Classification of clinical outcomes using high-throughput and clinical informatics.

Alexander Carswell Cambon
University of Louisville

CLASSIFICATION OF CLINICAL OUTCOMES USING HIGH-THROUGHPUT AND CLINICAL INFORMATICS

By

Alexander Carswell Cambon
B.S., Syracuse University, 1982
M. Eng., Pennsylvania State University, 1997

A Dissertation
Submitted to the Faculty of the
School of Public Health and Information Sciences
Of the University of Louisville
In Partial Fulfillment of the Requirements
For the Degree of

Doctor of Philosophy

Department of Bioinformatics and Biostatistics
University of Louisville
Louisville, Kentucky

December 2014

CLASSIFICATION OF CLINICAL OUTCOMES USING HIGH-THROUGHPUT

AND CLINICAL INFORMATICS

By

Alexander Carswell Cambon
B.S., Syracuse University, 1982
M. Eng., Pennsylvania State University, 1997

A Dissertation Approved on

November 19, 2014

by the following Dissertation Committee:

_____

Shesh N. Rai

_____

Kathy B. Baumgartner

_____

Guy N. Brock

_____

Nigel G.F. Cooper

_____

Dongfeng Wu

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Shesh N. Rai, for his guidance, patience, and valuable input. His open-ended style encouraged independent thought and development of the dissertation material. This inspired me to work harder, and as a result my confidence in ability to do independent statistical research has vastly increased. I would also like to thank the other committee members, Dr. Kathy Baumgartner, Dr. Guy Brock, Dr. Nigel Cooper, and Dr. Wu for comments and assistance over the past four years. Each of them provided significant contributions in various aspects of the dissertation work. I would like to thank my wife Anne for her patience and understanding during this time period. I was both working full time and working on the PhD for many years. I would also like to thank The University of Louisville Writing Center, and specifically Rebecca Hallman for invaluable input and comments to improve readability and integration of the manuscript. We would also like to thank the University of Louisville Health Sciences Kornhauser Library, and specifically Vida M. Vaughn, for performing and providing input concerning literature searches and methods. Research reported in this publication was supported by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institute of Health under grant number 5P20GM103436-13. The article contents are solely the responsibility of the authors and do not represent the official views of the National Institute of Health. Dr. Shesh Rai was generously supported by Dr. DM Miller, Director, the James Graham Brown Cancer Center and the Wendell Cherry Chair in Clinical Trial Research.

ABSTRACT

CLASSIFICATION OF CLINICAL OUTCOMES USING HIGH-THROUGHPUT AND CLINICAL INFORMATICS

Alexander C. Cambon

November 19, 2014

It is widely recognized that many cancer therapies are effective only for a subset of patients. However clinical studies are most often powered to detect an overall treatment effect. To address this issue, classification methods are increasingly being used to predict a subset of patients which respond differently to treatment. This study begins with a brief history of classification methods with an emphasis on applications involving melanoma. Nonparametric methods suitable for predicting subsets of patients responding differently to treatment are then reviewed. Each method has different ways of incorporating continuous, categorical, clinical and high-throughput covariates. For nonparametric and parametric methods, distance measures specific to the method are used to make classification decisions. Approaches are outlined which employ these distances to measure treatment interactions and predict patients more sensitive to treatment. Simulations are also carried out to examine empirical power of some of these classification methods in an adaptive signature design. Results were compared with logistic regression models. It was found that parametric and nonparametric methods performed reasonably well. Relative performance of the methods depends on the simulation scenario. Finally a method was developed to evaluate power and sample size needed for an adaptive signature design in order to predict the subset of patients sensitive to treatment. It is hoped that this study will stimulate more development of nonparametric and parametric methods to predict subsets of patients responding differently to treatment.

Key Words: classification; machine learning; dimension reduction; interaction; melanoma; clinical study.

TABLE OF CONTENTS

PAGE

ACKNOWLEDGMENTS iii

ABSTRACT iv

LIST OF TABLES ix

LIST OF FIGURES xi

NONPARAMETRIC CLASSIFICATION METHODS 1

1. Introduction 1

2. Notation and Definitions 8

3. Nested Cross Validation (NCV) to Estimate and Reduce Prediction Error 9

4. Nonparametric Dimension Reduction (DR) in Classification 11

5. Nonparametric Classification Methods 20

6. Generalizability of Classification Methods 39

7. Discussion and Conclusions 39

8. Challenges and Future Directions 43

PARAMETRIC CLASSIFICATION METHODS 44

1. Some Historical Developments of Parametric Classification Methods 44

2. Notation and Definitions 47

3. A Brief Introduction to Parametric Dimension Reduction with a Focus on Treatment Subset Prediction 48

4. Parametric Classification Models and Rules 54

5. Discussion and Conclusions 66

6. Challenges and Future Directions 71

ESTIMATING DESIGN PARAMETERS IN THE PRESENCE OF GENE AND GENE - TREATMENT INTERACTION 75

1. Introduction 75

2. Notation and Definitions 76

3. Background for the Adaptive Signature Design 77

4. An Adaptive Signature Design Model Similar to Freidlin and Simon 78

5. Power for the Adaptive Signature Design 82

6. Limitations and Future Work 95

PROPERTIES OF ADAPTIVE CLINICAL TRIAL SIGNATURE DESIGN IN THE PRESENCE OF GENE AND GENE –TREATMENT INTERACTION 99

1. Introduction 100

2. Background of ASD Model 101

3. Extensions and Modifications of ASD Method 103

4. Simulation Study 104

5. Results 107

6. Discussion 108

REFERENCES 118

CURRICULUM VITA 128

LIST OF TABLES

TABLE PAGE

Table 1: List of acronyms, definitions, and sections where acronyms are used. 9

Table 2: Nonparametric classification and related methods. 12

Table 3: Relationship between number of features and number of possible interactions. 16

Table 4: Some Nonparametric Dimension Reduction Methods. 19

Table 5: List of acronyms, definitions, and sections where acronyms are used. 48

Table 6: Parametric DR methods and software implementations. 53

Table 7: Some parametric classification methods and software implementations. 61

Table 8: Select parametric dimension reduction and classification methods – situation and conditions. 73

Table 9: PR_{11}=0.98, PR_{10}=PR_{01}=PR_{00}=0.25. (Expression-Treatment Interaction Effect Only); $\rho = 0$. 111

Table 10: PR_{11}=0.98, PR_{10}=PR_{01}=PR_{00}=0.25. (Expression-Treatment Interaction Only); $\rho = 0.6$. 112

Table 11: Small Sample Size Simulation Scenario PR_{11}=0.98, PR_{10}=0.25, PR_{01}=PR_{00}=0.25; n_1=n_2=100, $\rho = 0$. 113

Table 12: Gene Main Effect, and Gene-Treatment Interaction:PR_{11}=0.98, PR_{10}=0.35, PR_{01}=PR_{00}=0.25, $\rho = 0$. 114

Table 13: Sensitivity Main Effect, and Treatment Main Effect, and Sensitivity-Treatment Interaction: PR_{11}=0.98, PR_{10}=0.35, PR_{01}=0.35, PR_{00}=0.25, $\rho = 0$. 115

Table 14: Comparison of TWV Methods under simulation scenarios similar to Freidlin and Simon work (unequal variances for gene expression for predictive genes between sensitivity classes). PR_{11}=0.98, PR_{10}=0.25, PR_{01}=0.25, PR_{00}=0.25, $\rho = 0$. 116

Table 15: Comparison of LR_{TWV} and LDA_{TWV} methods when variances of gene expression for predictive genes are constrained to be equal PR_{10}=0.25, PR_{01}=0.25, PR_{00}=0.25, $\rho = 0$. 117

LIST OF FIGURES

FIGURE PAGE

Figure 1: Flowchart showing subdivisions of statistical learning methods. 2

Figure 2: Flowchart of Classification Process. 3

Figure 3a and b: Use of MDR to determine high-risk loci-genotype combinations. 21

Figure 4: A support vector machine for a 2-dimensional feature space. 29

Figure 5: Flowchart of ASD nested cross validation method. 32

Figure 6: Use of compound measure in FDA to allocate new subjects. 55

NONPARAMETRIC CLASSIFICATION METHODS

1. Introduction

Overview of Nonparametric Classification Methods

Classification is a subset of what Hastie, Tibshirani, & Friedman (2009 [75]) term the statistical learning field, and what Bishop (2006 [10]) refers to as pattern recognition or machine learning. Statistical learning can further be divided into two categories, supervised learning and unsupervised learning, as shown in Figure 1. “It is called ‘supervised’ because of the presence of the outcome variable to guide the learning process. In the unsupervised learning problem, we observe only the features and have no measurements of the outcome.” (Hastie et al., 2009 [75]).

In language often used in machine learning or pattern recognition, supervised learning methods have both inputs and outputs. Pattern recognition uses the term features for inputs. In the statistical field, the outputs are referred to as the outcome, response, or dependent variables, and the inputs are covariates, explanatory variables, or independent variables. These covariates can be high dimensional, such as genomic data, or low dimensional, such as age, gender, ethnicity, tumor thickness, or presence of ulceration.

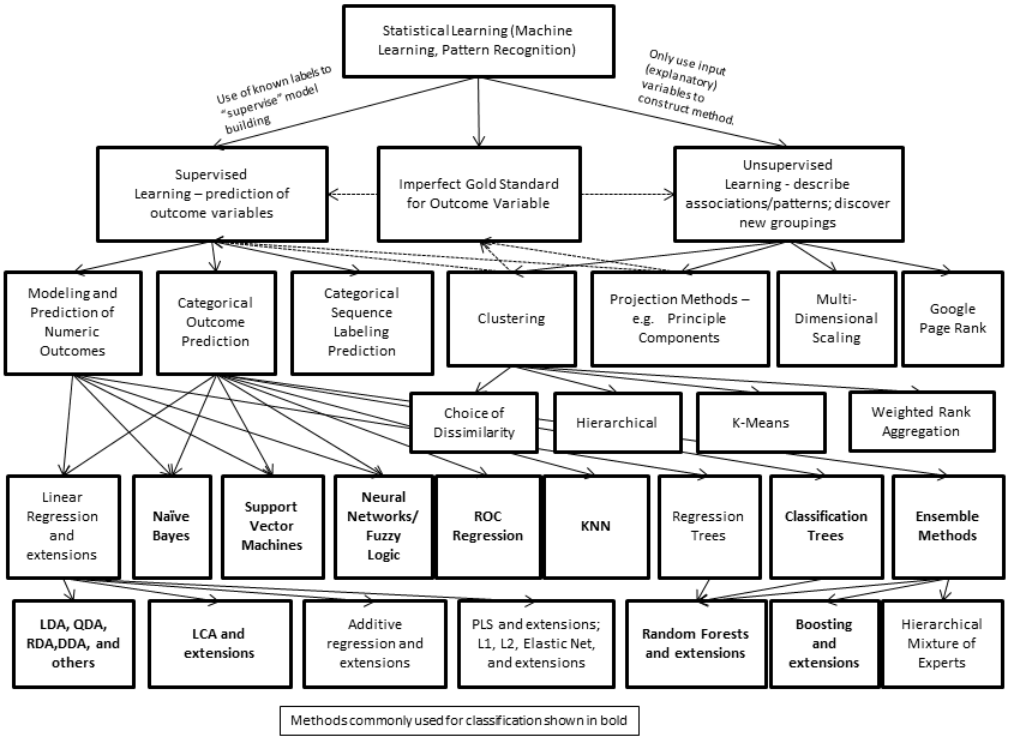

Figure 1: Flowchart showing subdivisions of statistical learning methods.

As seen in Figure 1, supervised learning can then be further subdivided into classification (modeling and predicting of categorical outcomes), and regression (modeling and prediction of continuous outcome variables). In classification, the categorical outcomes are the class labels which the classification method is predicting. However regression methods are also often included as a step in the classification process.

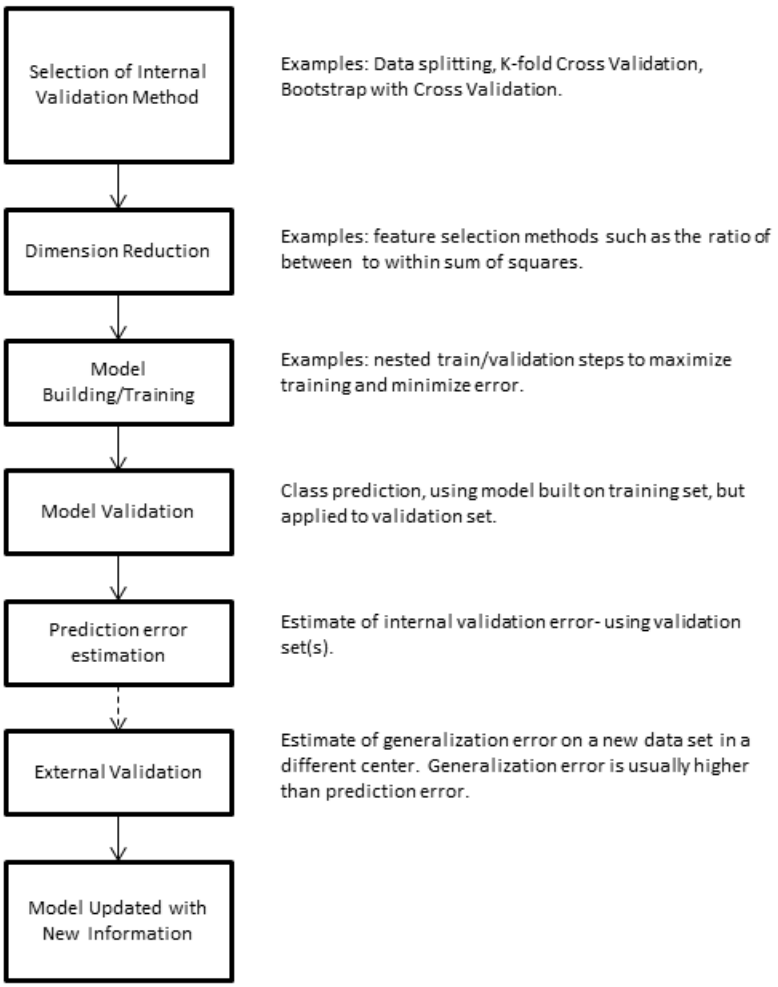

Figure 2: Flowchart of Classification Process. Steps shown separately such as dimension reduction and model building may be part of the same step.

Figure 2 shows a high level flowchart of the classification process. Model building is done on the training set, and prediction error assessed on the validation set. To prevent over fitting during model building, the training set itself is often divided up into nested training and validation steps. Within each paired training and validation set, the training and validation portions are usually non-overlapping.

The roots of discriminant analysis, a class of methods used to separate labeled groups of objects using covariates, date back to well before 1936, when Mahalanobis published his work on the "generalized distance" (1936 [111]), and Fisher published a closely related method, using the

same distance, to discriminate between species of iris (1936 [48]). Ensuing developments in nonparametric classification include K Nearest Neighbors (KNN) and Kernel Density Analysis (KDA) used for classification, both proposed by Fix & Hodges (1951 [49]).

These earlier methods were developed before the proliferation of high-throughput data (HTD), such as genomic or proteomic data. HTD typically have thousands or more of features, together with sample sizes on the order of 100 or less. This situation is often referred to as $p >> n$ where p refers to the features and n refers to the sample size of subjects. The early methods cannot handle $p >> n$ data without dimension reduction (DR) methods. Many DR methods have been originated for HTD to address this situation. Additional classification methods such as Random Forests (Breiman, 1999 [17]), Support Vector Machines (Boser, Buyon, & Vapnik, 1992 [12]) and Boosting (Freund & Schapire, 1995 [56]) have built-in (embedded) DR techniques. At the same time many of the earlier developed classification methods, when used in conjunction with DR techniques, compare favorably to these more recently developed methods. For example KNN outperformed Boosting and Random Forests in a comparison of methods by Dudoit, Fridlyand, & Speed (2002 [39]) involving genomic data.

Some Early Uses of Genomic Data, Clustering, and Classification

In the late 1990's and early 2000's unsupervised methods were used to show that genomic data could be used not only to identify known classes but also to identify or predict new classes. Golub et al. (1999 [66]) developed a method involving clustering and weighted voting of "informative" genes for cancer classification. Their results pointed in the direction of "a general strategy for discovering and predicting cancer classes for other types of cancer, independent of previous biological knowledge". This method was able to discriminate between two types of leukemia. Treatment regimens could then be targeted to the type of leukemia if it were known.

Development of Medical and Statistical Classification Methods in Melanoma

Following closely after the work of Golub et al. (1999 [66]), Bittner et al. (2000 [11]) applied clustering and multiple dimensional scaling (MDS) to a melanoma study and found that gene expression patterns correlated by specific subsets of melanomas. This finding pointed toward the possibility of classification of melanoma based on gene expression. It was only more recently that the prognostic system for melanoma as defined by the American Joint Committee on Cancer or AJCC (Balch et al., 2009 [5]) officially added mitotic rate to the other prognostic variables. Although no molecular marker was included with this latest update (Duncan, 2009 [40]), the addition of mitotic rate to the AJCC Index in 2009 was seen as evidence of "heightened interest in the utility of molecular markers" for melanoma (Segura et al., 2010 [147]). However, although factors such as number of metastatic nodes and tumor thickness explain "tremendous heterogeneity of prognosis among patients with stage III melanoma" (Balch et al., 2010 [6]), Duncan (2009 [40]) and Balch et al. (2004 [4]) pointed out that the AJCC melanoma cancer classification system "is a tumor-node-metastasis (TNM) based clinical and histologic scheme that segregates patients into prognostic categories that are not correlated with any specific therapeutic response". Duncan further stated that "integration of the genomic and clinical pathologic schemes may provide a future classification scheme that segregates tumors by predicted outcome and potential response to specific therapy."

The earlier classification systems for melanoma, as in other fields, underwent an evolution. Table 1 of the Rigel, Russak, & Friedman study (2010 [133]) showed the primary diagnosis of melanoma was, before the 1980's confined to gross features such as bleeding ulceration, but progressed to use of subsurface features in the 1990's and eventually to digital and subcellular feature after the turn of the century. The ABCD (asymmetry, border, color, differential structure) system was developed in the United States by Friedman, Rigel, & Kopf (1985 [61]) for self-examination and early detection. The 7-Point Checklist was also devised in the same year in the United Kingdom by MacKie (1985 [109]). It was based on seven features: sensory change,

diameter ≥ 1 cm, lesion growth, irregular edge, irregular pigmentation, inflammation, and crusting/oozing/bleeding. Argenziano et al. (1998 [3]) compared a 7-Point Checklist system based on simplified epiluminescence microscopy (ELM) pattern analysis to ABCD and found it improved sensitivity and specificity and required less experience to use. Henning et al. (2007 [77]) introduced the CASH system (color, architecture, symmetry, and homogeneity).

The 1990's and especially the 2000's decade saw increasing use of learning methods to distinguish melanoma tumors based on shape and other characteristics. For example, Claridge, Hall, Keefe, & Allen (1992 [28]) used shape analysis to classify melanoma, and Ercal, Chawla, Stoecker, Lee, & Moss (1994 [45]) used a machine learning method to distinguish melanoma from benign tumors using shape and relative tumor color. The more recent work by Lee & Claridge (2005 [100]) introduced the predictive power of the Irregularity Index in the diagnosis of malignant melanoma. Table 1 of the Rigel et al. study (2010 [133]) also identified systems including image analysis and pattern recognition in the 2000's. Table 2 of the same study compared sensitivity, specificity, and diagnostic accuracy of various dermoscopy diagnostic algorithms. MelaFind, a "noninvasive and objective computer-vision system designed to aid in detection of early pigmented cutaneous melanoma" (Monheit et al., 2011 [114]) was approved for use by the U.S. Food and Drug Administration (FDA) after a successful prospective phase III clinical trial. The last 10 to 15 years has also seen a proliferation of research incorporating genomic as well as genetic data for use in classification of melanoma. The findings by Viros et al. (2008 [166]) incorporated mutation status of oncogenes BRAF and NRAS along with histological features, and showed the BRAF mutations correlated well with morphological features and found that there were "significant survival benefit… for patients who, based on their age, were predicted to have BRAF mutant melanomas in 69% of the cases". In the previously cited Segura et al. work (2010 [147]), a microRNA (miRNA) signature was developed "whose overexpression was significantly correlated with longer survival". Several learning methods were used in this study including Nearest Shrunken Centroids or NSC (Tibshirani, Hastie, Narasimhan,

& Chu, 2003 [161]), Support Vector Machines (SVM), Adaboost with classification trees (Freund & Schachter, 1996 [57]), and Random Forests. This study also incorporated Pre-validation (PV), developed by Tibshirani & Efron (2002 [159]), to compare the added value of genomic covariates in classifiers using traditional prognostic indicators.

Use of Classification to Predict a Subset of Patients by Identifying Sources of Heterogeneity

Treatments specific to melanoma and applicable to a subset of patients have more recently been coming onto market. For example, subsequent to the findings in Viros et al. (2008 [166]), a treatment Vemurafenib, for a subset of patients having melanoma with BRAF V600e mutations (Chapman et al., 2011 [24]) was approved by the FDA. In the year prior, Ipilimumab was approved by the FDA for treatment of metastatic melanoma (Hodi et al., 2010 [79]). Saenger & Wochok (2009 [139]) had previously shown that heterogeneity is present in patient response to Ipilimumab. Moreover Freidlin, Jiang, & Simon (2010 [55]) stated that "due to the molecular heterogeneity of most human cancers, only a subset of treated patients benefit from a given therapy. This is particularly relevant for the new generation of anticancer agents that target specific molecular pathways." At the same time, relapse of patients or a subset of patients under these new treatments remains a challenge. Treatment combinations are one avenue being explored (Tuma, 2012 [162], and Vanneman & Dranoff, 2012 [163]). As pointed out by Freidlin et al. (2010 [55]), "Genomic … technologies … provide powerful tools for identifying a genetic signature for patients who are most likely to benefit from a targeted agent".

Purpose of Chapter

The purpose of this study is to carefully review and examine nonparametric classification methods in order to understand and best use these methods for treatment subset prediction. Each method uses a measure of distance to make classification decisions. Proper understanding of these classification-specific distances facilitates their use in prediction of sensitive patients.

Organization of Chapter

The remainder of this chapter is organized as follows: Section 2 addresses notation and definitions. Section 3 covers nested cross validation methods to estimate and reduce prediction error. Section 4 outlines nonparametric DR methods for classification. Nonparametric classification methods are in Section 5. Section 6 highlights generalizability of classification methods. Discussion and conclusions are in the penultimate section, and Section 8 highlights challenges and future directions.

2. Notation and Definitions

Subscripts *D* and *V* denote training and validation sets respectively, and n_D and n_V denote their sample sizes; y_i denotes a categorical or a continuous response for the i^{th} subject, $i = 1,...,n;\ n = n_D + n_V$. However when class is the outcome variable, then $g_i = 1$ denotes the disease or relapsed class, and $g_i = 0$ or -1 denotes the other class; covariate vectors $\boldsymbol{x}_i = \{x_{i1}, x_{i2},..., x_{im}\}$ and $\boldsymbol{z}_i = \{z_{i1}, z_{i2},..., z_{i\ell}\},$ where $\ell = p - m$, denote respectively the clinical and high-dimensional covariate vectors for the i^{th} subject. The covariate vector for the i^{th} subject is then $\{\boldsymbol{x}_i, \boldsymbol{z}_i\}$, and the vector for the i^{th} subject consists of the class-covariate vector pairing $\{g_i, \{\boldsymbol{x}_i, \boldsymbol{z}_i\}\},\ i = 1,..., n_D$. In the training set $\hat{\mu}_k$ denotes the average of the k^{th} feature, $\hat{\mu}_{gk}$ denotes the average of feature k for class *g*, and $\hat{\mu}_{gtk}$ denotes the average of feature k for class *g* and treatment arm *t*, where $t = 1$ denotes an enhanced treatment, and $t = 0$ denotes standard treatment or control.

The term feature refers either to covariates or functions of covariates. For example in the model $E(Y_i) = \beta_0 + \beta_1 \log(x_{i1}),$ the term $\log(x_{i1})$ is used in place of the identity term and is therefore both a covariate and a feature. However in the equation: $E(Y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 \log(x_{i2}),$ the term $\log(x_{i1})$ is used in addition to the identity term and is therefore only a feature. Features may also be functions of several covariates. For example, in the

equation $E(Y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ...\beta_m x_{im} + \gamma_1 z_{i1} + \ldots \gamma_\ell z_{i\ell}$, a Feature Extraction (FE) method could extract a smaller number of features from the covariates.

The term interaction refers to a statistical interaction. A statistical interaction between two features indicates that the effect of one feature on the response depends on the level of the other feature. Higher-order interactions are interactions between more than two features. Epistasis refers to situations in which gene interactions are present while the corresponding main effects are small. Cordell (2009 [19]) provides an in-depth review of gene-gene interactions. Table 1 provides a list of acronyms, together with definitions and sections where they are used.

Table 1: List of acronyms, definitions, and sections where acronyms are used.

Acronym	Associated Words	Description	Sections
ASD	Adaptive Signature Design	Class of Methods to predict patients sensitive to treatment	1,3-6
CV	Cross Validation		1,3-6
DLLR	Difference in LLR		5
DR	Dimension reduction		1,3-5,7-8
FE	Feature Extraction	Class of DR Methods	2,3
FS	Feature Selection	Class of DR Methods	3
HTD	High throughput data		1,4
KDA	Kernel Density Analysis	Can be used as a nonparametric classification method	1,5.3,7
KNN	K Nearest Neighbors	Nonparametric Classification Method	1,5.2,7
LLR	Log Likelihood Ratio		4,5
LR	Likelihood Ratio		4
LRT	Likelihood Ratio Test		5
MDR	Multifactor Dimensionality Reduction	Nonparametric Classification Method	5.1
mRNA	Messenger RNA		6
miRNA	microRNA		6
NSC	Nearest Shrunken Centroids	HTD classification method	1,5.2
OR	Odds Ratio		4-5
PLS	Partial Least Squares		3
PV	Pre-Validation	Method of assessing added value of genomic features to clinical features	1,5
SNP	Single Nucleotide Polymorphism		6
SVM	Support Vector Machine	Nonparametric Classification Method	1,5.4, 7

3. Nested Cross Validation (NCV) to Estimate and Reduce Prediction Error

Data splitting involves dividing the data into a training set and a non-overlapping validation set in order to avoid a downward bias of prediction error (PE), which is the probability of an

incorrect classification. The expected prediction error can be defined in terms of the joint distribution of G, X and Z as follows (Hastie et al., 2009 [75]):

$$E(PE) = E[L(G, \hat{G}(X, Z))],$$

where L is a loss function associated with incorrect classification. If the loss function is 0 for correct classification and 1 for misclassification, then prediction error can be estimated on the validation set by:

$$\hat{p}_E = \sum_{i=n_D+1}^{n_D+n_V} I(\hat{g}_i \neq g_i) / n_V .$$

where g_i is class membership for subject *i*, and $\hat{g}_i$ is predicted class membership.

The classification model is built using data on the training set only, and then this rule is applied to the validation set to predict class outcome, without use of class information on the validation set even if it is available. Only after the predicted outcomes have been made is class information used to estimate prediction error. This helps to achieve an unbiased estimate of prediction error.

However the goal of machine or statistical learning is not just to avoid a downward bias in prediction error, but also to minimize prediction error. The Adaptive Signature Design or ASD (Freidlin & Simon, 2005 [54]) incorporates this learning process by optimizing a set of tuning parameters on the training set. To avoid overfitting, cross validation is performed on the training set for each set of tuning parameters. A grid search is then used to select the optimal set of tuning parameters. This final tuning parameter set is then used for prediction on the final validation set. This approach is called nested cross validation, or NCV.

Methods such as the bootstrap (sampling with replacement, Efron, 1994 [42]) can also be used. If a bootstrap sample of size n_D is used on the training set, an average of approximately 63% of the training set subjects are placed into a nested training set. Some of the training set subjects are selected more than once, and some are not selected at all. The subjects not selected in

the bootstrap sample (the out-of bag or OOB sample) are then placed in the corresponding nested validation set. This process is then repeated B times, with some method to assure than all subjects in the training set are included in a nested validation set at least once. The number B is chosen depending on the classification task at hand. See Pattengale, Alipour, Bininda-Emonds Moret, & Stamatakis (2010 [122]) for guidance. Since class outcomes for many of the training set subjects are predicted more than once, a voting or averaging rule is needed to make the final classification decision. Breiman (2001 [18]) used OOB prediction error together with the bootstrap in his Random Forests method. Carpenter & Bithell (2000 [21]) provide a tutorial on both parametric and nonparametric bootstrap methods. Table 2 shows classification-related methods, including the bootstrap, along with strengths and weaknesses, and related software packages.

4. Nonparametric Dimension Reduction (DR) in Classification

There is extensive ongoing research involving DR, and as a title of recent review by Fan & Lv (2010 [47]) implied, the literature involving DR is so extensive that any review of this topic itself requires a DR technique. The DR step in classification should be in alignment with the goals of the specific classification problem. Specifically, the goal of DR is to reduce as much as possible presence of noise while at the same time preserving information relevant to class outcome. The presence of many noisy features can have a large negative impact on classification accuracy, and can even reduce accuracy to that of random guessing (Fan & Fan, 2008 [46]).

DR methods can be subdivided into feature selection (FS) methods and FE methods. FS methods do not change individual features, but merely select a subset of them. FE methods reduce dimensionality by extracting a smaller number of new features from the existing features. FE may be needed if pairwise correlations between features are high, or if prior knowledge favors grouping of features (Fan & Lv, 2010 [47]).

Table 2: Nonparametric classification and related methods.

Method	Strengths	Drawbacks	R packages	Other Packages
MDR	Incorporates interactions agnostically, including higher order gene-gene and gene environment interactions; makes few assumptions; often attains good classification performance.	Requires special modifications to handle continuous covariates; since it uses best subset method, it requires prior DR step in HTD setting.	MDR	MDR (open source on sourceforge.net)
KNN	Often attains good classification performance; one of most simple and intuitive methods, one of most widely used for low-dimensional data	Does not naturally handle categorical covariates; requires prior DR step in HTD; usually requires scaling of covariates.	knn(class);kknn,	Proc discrim (SAS) Enterprise Miner (SAS)
Naïve Bayes	Often improves classification performance by reducing variance at the expense of some increase in bias	May not work well when independence assumption is severely violated.	e1071	SAS macros available
KDA	Adaptable to different underlying distributions, while still retaining desirable properties of statistical densities- naturally handles unequal class sizes and different misclassification costs	Poor small sample size performance; does not naturally handle categorical covariates, requires prior DR step; comparatively high computational burden inhibits use in DR steps.	density(base*), kknn,	Proc kde (SAS)
SVM	Robust (places more emphasis on observations near class boundary); includes embedded DR methods; can handle nonlinear boundaries	Requires extensive tuning; SVM's using kernel functions with inner products may need a prior DR step since they increase the dimensional space; SVM does not handle unequal class sizes as well as KDA; usually requires scaling of covariates.	e1071, svmpath	SHOGUN (http://shogun-toolbox.org/) Enterprise Miner (SAS)
Classification Trees	Easy to interpret, handles mixed covariates, naturally incorporates interactions	May loose information from continuous covariates; single trees often have poor classification performance.	rpart	Enterprise Miner (SAS), CART® Salford Systems), GUIDE, LOTUS
Random Forests	Smooths out cutpoints for continuous covariates, improved classification performance compared to classification trees, naturally incorporates interactions, including treatment interactions;	Loses some interpretability compared to single classification trees.	randomForest	Random Forest (Salford Systems) Enterprise Miner (SAS)
Bumping (or Bump Hunting)	Uses bootstrapping to build an ensemble of trees, but selects tree with lowest prediction error. Useful for finding interactions in absence of main effects.	Needs prior DR step in HTD setting. May lose information from continuous covariates.	prim	
Boosting	Can control order of interactions, often attains good classification performance, has natural embedded DR, acceptable speed/good performance in HD with off – the-shelf software	Loses some interpretability compared to single classification trees.	gbm, mboost	Enterprise Miner (SAS)
Bootstrap	Not a classification method in itself, but incorporated into many classification methods such as Random Forests to improve classification performance	Requires some increase in computational costs, though this continues to be less of a concern with increased computing power.	rms, bootstrap, boot	Enterprise Miner (SAS)

*base-base package in R

Feature Selection Methods for Treatment Subset Prediction

FS methods choose a small number of features, say p^*, from the p features. If the goal is to predict two groups of subjects categorized as "Relapse within 5 years" and "No relapse within 5 years", a Wilcoxon or t test comparing the two classes for each feature could be used to select the continuous features which are most highly differentially expressed between the two classes. These features would then be selected for use in the classification method. However this method may not be optimal if the goal is to identify a subset of patients whose tendency to relapse depends on treatment. In this case, a method is desired which specifically preserves information concerning treatment-gene interactions. For example one could, for each gene, subtract the difference in gene expression between relapsed and non-relapsed subjects in the control group from that in the treatment group, and then divide by the standard error. The equation is:

$\{(\hat{\mu}_{11k} - \hat{\mu}_{01k}) - (\hat{\mu}_{10k} - \hat{\mu}_{00k})\} / \sqrt{\hat{\sigma}_k^{\,2}(1/n_{11} + 1/n_{01} + 1/n_{10} + 1/n_{00})}$, $k = 1, \ldots, p$, where $\hat{\sigma}_k^{\,2}$ is the estimate for the pooled variance for gene *k* on the training set, $\hat{\mu}_{gtk}$ are class and treatment arm-specific means for feature *k* on the training set ($g \in \{0,1\}, t \in \{0,1\}$), and n_{gt} is the sample size of subjects specific to class *g* and treatment arm *t* on the training set*;* i.e. - $n_{00} + n_{01} + n_{10} + n_{11} = n_D$.

Then a tuning parameter could be used to select genes which exceed a specified value (or absolute value) for this quantity. Such an approach was used in simulations in Freidlin et al. (2010 [55]).

Categories of Feature Selection Methods

Categories that Saeys, Inza, & Larrañaga (2007 [140]) use for FS methods are the following:

1. Filter, wrapper, and embedded methods

Filter methods such as the Wilcoxon rank sum test, and the ratio of between group to within group sum of squares (Dudoit et al., 2002 [39]), are independent of the classification method. On

the other hand, wrapper methods use a classifier method and attempt to achieve optimization by successively selecting a subset of features for classification. Sequential forward selection (SFS) (Kittler 1978 [94]), and Recursive Feature Elimination using SVM (RFE-SVM, Guyon, 2002 [156]) are two examples of wrapper methods. The methods described in Section 4 for MDR and for SVM are also wrapper methods. Finally, embedded methods such as those in NSC, Random Forests and Boosting, are built into the classification method. Since they are "embedded" in the classification method, they are more naturally described along with the classification method in Section 4.

Filter methods are popular because they can be used for any classification method and often require less computing costs than wrapper methods. Filtering methods may even be needed prior to a wrapper method being applied for classification methods such as MDR and KNN, which are more strongly affected by the curse of dimensionality (see Section 4). Many more examples of each of these three types of DR methods are given in Saeys et al. (2007 [140]).

2. Multivariable and univariable methods

Saeys et al. (2007 [140]) also make a distinction between multivariable and univariable FS methods. Unlike univariable methods such as the Wilcoxon rank sum test, multivariable methods, like FE methods, address feature dependencies. They sometimes also take into account feature interactions. For example the variable importance measure (VIM) used in Classification Trees, Random Forests and Boosting (all outlined in Section 4) can be used as a multivariable FS method that attempts to rank the importance of each feature. As used in trees and Boosting, VIM ranks each variable/feature based on the number of times it is selected in each tree and the amount of reduction in prediction error that results. Further details are given in Section 4.5. This VIM can be easily extended to Random Forests (Section 4.6.1) by averaging over all the trees in the forest. However Random Forests uses a different VIM which involves permuting each splitting variable in a tree and recording the decrease in accuracy. This permutation method nullifies the effect of the feature while keeping the other features. Hastie et al. (2009 [75])

showed that this permutation method differentiates between features less clearly than the other measure. Also they pointed out that permuting a feature is not the same as leaving it out, since when a feature is permuted, there is no opportunity for another feature to take its place in the split.

Lunetta, Hayward, Segal, & Van Eerdewegh (2004 [108]) claimed that this VIM permutation method "takes into account interactions among variables without requiring model specification". At the same time they also discovered that a much larger number of trees are needed to achieve stable estimates of VIM with Random Forests. However Winham et al. (2012 [173]) found that the permutation method had power to detect interactions in low dimensions, but did increasingly poorly as the dimension increases. For HTD, it captured only those interactions associated with strong main effects. They also pointed out that in the Lunetta et al. study "the multiplicative models …have strong marginal components, indicating that the improved performance may be due to the marginal rather than non-linear association".

The RFE-SVM method mentioned earlier is a multivariable FS method. This method recursively eliminates features in an SVM one or more at a time using weights derived from the SVM. In the Guyon study (2003 [69]) it outperformed univariable FS methods. A contributing factor was that it avoids selecting a redundant set of features. Wang et al. (2005 [169]) reviewed multivariable methods including CFS (correlation based feature selection, Hall, 1999 [70]), and ReliefF (Robnik-Siko & Kononenko, 2003 [137]). They found that ReliefF accounts for gene-gene interactions. This method is also briefly highlighted in Section 4.2. CFS is also a multivariable method, but it seeks a subset of features highly correlated with class outcome and uncorrelated with each other (Liu, Li, & Wong, 2002 [102]).

It is important to distinguish between feature-feature dependency, association of features with class outcomes, and feature-feature interactions. Features may be highly dependent with respect to each other and yet have no association with class outcome. CFS seeks out the reverse situation, which can also be true. In turn, features associated with class outcomes may or may not include

feature-feature interactions. That is, an individual feature in a group may have a strong association with class outcome, but changing the level of that feature may not influence the effect that other features have on class outcome. Finally features which are independent of each other may still have highly significant feature-feature interactions. These distinctions are sometimes lost, or at least not sufficiently clarified, in the literature (Saeys et al., 2007 [102]).

Multivariable methods, since they take into account feature dependencies or feature interactions, search over a much higher dimensional space than p unless the search is restricted in some way. Therefore unrestricted multivariable methods may not be appropriate for an initial large scale DR step. Table 3 shows the relationship between number of covariates p and number of possible 1^{st}, 2^{nd} and 3^{rd} order interactions. This relationship also holds for best subsets selection and other unrestricted multivariable DR methods. Fan & Fan (2008 [46]) and Fan & Lv (2010 [47]) provide more guidance and details for selection of FS methods.

Table 3: Relationship between number of features and number of possible interactions.

	Order of interaction		
P	1	2	3
5000	$12.5*10^6$	$20*10^9$	$26*10^{12}$
10000	$50*10^6$	$166*10^9$	$420*10^{12}$
20000	$200*10^6$	$1.3*10^{12}$	$6.7*10^{15}$
40000	$800*10^6$	$11*10^{12}$	$107*10^{15}$

Nonparametric FE

FE methods extract a smaller set of features by combining information from the existing features, while containing that information in fewer dimensions/features. FE can be traced back at least to principle components, introduced by Pearson (1901 [124]). This is perhaps one of the oldest and most well-known and commonly used methods. Several nonparametric FE approaches are briefly described below.

1. Multidimensional Scaling

Multidimensional scaling (MDS), proposed by Richardson (1938 [132]), extracts a lower-dimensional representation of a pairwise distance between features. An example of a pairwise distance is difference in wine tasting scores between raters. Bittner et al. (2000 [11]) used MDS with gene expression data. A thorough discussion of MDS and related methods such as classical scaling can be found in Hastie et al. (2009 [75]).

2. Use of Biological Information for FE

Biological information is increasingly being used to facilitate FE. One of many examples include Gene Set Enrichment Analysis, or GSEA (Subramanian et al., 2005 [154]). GSEA reduces dimensionality by grouping genomic data into biological pathways or gene sets. Statistical tests are then conducted on the gene set instead of the individual genes. Efron & Tibshirani proposed an interesting modification of GSEA called Gene Set Analysis or GSA (2006 [43]), which is implemented in the R package GSA. Tarca, Draghici, Bhatti, & Romero (2012 [155]) introduced a gene set analysis method which down-weights overlapping genes. The method is implemented in Bioconductor package PADOG. Another widely used publicly available tool is DAVID (Huang, Sherman, & Lempicki, 2008 [85]). DAVID includes a "Gene Functional Classification Tool" that can be used as part of an FE method. Winter et al., (2012 [174]) recently proposed an adaptation of Google's PageRank (Page, Brin, Motwani, & Winograd, 1999 [119]) called NetRank which combines correlation of genes with class together with their importance in the TRANSFAC transcription factor network outlined by Matys et al., (2006 [112]) to rank genes as part of a DR step in a classification method. For a review of use of biological information for DR, see Moore, Asselbergs, & Williams (2010 [115]).

3. Clustering Used Together With Feature Extraction

One method often used as part of an FE approach is clustering. Clustering methods group similar objects together. See Gan, Ma, & Wu (2007 [62]) for an in-depth review of clustering methods. Some clustering algorithms are identified in Figures 1a and 2. Clustering is often used

with HTD. Pittman et al. (2004 [126]) used clustering to group genes. Metagenes were then extracted from these groups using FE. The metagenes were then used as features in the classification method. References for development of clustering methods shown in Figure 2 include [50, 52, 67, 103, 151, 170].

Table 4 provides a list and description of some nonparametric DR methods, including software implementations. For more detail about FS methods including filter, wrapper, and embedded methods, and univariable and multivariable FS methods, see Saeys et al. (2007 [59]). For a review of biological information for DR, see Moore, Assselbergs, & Williams (2010 [52]).

Table 4: Some Nonparametric Dimension Reduction Methods.

DR Method	Examples	Description	R/Bioconductor Software	Other Software
FS Univariable	Wilcoxon test	Ranks features based on Wilcoxon test.		
	BSS/WSS	Ranks features based on ratio of Between-Sum-of-Squares to Within-Sum-of-Squares (Dudoit, 2002).		
FS Multivariable	CFS	Correlation Based FS. Seeks a subset of features highly correlated with outcome and uncorrelated with each other (Liu, Li, Wong 2006).	RWeka,	WEKA
	ReliefF	See Robnik-Siko and Kononenko, 2003 and Wang, Tetko, Hall and others, 2005.		
Other FS and FS software		FSelector (Romanski, 2012), Sure Independence Screening (Fan and Lv, 2008), caret (Kuhn, 2013).	FSelector, SIS, caret	
Wrapper Methods	Best Subset Selection	Uses a classifier method to select the best subset of features from a larger set of features (Richardson, 1938), also see Hastie et al., 2009, Bittner et al., 2000. Computationally impractical for large p.	leaps	Proc phreg, proc reg (SAS)
Embedded Methods		The classification method itself incorporates a feature ranking or dimension reduction method.		
	Variable Importance Measure	For classification and regression trees; easily extendable to Random Forests and Boosting; bases importance of each variable/feature on the number of times it is selected in each tree and the amount of reduction (improvement) in prediction error. The relative importance of each variable uses the square root of the VIM and scales them by using 100 for the most important variable.	randomForest	Proc split (SAS)
FE	MDS	Multi-Dimensional Scaling. Extracts a lower-dimensional representation of a pairwise distance between features.	cmdscale (R base), isoMDS(MASS)	Proc mds (SAS)
Biologically Based FE	GSEA	Gene Set Enrichment Analysis. Reduces dimensionality by grouping genomic data into gene sets or biological pathways (Subramanian, Tamayo, Mootha and others, 2005).	GSEABase, GSEAlm, limma	
	GSA	Gene set Analysis. Adaptation of GSEA, Efron and Tibshirani	GSA	
	PADOG	Pathway Analysis with Down-weighting of Overlapping Genes, (Tarca, Draghici, Bhatti and others, 2012).	PADOG	
	DAVID	Includes a Gene Functional Classification Tool that can be used as part of an FE method, (Huang, Sherman, and Lempicki, 2008).		DAVID
	NetRank	Adaptation of Google's PageRank which combines correlation of genes with class together with their importance in the TRANSFAC transcription factor network (Matys, Kel-Margoulis, Fricke and others, 2006).		GUILD

DR- Dimension Reduction; FS – Feature Selection; SVM-Support Vector Machine; VIM-Variable Importance Measure

5. Nonparametric Classification Methods

5.1 Multifactor Dimensionality Reduction (MDR)

MDR (Ritchie et al. (2001 [135]) was originally developed to detect multi-locus gene-gene and gene-environment interactions associated with disease. Lou et al. (2007 [47]) extended it to include continuous covariates. Figure 3a shows an example where the number of loci is equal to 2, and each locus has 3 genotypes. In this example there are nine possible loci-genotype combinations. The upper left cell of Figure 3a represents the AA-BB combination. The steps for MDR are:

1. In each cell the proportion of disease versus disease-free individuals are calculated.
2. A single cut-off is then found to identify high-risk vs. low-risk cells using nested CV.
3. The optimal number and combination of loci on the training set is determined using best subsets.
4. Then prediction error is estimated on the final validation set using the final model chosen.

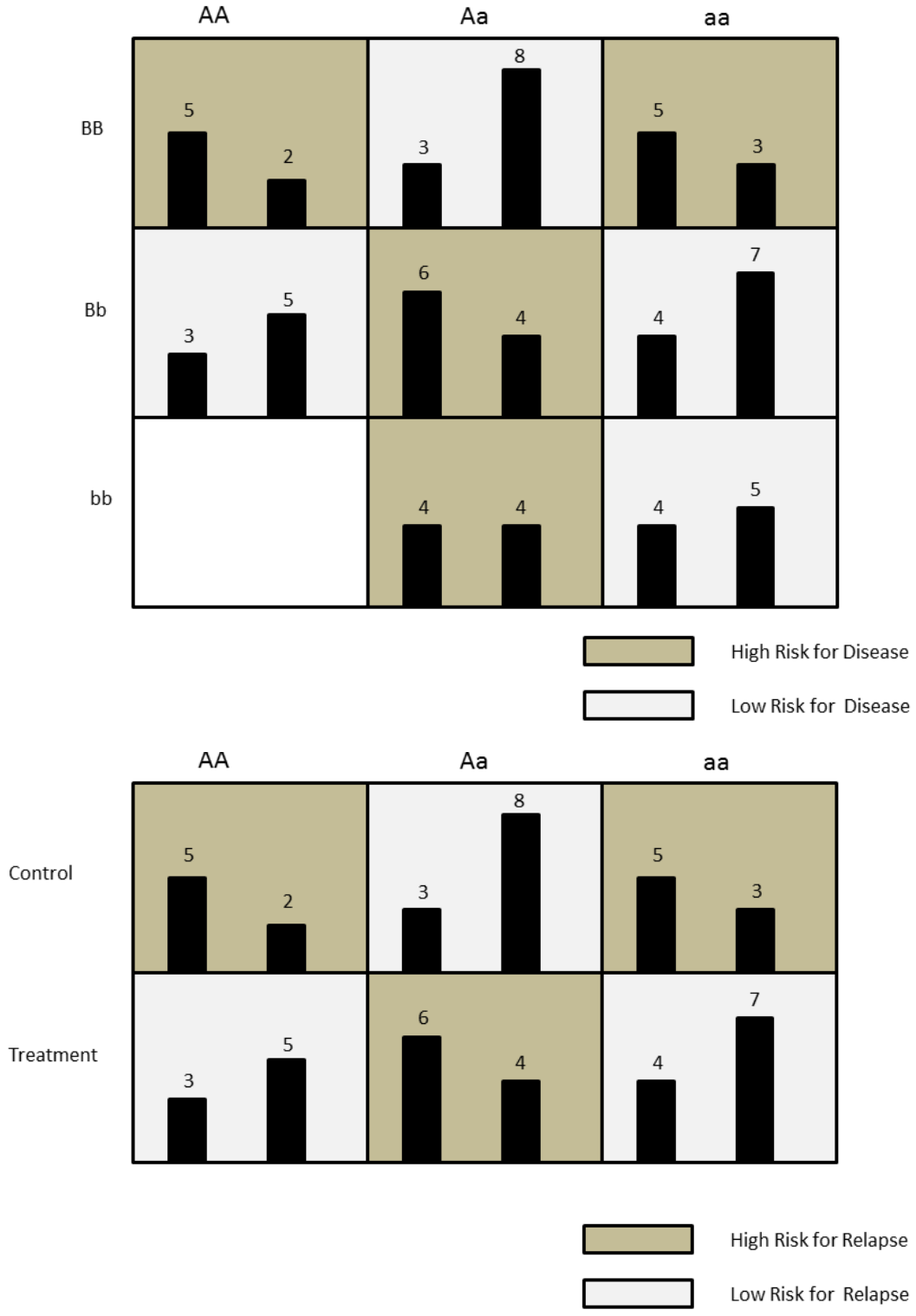

Figure 3a and b: Use of MDR to determine high-risk loci-genotype combinations on a training set, for purposes of classification.

MDR and DR

Richie et al. (2001 [135]) claimed that MDR is a DR method which reduces a p-dimensional factor space to a one-dimensional space by categorizing each possible multi-locus genotype combination as high-risk or low-risk. However Park & Hastie (2008 [120]) challenged this assertion on three fronts:

1. For every *k* less than or equal to *p*, MDR searches for the optimal *k* loci-genotype combination.
2. MDR searches for the optimal number of *k* factors.
3. MDR searches for the optimal cutpoint to classify individuals in each loci-genotype combination into high or low risk.

They showed through simulation that the effective degrees of freedom are actually much greater than 1.

Advantages and Disadvantages of MDR

1. Since proportions are estimated for each combination, there is no assumption of linearity, and the degree of interaction is determined agnostically- that is without any linearity assumptions or hierarchy regarding main effects and interactions.
    a. The advantage of this is that higher order interactions between loci are estimated without first having to account for main effects. This is a claimed advantage in the presence of epistasis.
2. The order of interaction is restricted to *k*-1 by the number of features k included in the model.
    a. There may be some advantages in restricting the order of interactions to be much less than the number of features. In Section 4.6.2 an ensemble method is discussed which follows this approach.
3. Figure 3b shows how MDR can be extended from detection of gene-gene interactions or gene-environment interactions in a case-control setting to gene- treatment interactions in a clinical study.

4. Generalized MDR, or GMDR (Lou et al., 2007 [106]) broadened use of MDR by allowing for both continuous and discrete covariates, as well as continuous phenotypes.
5. Ritchie, Hahn, & Moore (2003 [136]) found that in the presence of 50% genetic heterogeneity, the power of MDR is very limited and stated that "extending MDR to address genetic heterogeneity should be a priority".
6. MDR is strongly affected by the curse of dimensionality. If each of *p* loci has 3 genotypes, then the number of table cells for which proportions need to be calculated can be as large as 3^p. Therefore in presence of HTD DR methods may be needed before applying MDR (Cordell, 2009 [30]). Methods such as Genetic Algorithms (Holland, 1975 [82]) which do not require an exhaustive search have also been used with MDR in an attempt to address this issue.
7. Software is readily available for MDR and GMDR (http://sourceforge.net/projects/mdr/ accessed 10/28/2012).
8. The method is straightforward, and it makes few assumptions regarding interactions. It is often used as a standard for gene-gene interactions involving SNP's (single nucleotide polymorphisms). It has also been used as a complement to logistic regression.

MDR and Logistic Regression

Ritchie et al. (2001 [135]) claimed an advantage of MDR over logistic regression in the presence of epistasis since main effects do not need to be accounted for first. Logistic regression used for classification (LCA) is a parametric method. However at this point it is interesting to note that LCA using one loci-genotype cell as a reference, and dummy variables for every other loci-genotype combination, could be used to mimic the MDR approach. Cut-offs could be determined in a similar manner, and inclusion of continuous covariates could be included without need for development of additional methods. Comparisons with other genetic models could then also be made using logistic regression; for some guidance see for example Sasieni (1997 [141]) and Freidlin, Zheng, Li, & Gastwirth (2002 [53]). Methods to account for heterogeneity, such as

the quasi-likelihood approach for inclusion of a dispersion parameter (McCullagh & Nelder, 1989 [113]), are also available.

MDR requires no modification to be used as a treatment subset prediction method for categorical factors. One simply uses treatment as an environmental factor, as shown in Figure 3b. See Table 2 for further information on MDR. Park & Hastie (2008 [120]) provide an in-depth review of MDR and also propose a penalized logistic regression method for detection of gene-gene interactions.

5.2 K Nearest Neighbors (KNN)

While the MDR method was first developed for categorical data, K Nearest Neighbors (KNN) is more naturally used with continuous data. It classifies the i^{th} subject with covariate or feature vector $\{\boldsymbol{x}_i, \boldsymbol{z}_i\}$ in the validation set based on the *K* subjects with feature vectors nearest to it in the training set. Nearness is based on a distance metric for the subject-specific feature vectors, such as Euclidian distance, Mahalanobis distance, absolute value (Manhattan) distance, or 1 minus correlation (Dudoit et al., 2002 [39]). The class assigned to the subject in the validation set is based on majority class vote of the nearest subjects in the training set. For this reason *K* is usually chosen as an odd integer to avoid ties. Pre-processing of data is important for KNN since most distance measures used do not take into account variance. One method is to scale each feature so that it has overall mean 0 and unit variance (Hastie et al., 2009 [75]); *K* can be chosen through methods such as nested CV; *K*=1 or *K*=3 are common choices.

As with MDR, a separate DR step is needed in an HTD setting before applying KNN. Used in this way, KNN outperformed most alternatives in the Dudoit et al. (2002 [39]) comparison study. See Table 2 for KNN strengths, weaknesses and software.

Distance Measures for K Nearest Neighbors

The proportion of nearest neighbors in class $g = 1$ is used as a distance measure for classification in the KNN method. R Code in supplemental material shows how this distance can

be extracted. The distance of the *K* nearest neighbors for each class may also be used as a distance measure. If the *K* nearest neighbors in class $g = 1$ are closer to subject i than the *K* nearest neighbors in class $g = 0$, then this distance could be used to allocate subject i to class $g = 1$. Section 7 shows how these distances can be applied to treatment subset prediction.

Classification Methods and Clinical Covariates

Handling of mixed variables is not regarded as a strength for KNN. However the PV example in the work by Tibshirani & Efron (2002 [159]) used the following approach to combine genomic and clinical covariates:

1. An NSC predictor (Tibshirani e al. 2003 [161]) is developed using CV on the genomic data only.
2 The NSC predictor is then included as a feature, along with clinical covariates in a logistic regression model over all the data set.
2. The added value of the genomic NSC feature is then evaluated in the logistic regression model.

The authors proposed a bootstrapping procedure to address issues associated with effective degrees of freedom. Hofling & Tibshirani (2008 [81]) and Chang et al. (2005 [23]) also proposed modified PV methods. Approaches such as PV could be used to combine clinical covariates and HTD features in a classification method for treatment subset prediction. There is nothing restricting PV to HTD classifiers such as NSC.

5.3 Kernel Density Analysis (KDA) Used for Classification

KDA can be thought of as an extension of KNN (Hastie et al., 2009 [75]). A nearest neighbor kernel density is a non-smooth kernel density.

Univariable Kernel Density Estimation

Kernel Densities are nonparametric densities formed through use of kernel functions. A one-dimensional kernel density can be expressed as:

$$\hat{f}(x_0) = \frac{1}{nb}\sum_{i=1}^{n} K\left(\frac{x_{i1} - x_0}{b}\right),$$

where n is the sample size, K is the kernel, b is the bandwidth, x_0 refers to a point in this one-dimensional covariate space, and x_{i1} refers to a one-dimensional covariate vector for the i^{th} subject. The Gaussian kernel is $K_b(x_0) = \phi\left(\frac{|x_{i1} - x_0|}{b}\right)$, where ϕ is the probability density function for the Gaussian distribution. For more details see Venables & Ripley (2002 [165]) and Hastie et al. (2009 [75]).

Univariable Kernel Density Analysis for Classification

To use KDA for classification with one feature, the densities and prior class probabilities must be calculated for each of the classes. If the feature is genomic, the equation is

$$LR_i = \frac{\Pr(g=1, z_{i1})}{\Pr(g=0, z_{i1})} = \frac{\Pr(g=1)\Pr(z_{i1} \mid g=1)}{\Pr(g=0)\Pr(z_{i1} \mid g=0)}.$$

If class densities and probabilities are estimated from the training set, the ratio is estimated as

$LR_i = \frac{\hat{\pi}_1 \hat{f}(z_{i1} \mid \hat{\boldsymbol{\theta}}_1)}{\hat{\pi}_0 \hat{f}(z_{i1} \mid \hat{\boldsymbol{\theta}}_0)}$, where $\hat{\pi}_g \in \{\hat{\pi}_1, \hat{\pi}_0\}$ are the prior class probabilities estimated from the training set, and $\hat{\boldsymbol{\theta}}_g \in \{\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_0\}$ are the vectors of class-specific parameters used to form the density, such as bandwidth and sample size, again calculated from the training set for each class *g*. A cut-off of 1 can be used for the two classes, or it can be based on CV.

Note that this ratio is the distance measure used to make a classification decision. A ratio of 1 indicates a subject on the border between the two classes, and ratios near 0 or much greater than 1 indicate stronger evidence for the subject being in one of the two classes. The log of this ratio, LLR_i, can also be used, in which case the class border is 0, and large positive or negative values indicate stronger evidence.

Numeric Example

In the training set, 30 patients responded to treatment and 30 patients did not. There is one value for gene expression for each patient. Gene expression for non-responders is, after appropriate normalization, normally distributed with mean 0 and standard deviation 1. For responders (class $g = 1$), it is normally distributed with mean 1 and standard deviation 2. Now on the validation set class is unknown for patient i, but gene expression=1.5. Using set.seed(234) and default settings for functions density (for kernel density calculation) and approxfun in R (version 3.02) for interpolation between density values:

$$LR_i = \frac{\hat{\pi}_1 \hat{f}(z_{i1} = 1.5 \mid \hat{\boldsymbol{\theta}}_1)}{\hat{\pi}_0 \hat{f}(z_{i1} = 1.5 \mid \hat{\boldsymbol{\theta}}_0)} = \frac{0.5 * 0.1915}{0.5 * 0.1408} = 1.36.$$

Since the ratio is greater than 1, and class sizes are equal, the patient can be predicted to respond to treatment. The class boundary ($LR_i = 1$) occurs at about $z = 1.298$.

Multivariable Kernel Density Analysis for Classification

The naïve Bayes assumption is that features within a class are independent. Extension of KDA to a multi-dimensional feature space is made easier if this assumption is employed. Various authors including Hastie et al. (2009 [75]) and Bickel & Levina (2004 [9]) have noted the success of naïve Bayes methods, despite the fact that the independence assumption is not generally true. Using this assumption, densities can be calculated conditional on class for each feature individually, and then the *p*-dimensional density is the product of the individual densities. The density conditional on class g for a new subject with genomic feature vector $\boldsymbol{z}_i = z_{i1}, ..., z_{ip}$ is then $\hat{f}(\boldsymbol{z}_i \mid \hat{\boldsymbol{\theta}}_g) = \prod_{k=1}^{p} \hat{f}(z_{ik} \mid \hat{\boldsymbol{\theta}}_{gk})$, assuming independence. Individual subjects can be allocated as before using the plug-in likelihood ratio $LR_i = \frac{\hat{\pi}_1 \hat{f}(\boldsymbol{z}_i \mid \hat{\boldsymbol{\theta}}_1)}{\hat{\pi}_0 \hat{f}(\boldsymbol{z}_i \mid \hat{\boldsymbol{\theta}}_0)}$. Densities can also be based on a

combination of clinical and genomic features $\{\boldsymbol{x}_i, \boldsymbol{z}_i\}$, and categorical features as well as continuous features. Categorical features can be incorporated using histogram estimates (Hastie et al., 2009).

Kernel Density Classification for Two or More Classes

If there are two or more classes $g = 0,1,...,G-1,$ then subject *i* can be allocated to the class with the highest posterior probability; i.e. -

$\hat{g}_i = \arg\max_g \{Pr_{post,gi}\} = \arg\max_g \left\{ \dfrac{\hat{\pi}_g \hat{f}(\boldsymbol{z}_i \mid \hat{\boldsymbol{\theta}}_g)}{\sum_{g=0}^{G-1} \hat{\pi}_g \hat{f}(\boldsymbol{z}_i \mid \hat{\boldsymbol{\theta}}_g)} \right\}$, where π_g is the prior probability for class *g*. Details can be found in Hastie et al. (2009 [75]). Now denote $\Pr_{post,i}$ as the posterior probability for class *g*=1 for subject *i*. Then $Pr_{post,i}$ is a distance used to make a classification decision. For two classes, a $Pr_{post,\mathrm{i}} = 0.5$ is on the border between the two classes. Probabilities away from this boundary in either direction indicate stronger evidence for one of the two classes. Approaches employing both LLR_i and $Pr_{post,i}$ distances are applied to treatment subset prediction in Section 7.

Other KDA Methods

There are also other approaches besides naïve Bayes to obtain multivariable kernel densities. Wand and Jones (1993 [168]) compared smoothing parameterizations for bivariable kernel densities. The R package ks (Duong, 2007 [41]) allows kernel smoothing from one to 6 dimensions. The function kda.kde in package ks performs KDA for up to 6 dimensions. Still another KDA approach involves difference in densities. This approach can be traced back to Hall and Wand (1988 [71]). More recently Kim and Scott (2010 [93]) introduced L_2 kernel classification for difference in densities. For dimensions $p \geq 15$, they introduced a regularization

parameter, which their simulations showed enables performance similar to other HTD methods such as SVM.

5.4 Support Vector Machines

The KDA method uses kernel functions to estimate class densities, which can then be used to find a decision boundary. SVM's, on the other hand, use kernel functions to find a decision boundary between the two classes. When two classes are linearly separable, a decision boundary is found that maximizes the separation between the two classes. See Hastie et al. (2009 [75]) and Cristianini & Shawe-Taylor (2000 [32]) for more detail.

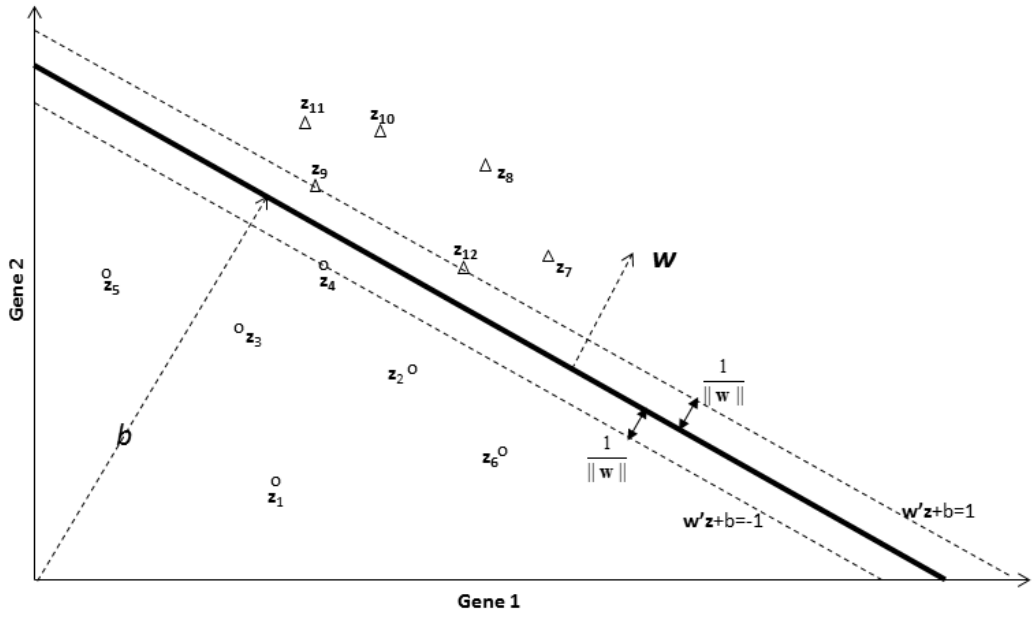

Figure 4: A support vector machine for a 2-dimensional feature space: (gene expression for gene 1 and gene 2) which linearly separates the two classes, which are represented by the "Δ" subjects and the "o" subjects. In this case there is a separating hyperplane, meaning that there are no subjects on the wrong side of the line. Therefore only three subjects-specific feature vectors are needed as support vectors. In this case they are $\mathbf{z}_4$, $\mathbf{z}_9$, and $\mathbf{z}_{12}$.

Support Vector Machines for Two Linearly Separable Classes

1. Figure 4 shows training set data points consisting of the fixed $\mathbf{z}_i$ vectors, where $\mathbf{z}_i \in \mathcal{R}^2$.

   Extension to higher dimensions $\mathcal{R}^p$ is straightforward. The $\mathbf{z}_i$ are the feature vectors for each

subject i in the training set, $i = 1,...,n_1; n_1 = 6$ for the o class, and $i = n_1 + 1,...,n_1 + n_2$ for the Δ class.

2. The horizontal and vertical axes represent gene expression for gene 1 and gene 2. Quadratic programming is used to conduct a search to find the three support vectors, which are the three subject-specific feature vectors that minimize Euclidian distance $||\boldsymbol{w}||$ in the equations for the two margin boundaries passing through the three support vectors ; the equations are $\boldsymbol{w}'\boldsymbol{z} + b = 1$ and $\boldsymbol{w}'\boldsymbol{z} + b = -1$.
3. In Figure 4 the three support vectors are $\boldsymbol{z}_4$, $\boldsymbol{z}_9$ and $\boldsymbol{z}_{12}$.
4. The decision boundary is then the line $\boldsymbol{w}'\boldsymbol{z} + b = 0$.
5. The classification decision for subject i in the validation set is then $\hat{g}(\boldsymbol{z}_i) = \text{sign}(\boldsymbol{w}'\boldsymbol{z}_i + b)$.

From this it follows that the distance used to make a classification decision for subject *i* in the validation set is $d_i = \boldsymbol{w}'\boldsymbol{z}_i + b.$ A large magnitude distance in the negative/positive direction indicates subject i is further away from the class boundary, and offers stronger support for the subject being in one of the two classes.

Numeric Example

Subject i in validation set has gene expression 1.1 and 1.3, for two genes, after appropriate normalization and scaling. Also from the training set are calculated coefficient vector $\boldsymbol{w} = (-0.5, 1.8)'$, and $b = 3.$ Then $\boldsymbol{w}'\boldsymbol{z}_i + b = 4.79,$ and $\hat{g}_i = \text{sign}(4.79) = 1.$ R code which calculates coefficients and classification distance is included in supplemental material.

Support Vector Machines for Non-Linearly Separable Classes

In the case where it is not possible to linearly separate the two classes, it is possible to allow additional support vectors on the wrong side of their respective margin boundaries with slack variables which are the distances of the corresponding support vectors from their respective class margin boundary. In this situation the same approach is used to find the support vectors as before,

but with the additional constraint of minimizing the slack variable distances using a penalty term. More detail is given in Hastie et al. (2009 [75]). If the decision function itself is nonlinear one can often find a linear separating hyperplane in a higher dimensional space using specific kernel functions. See Vapnik (1998 [164]) for more details. Chen et al. (2008 [25]) compared various SVM methods with MDR in their ability to detect gene-gene interactions. More detail is given in Hastie et al. (2009 [39]).

5.5 Classification and Regression Trees

Classification and regression trees are a way of recursively partitioning data, one feature at a time, into more homogeneous subgroups. The root node is at the top of the tree and contains the entire data set. Starting from this root node, the data is recursively partitioned into daughter nodes which contain finer and finer subsets of the data, one feature at a time, based on which feature subsets the previous node into the most homogeneous subgroups. The terminal nodes contain the final partitions for all the data. Each subject in a validation set can then be allocated into one and only one of these final partitions based on a classification tree grown on a training set.

Figure 5: Flowchart of ASD nested cross validation method. *Training set can refer both to Stage 1 patients, and to the training sets nested within the Stage 1 data. Similarly, validation set can refer to the corresponding nested validation sets in the Stage 1 data, or to the final validation set – the Stage 2 patient data.

Feature and Cutpoint Selection

Feature selection for partitioning a node is performed using criteria such as reduction in impurity. Impurity measures are at a maximum when the two partitioned daughter nodes have the same proportion of subjects in each of the two classes. Impurity is at a minimum when the two daughter nodes perfectly separate the two classes. Reduction in impurity is measured by the difference in impurity between the parent node and the average impurity of the daughter nodes.

Examples of impurity measures include the Gini index (Gini, 1912 [64]) and entropy (Hastie et al., 2009 [75]).

Statistical Interactions in Classification Trees

One advantage of classification trees often pointed out is that they are able to identify complex interactions, including interactions involving different types of covariates, such as interactions between treatment, clinical covariates, and HT covariates. In fact early development of classification trees were motivated by interaction detection (Morgan & Sonquist, 1963 [75]). Strobl, Malley, & Tutz (2009 [152]) provide an in-depth review of statistical interactions in trees.

If a tree contains only one split, then no interaction can be included since only one feature is used in a split. Such a tree is often called a stump and has a depth $h = 1$. However if a tree contains two splits, then a two-way interaction is involved if the change in class proportions due to the second split depends on the partitions for the first split. A concise illustration is shown in Figure 4 of Strobl et al. (2009 [152]). Interactions may also be present when a split occurs in one daughter node, but not the other. Trees with greater depth (more recursive partitions) are capable of including successively higher order interactions.

If specific interest lies in treatment-covariate interactions to identify a subset of patients more prone to relapse, trees based on unmodified CART® (Breiman, Friedman, Olshen & Stone, 1984 [15]) have limitations since they may model interactions that are not of primary interest. However Schmoor, Ulm, & Schumacher (1993 [145]) modified the approach. They first used CART to identify covariates associated with disease prognosis. They then tested for treatment heterogeneity within the subgroups identified by these covariates, but results were not significant. Finally they modified the splitting procedure and based it on treatment interactions. Using this approach they found that Karnofsky Index and age had significant treatment interactions. Karnofsky Index was not included in the first approach because its main effect was not strongly associated with prognosis. Later Loh (2002 [104]) proposed a method called GUIDE which enables detection of local interactions in regression trees. Su, Tsai, Wang & Li (2009 [153])

further developed use of interaction trees. See Loh (2011 [105]) for a review of classification and regression trees which includes interaction trees.

5.6 Ensemble Methods

"Plans fail for lack of counsel, but with many advisers they succeed" (New International Version, Prov. 15:22). This is the basic idea behind ensemble methods, which combine classifiers in different ways to improve stability and performance of individual classifiers. Some ensemble methods are reviewed below.

5.6.1 Random Forests

Individual classification trees can be unstable, since small changes in the data can have a large effect on the predicted class labels. Also the order of the covariates or features used to split nodes can have a large influence on the predicted class label. Random Forests is an ensemble classification method that incorporates tools such as bootstrapping and Random Features (random sampling of features selected for potential splitting of a node; Breiman, 1999 [17]) to improve stability of classification trees. The steps are as follows:

1. Divide data with n subjects into training set with n_D subjects and validation set with n_V subjects.
2. From training set with sample size n_D, use bootstrapping to generate a nested training set of size n_D.
3. From the nested training set grow a tree using an impurity measure to select from a small set of m candidate random features at each node.
4. Classify each subject not included in the bootstrap sample using this tree.
5. Repeat the above steps to grow B trees.
6. Using the B trees, classify each subject using the majority vote from all the trees.

If the data set is divided up into a training and a validation set, then B trees can be grown on the training set and used to classify each subject i in the validation set. In that case the

classification decision for subject i can be written as $\hat{g}_i^{(ensemble)}(\boldsymbol{x}_{i,}\boldsymbol{z}_i) = I(\bar{p}_{i1} > 0.5)$, where I is the indicator function and $\bar{p}_{i,G=1}$ is the proportion of the B trees that classified subject i into $g = 1$. In this case the distance used to make the classification decision is $\bar{p}_{i,G=1}$. A proportion away from 0.5 and near 0 or 1 indicates stronger support for $g_i = 0$ or $g_i = 1$. R code is included in supplementary material which automatically calculates this proportion. The classification decision can also be written as $\hat{g}_i^{(ensemble)} = \text{sign}\left(\sum_{b=1}^{B} \hat{g}_i^{(b)}(\boldsymbol{x}_{i,}\boldsymbol{z}_i)\right)$, where $\hat{g}_i^{(b)} \in \{-1,1\}$, and $b = 1,...,B$ denotes the b^{th} successive classifier. The classification distance is then: $\sum_{b=1}^{B} \hat{g}_i^{(b)}(\boldsymbol{x}_{i,}\boldsymbol{z}_i)$. A higher magnitude of this quantity in either the positive or negative direction is then an indication of how far away subject i is from the border between the two classes.

Section 7 shows how these distances can be extended to treatment subset prediction. See Foster, Taylor, & Ruberg (2011 [51]) for an existing Random Forests method which uses clinical covariates, including demographic and survey data, to predict a subset of patients having an enhanced treatment effect.

Bumping and Interaction Effects

Bumping (Tibshirani & Knight (1999 [158]), like Random Forests, uses bootstrapping to build an ensemble of trees, but unlike Random Forests it selects the one tree with the lowest prediction error on the training set. If enough trees are used, Bumping can often find a split to minimize prediction even if are no main effects. An example is given in Hastie et al. (2009 [75]). Lipkovich, Dmitrenko, Denne, & Enas (2011 [101]) developed a method closely related to Bumping to predict multiple treatment subgroups in clinical trials.

5.6.2 Boosting

Boosting employs an ensemble of individual classifiers, and each classifier after the first one is reweighted based on the previous classifier.

1. The individual classifier is selected from a category of classifiers; for example, logistic classification trees (Chan & Loh, 2004 [22]).
2. The first individual classifier, $\hat{g}_i^{(1)}(\boldsymbol{x}_i, \boldsymbol{z}_i)$, has weights initialized for each subject i in the training set: $w_i^{(1)} = 1/n_D, i = 1,...,n_D$. This classifier is then fit to the training set using these $w_i^{(1)}$ weights.
3. If subject *i* is misclassified for the b^{th} classifier, its weight ($w_i^{(b+1)}$) is increased, and subjects classified correctly have their weights decreased.
4. A coefficient $\beta^{(b)}$ is determined for the b^{th} classifier based on weighted classification performance. Individual classifiers with better performance are given larger positive coefficients.
5. Subjects in the validation set are then allocated to one of two classes based on the ensemble classifier: $\hat{g}_i^{(ensemble)} = sign\left(\sum_{b=1}^{B} \beta^{(b)} \hat{g}_i^{(b)}(\boldsymbol{x}_i, \boldsymbol{z}_i) \right),\ i = n_D + 1,..., n_D + n_V.$

A shrinkage constant γ much smaller than 1 has been shown to prevent overfitting. The equation then becomes:

$$\hat{g}_i^{(ensemble)} = sign\left(\gamma \sum_{b=1}^{B} \beta^{(b)} \hat{g}_i^{(b)}(\boldsymbol{x}_i, \boldsymbol{z}_i) \right).$$

The distance used in Boosting for classification is then $\gamma \sum_{b=1}^{B} \beta^{(b)} \hat{g}_i^{(b)}(\boldsymbol{x}_i, \boldsymbol{z}_i)$. Large negative distances indicate support for allocating subject i into class $g = -1,$ and large positive distances indicate support for class $g = 1.$ Differences in this distance across treatment arms may then be evidence for subject i being sensitive to treatment. R code is included in supplementary material which calculates the Boosting classification distance for a subject in the validation set.

More information about Stochastic Boosting and Gradient Boosting can be found in Hastie et al. (2009 [75]), Friedman (2002 [60]), and Elith, Leathwick, & Hastie (2008 [44]).

Boosting and Interaction Order

As explained in Section 5.5, the order of interactions in a tree is restricted by the depth of the tree. A tree of depth *h* allows interactions up to order $h-1$. This fact can be exploited in Boosting to compare models allowing different interaction orders. This can be done by comparing prediction error of boosted trees of a specific depth to that of boosted tree stumps, which allow no interactions. Care must be taken to allow a sufficient number of trees to minimize prediction error. Several authors have indicated that trees with depth in the range of 5 to 9 can be used in this manner to assess order of interactions.

Boosting and Interpretability

The VIM developed for classification trees can be easily used in Boosting to find the most influential variables. Elith et al. (2008 [44]) showed partial dependence plots of main effects and interactions based on VIM. The main effects plots showed the most influential main effects with other variables fixed at their average value. The same approach was used for contour plots of two variables to assess interactions. Contour plots which include the two-variable interaction are compared to contour plots from boosted stump models which cannot include interactions. The difference in the two plots gives a visual comparison of the importance of the interaction on the marginal effect.

5.7 Other Methods

A brief mention is given here of some other classification methods which may be used to identify subsets of patients responding differently to treatment.

Ensemble methods reviewed in this study, such as Random Forests and Boosting, employ individual classifiers of the same category. Stacking (Wolpert, 1992 [175]) uses an ensemble of different classifiers, and this approach has been shown to improve performance under many

situations. Leblanc & Tibshirani (1996 [98]) developed a Bootstrap approach for Stacking. A related approach is Bundling (Hothorn and Lausen, 2005 [84]), implemented in the R package ipred (Peters & Hothorn, 2012 [125]). For a recent in-depth review of ensemble methods, including combination methods, combination learning, ensemble diversity, and many other approaches, see Zhou (2012 [182]).

Cordell (2009 [30]) reviewed a Bayesian method to assess gene interactions (BEAM, or Bayesian Epistasis Association Mapping, Zhang & Liu, 2007 [180]). Software implementations of BEAM can be found at http://sites.stat.psu.edu/~yuzhang/ (accessed 10/13/2012). Huang et al. (2004 [86]) introduced FlexTree, a tree-structured methods for detecting gene-gene interactions. This was also compared to MDR and the penalized logistic regression method in the Park & Hastie (2008 [120]) work. Logic regression (Ruczinski, Kooperberg, & LeBlanc, 2004 [138]) has been used in sequence analysis and to explore high-order interactions in HTD. Foster et al. (2011 [51]) introduced the "Identical Twins" approach. This method borrows an approach from counterfactual models (Ginsberg, 1986 [65]). Specifically, it assumes "there are two possible outcomes for each person (one under each treatment assignment), only one of which can be observed (Foster et al. 2011 [51])." The KDA classification method described in Section 4.3 used a counterfactual approach, as did Rai, Pan, Cambon and others (2013 [129]).

For an introduction to Neural Networks, see Duda et al. (2001 [38]), Hastie et al. (2009 [75]), and Venables & Ripley (2002 [165]). For an application of Neural Networks to cancer informatics, see Cruz & Wishart (2006 [33]). For Bayesian Neural Networks ,see Neil & Zhang (2006 [116]). A Bayesian Neural Network model entered in the NIPS 2003 challenge (http://nips.cc/) had superior classification performance to Boosting and Random Forests models (Hastie et al. 2009 [75]). One disadvantage of Neural Networks in a clinical setting is difficulty of interpretation (Hastie et al., 2009[75]). Chen et al. (2008 [25]) argued that Neural Networks are not appropriate for study of gene-gene interactions. On the other hand, ensemble methods such as Boosting can employ Neural Networks as individual classifiers in a way similar to trees. An

implementation of Neural Networks is found in R package nnet (Venables & Ripley, 2002 [165]). There is also extensive literature on rule-based and fuzzy rule-based classifiers (Kosko 1993 [96]).

6. Generalizability of Classification Methods

Split sample, CV, and bootstrap methods discussed above are internal validation methods. In these methods, the same data set used for training is also used for validation and prediction error, even though paired training and validation sets are themselves non-overlapping. Internal validation is important and necessarily precedes other more all-encompassing forms of validation. Altman, Vergouwe, Royston & Moon (2009 [1]) provide an in-depth review of internal and external validation.

High dimensional genomic and proteomic data magnify challenges associated with external validation because the field is changing rapidly and platforms, technologies, and methods become obsolete. For example in the early 2000's HTD commonly involved messenger RNA (mRNA) expression. However it has increasingly been shown that miRNA's, copy number, SNP's, proteins, and methylation status play an important role in cancer and melanoma. Next Generation Sequencing (NGS) platforms are replacing traditional microarray platforms in many applications. Since NGS often involves count data, this may mean that, in these situations, models for count data, such as Poisson models, may be used in place of other classification methods.

While developments in all these areas are bringing tremendous opportunity for advancement in treatments of melanoma and other cancers, they also pose challenges regarding external validation. Classification methods that are robust to changes in platforms may be more generalizable. To that end, Maglietta et al. (2010 [110]) proposed a method which involves rules for selecting sets of genes/miRNA's.

7. Discussion and Conclusions

The specific ASD described in Freidlin & Simon (2005 [54]) uses weighted voting and single gene logistic models which include treatment-gene interaction. The tuning parameter set consists

of a tuning parameter η to select predictive genes from the training set, a tuning parameter R to determine the cutoff for the magnitude of the treatment arm odds ratio (OR) for a selected gene, and a tuning parameter H to determine the number of selected genes with OR exceeding R necessary to predict a patient to be sensitive. More details are given in Figure 5.

In this specific ASD design, the treatment arms OR is used as a distance measure to compare response to treatment across the two arms. In the ASD implementation by Scher, Nasso, Rubin, & Simon (2011 [144]) the log hazard ratio was used in the same way. Below are examples of nonparametric distances used in treatment subset prediction methods.

Treatment Subset Prediction Using Kernel Density Analysis and Random Forests

Distances for KDA used for classification have been described in Section 5.3. For example an LLR close to 0 indicates a subject on the borderline between the two classes, and large positive or negative values indicates the extent to which a subject may be in one class or the other. The difference in the LLR $(DLLR)$ between the treatment and control group is then a measure of how sensitive the subject in the validation set is to treatment. The single-gene model is:

$$DLLR_{ik} = \left(LLR_{i1k}\right) - \left(LLR_{i0k}\right) = \log \frac{\hat{\pi}_{11}\hat{f}(z_{ik} \mid \hat{\boldsymbol{\theta}}_{11k})}{\hat{\pi}_{01}\hat{f}(z_{ik} \mid \hat{\boldsymbol{\theta}}_{01k})} - \log \frac{\hat{\pi}_{10}\hat{f}(z_{ik} \mid \hat{\boldsymbol{\theta}}_{10k})}{\hat{\pi}_{00}\hat{f}(z_{ik} \mid \hat{\boldsymbol{\theta}}_{00k})}.$$

The method for calculating LLR_{itk} is the same as in Section 5.3, except that it is now calculated separately for each treatment arm, $t = 1$ and $t = 0,$ and for each selected gene *k*. This in turn necessitates estimated prior probabilities $\hat{\pi}_{gt}$ and parameters $\hat{\boldsymbol{\theta}}_{gtk}$ in place of $\hat{\pi}_g$ and $\hat{\boldsymbol{\theta}}_g$, as these quantities are now calculated separately not only for each class *g*, but also for each treatment arm *t* (and $\hat{\boldsymbol{\theta}}_{gtk}$ for each selected gene *k)*, using class and treatment arm-specific information from the training set. Alternatively, the estimate for the single gene posterior OR is

$$OR_{post,ik} = \frac{Pr_{post,i1k} / \left(1 - Pr_{post,i1k}\right)}{Pr_{post,i0k} / \left(1 - Pr_{post,i0k}\right)}, \text{ where } Pr_{post,itk} \text{ is similar to } Pr_{post,i} \text{ in Section 5.3 except}$$

it is calculated for each selected gene *k* and for both treatment arms $t = 1$ and $t = 0$. A high positive value of $DLLR_{ik}$ or of $\log(OR_{post,ik})$ is indicative of a positive response in the treatment arm compared to the control arm for subject *i*. In a weighted voting method, $DLLR_{ik}$ or $\log(OR_{post,ik})$ is calculated for each gene *k* and for each subject *i* in the validation set, and if the number of genes with $DLLR_{ik} > R$ is at least H, that subject would be classified as sensitive to treatment. The same nested CV approach used in ASD could then be employed in this method. Since the classification distance for Random Forests can also be expressed as a proportion (of votes), a Random Forests OR can also be employed in the same manner for treatment subset prediction. However only tuning parameter R may be needed for Random Forests, since one OR is calculated for each subject, and since the Random Forests embedded DR method may eliminate need for η.

K Nearest Neighbors and Treatment Subset Prediction

The distance employed in KNN is calculated over all the selected features for a subject. This eliminates need for the tuning parameter H in ASD. However a tuning parameter is still needed for K, η, and R. To make a classification decision, KNN uses the proportion of nearest neighbors in the response class *g*=1, say $\bar{p}_i$ as a distance. If the proportion is greater than 0.5, the subject is assigned to the response class, and to the non-response class otherwise. This can be expressed as $\hat{g}_i = I(\bar{p}_i > 0.5)$. A simple method for treatment subset prediction is to compare this proportion across the two treatment arms. In this method, the *K* nearest neighbors are calculated for both the treatment arm and the control arm. A greater difference in this proportion over the two treatment arms $(\bar{p}_{i1} - \bar{p}_{i0})$ is then indicative of patient sensitivity to treatment.

An alternative approach that makes direct use of the distance measure used to select the nearest neighbors is as follows:

1. The *K* nearest neighbors for a subject in the validation set are found for both responders and non-responders over both training set treatment arms. The total nearest neighbors for a subject is $4K$.
2. The average distance, say $\bar{d}_{gti}$, is calculated for both the *K* nearest responders and the *K* nearest non-responders in each treatment arm.
3. The average distance of the responders is then subtracted from the average distance of the non-responders for both treatment arms: $\left(\bar{d}_{01i} - \bar{d}_{11i}\right) - \left(\bar{d}_{00i} - \bar{d}_{10i}\right)$.
4. If this distance exceeds tuning parameter R, the subject is predicted to be sensitive to treatment.

Numeric Example

After appropriate DR and standardization for gene expression, it is found, using nested CV on the training set, that the optimal values for *K* and R are 2 and 2.2. Then for subject i in the validation set, the average distance of the two closest responders on the training set treatment arm is 2.3, using Euclidian distance. The average distance of the two closest non-responders is 4.4. On the training set control arm, the average distance for the two closest responders is 3.5, and for non-responders it is 3.3. Therefore $\left(\bar{d}_{01i} - \bar{d}_{11i}\right) - \left(\bar{d}_{00i} - \bar{d}_{10i}\right) = 2.1 - (-0.2) = 2.3$. Since $2.3 > R$, subject i is predicted to be sensitive to treatment.

Treatment Subset Prediction Using Support Vector Machines and Boosting

The numeric example given for SVM in Section 5.4 is easily extendable to treatment subset prediction. The distance d_{ti} is calculated in the same manner as d_i in Section 5.4, but with $\boldsymbol{w}$ and *b* calculated separately for each treatment arms in the training set. The difference in this distance between treatment arms is then:

$d_{1i} - d_{0i} = \left(d_i \mid t = 1\right) - \left(d_i \mid t = 0\right) = \left(\boldsymbol{w}_1{}'\boldsymbol{z}_i + b_1\right) - \left(\boldsymbol{w}_0{}'\boldsymbol{z}_i + b_0\right)$. A tuning parameter similar to R can be selected using nested CV or bootstrapping on the training set. If this difference in

distance exceeds R, subject i in the validation set is then predicted to be sensitive to treatment. The same approach can be used for Boosting, using the Boosting distance. Note that if an embedded DR method is used, then only tuning parameter R is needed for these methods. An additional advantage for Boosting (and Random Forests) is that these methods easily handle mixed covariates, meaning that clinical as well as genomic data can be easily incorporated.

8. Challenges and Future Directions

The no-free-lunch theorem states that no one classification method is optimal under all situations. Therefore it is beneficial to have a range of different methods for treatment subset prediction. Treatment interactions have been incorporated into each of the nonparametric methods reviewed above, but since the classification distance used for each method is unique, the type of treatment interaction is different for each method, as is the type of DR required. Many aspects of ASD require further research. For example, instead of a weighted voting method, an average score could be calculated over all the selected genes. This would eliminate need for tuning parameter H. Also sensitivity status is a latent variable since some sensitive patients do not respond to treatment, while some nonsensitive patients do respond to treatment. This results in a mixture of normal distributions for each class-specific and treatment arm-specific status. An EM algorithm (Dempster, Laird, and Rubin, 1977 [34]) developed by Bishop (2006 [10]) addresses this situation. He also developed a Bayesian method which addresses some of the shortcomings of the maximum likelihood method for mixtures. Finally, many works also highlight the importance of pathways in heterogeneity. It would be beneficial then to more directly incorporate pathway analysis into ASD methods. Methods such as these warrant further research using real data sets and simulation studies to elucidate their performance under various scenarios.

PARAMETRIC CLASSIFICATION METHODS

1. Some Historical Developments of Parametric Classification Methods

The roots of discriminant analysis date back to before 1936, when Mahalanobis (1936 [111]) introduced the Mahalanobis distance, and Fisher outlined a method which used this same distance to discriminate between species of iris (1936 [48]). This became known as Fisher's Discriminant Analysis (FDA). Welch (1939 [171]) used the likelihood ratio (LR) to show that FDA was equivalent to Linear Discriminant Analysis (LDA) when maximum likelihood (ML) estimates are used in place of true parameters. LDA assumes a multivariate normal distribution for the covariates. Wald (1944 [167]) incorporated prior population probabilities and misclassification costs.

Other statisticians who played important roles in the early development of discriminant analysis include Pearson, who proposed a mixture of normal distributions for clustering (1894 [123]), Neyman & Pearson (1933 [117]), who introduced the likelihood ratio test (LRT), and Rao (1947 [130]). Cox (1966 [31]) played an early role in the development of Logistic Classification Analysis or LCA (Anderson, 1972 [2]). Quenouille (1949 [128]) introduced the jackknife which reduces bias by successively leaving out one observation at a time, and Hills (1966 [78]) and Lachenbruch & Mickey (1968 [97]) introduced a method, closely related to the jackknife, of estimating error rates by omitting one observation from the computation of the discriminant function and using that observation to estimate error. This is the leave one out cross validation (LOOCV) method, which is one example of the more general method of cross validation (CV) which divides data into non-overlapping training and validation sets to build a model and estimate error. Lachenbruch and Mickey also pointed out the bias and optimism of the resubstitution method, which uses the same sample both to build a model and to estimate error.

These early methods were developed at a time when high throughput data (HTD) was not as prevalent. HTD typically have tens of thousands or more of features, together with sample sizes on the order of 100 or less. This situation is often referred to as $p >> n$ where p refers to the features and n refers to the sample size. The early parametric methods mentioned above are not HTD-capable without some outside help. Many dimension reduction (DR) methods have been developed for HTD to address this situation. More recent nonparametric machine learning methods such as Random Forests and Boosting were developed specifically for HTD and have built-in DR techniques. At the same time, the earlier developed parametric classification methods, when used in conjunction with suitable DR techniques, often compare favorably to these more recently developed machine learning methods. For example a modified version of LDA, Diagonal Linear Discriminant Analysis (DLDA), which ignores correlation between features, outperformed Boosting and Random Forests in a comparison of methods by Dudoit et al. (2002 [39]) involving genomic data. Further work by Bickel & Levina (2004 [9]) and Fan & Fan (2008 [46]) showed reasons why DLDA often outperforms LDA in an HTD setting.

L1 Regularization or L1R (Tibshirani, 1996 [157]) selects at most n features and shrinks coefficients of selected features towards 0. Thus it provides a way to naturally integrate the DR and classification steps. In contrast, L2 Regularization or L2R (Hoerl & Kennard, 1970 [80]) shrinks parameter estimates towards 0 or towards a common value, without eliminating any of them. L2R methods such as those introduced by Ledoit & Wolfe (2004 [99]) improve stability of covariance matrices. Developments such as these further enabled use of parametric classification methods in an HTD setting. Applications involving classification include Penalized Discriminant Analysis (Hastie, Buja, & Tibshirani, 1995 [74]), NSC (Tibshirani et al., 2003 [161]), Regularized Discriminant Analysis (Friedman, 1989 [58]), MLDA (Modified LDA, Xu, Brock, & Parrish, 2009 [177]) and penalized logistic regression for gene interactions (Park & Hastie, 2008 [120]).

Treatment Subset Prediction

In phase III clinical trials, it has increasingly been recognized that treatment agents only benefit a subset of patients enrolled in these studies. In cases where predictive assays cannot be developed before a phase III clinical trial, Freidlin & Simon (2005 [54]) proposed the Adaptive Signature Design (ASD) which predicts a subset of patients more sensitive to treatment based on gene expression data. Their specific ASD model used weighted voting and single gene logistic models which include gene-treatment interaction. The tuning parameter set consists of a parameter η to select predictive genes from the training set, a parameter R to determine the threshold for the magnitude of the treatment arm odds ratio (OR) for a selected gene and patient in the validation set, and a parameter H to determine the threshold for the number of selected genes for a patient in the validation set having a treatment arm OR which exceed *R*. Subjects which have at least *H* selected genes meeting criteria for *R* are predicted to be sensitive to treatment. Type 1 error is partitioned between a test for overall treatment effect involving all the subjects and a test for treatment effect involving only the stage II patients predicted to be sensitive to treatment. Figure 5 gives more information.

Other recent works in this area include Zhang, Tsiatis, Laber, & Davidian (2013 [179]), and Zhao, Tian, Cai, Claggett, & Wei (2013 [181]).

Purpose of Chapter

The purpose of this study is first to review and examine parametric classification methods, focusing on specific distance measures used to make a classification decision. These are specific to the classification method and are what make the method unique. Then extensions of these methods to treatment subset prediction can often be made by comparing these distances across treatment arms. Since no one classification method is optimal in all situations, it is advantageous to have available different methods for treatment subset prediction.

Organization of Study

Section 2 covers notation and definitions. Section 3 is a brief introduction to parametric DR methods, with a focus on treatment subset prediction. Parametric classification models and rules are outlined in Section 4. Discussion and conclusions are in Sections 5, and challenges and future work are in Section 6.

2. Notation and Definitions

Notational convention follows that of Part 1 of this review. Subscripts *D* and *V* denote training and validation sets respectively, and n_D and n_V denote their sample sizes; y_i denotes a categorical or a continuous response for the i^{th} subject. For two classes, $g_i = 1$ denotes the disease or relapsed class, and $g_i = 0$ or -1 denotes the other class; for more than two classes, $g_i \in \{0,1,...,G-1\}$; covariate vectors $\boldsymbol{x}_i = \{x_{i1}, x_{i2}, ..., x_{im}\}$ and $\boldsymbol{z}_i = \{z_{i1}, z_{i2}, ..., z_{i\ell}\}$, where $\ell = p - m$, denote respectively the clinical and high-dimensional covariate vectors for the i^{th} subject. The covariate vector for the i^{th} subject is then $\{\boldsymbol{x}_i, \boldsymbol{z}_i\}$. In the training set $\hat{\mu}_k$ denotes the average of the k^{th} feature, $\hat{\mu}_{gk}$ denotes the average of the k^{th} feature over subjects that belong to the specific class g, and $\hat{\mu}_{gtk}$ denotes the average over the subset of subjects in class g that have also been assigned to treatment arm t, where $t \in \{0,1\}$.

The term feature refers either to covariates or functions of covariates, and the term interaction refers to a statistical interaction [30]. More detail can be found in Part 1 of this review. Table 5 provides a list of acronyms and definitions used in this chapter.

Table 5: List of acronyms, definitions, and sections where acronyms are used.

Acronym	Associated Words	Description	Sections
ASD	Adaptive Signature Design	Method to predict subset of patients responding differently to treatment; see Figure 5	1,3-6
CS	Compound Symmetry	See Table 7	
CV	Cross-Validation		1,3-6
DLDA	Diagonal Linear Discriminant Analysis	See Table 7	4
DLLR	Difference in LLR		5
DR	Dimension Reduction		3-4
ECM	Expected Cost of Misclassification		4
FE	Feature Extraction	Class of DR methods	3
FDA	Fisher's Discriminant Analysis	See Table 7	4
FS	Feature Selection	Class of DR methods	3
HTD	High Throughput Data		1,4
LASSO	Least Absolute Shrinkage and Selection Operator	Same as L1R- see Table 6	1,3
L1R	L1 Regularization	Same as LASSO - see Table 6	1,3
L2R	L2 Regularization	Same as Ridge Regression - See Table 6	1,3
LCA	Logistic Classification Analysis	See Table 7	1,4,5
LDA	Linear Discriminant Analysis	See Table 7	1,4
LOOCV	Leave-One-Out Cross-Validation		1
LR	Likelihood Ratio		1,4
LRT	Likelihood Ratio Test		1,4,5
LLR	Log Likelihood Ratio		4,5
MAD	Median Absolute Deviation	Robust regression method	4
ML	Maximum Likelihood		1,4,5
MLDA	Modified Linear Discriminant Analysis	See Table 7	4
mRNA	Messenger RNA		6
miRNA	microRNA		6
NSC	Nearest Shrunken Centroids	HTD classification method	5.2
OR	Odds Ratio		1,4-5
PC	Principal Components		3
PCA	Principal Components Analysis	See Table 6	3
PDA	Penalized Discriminant Analysis	See Table 7	4
PLS	Partial Least Squares	See Table 6	3
QLDA	Quadratic Linear Discriminant Analysis	See Table 7	4
QDA	Quadratic Discriminant Analysis	See Table 7	4
RDA	Regularized Discriminant Analysis	See Table 7	4
SCRDA	Shrunken Centroids RDA		4
SNP	Single Nucleotide Polymorphism		6
SPCA	Sparse PCA	See Table 7	3
SPLS	Sparse PLS	See Table 7	3

3. A Brief Introduction to Parametric Dimension Reduction with a Focus on Treatment Subset Prediction

The application of DR methods in an HTD setting is an area still undergoing rapid development. As a recent title of a review of DR methods by Fan & Lv (2010 [47]) suggests, selectivity is needed for a review of DR itself.

Univariable Feature Selection for Large Scale Dimension Reduction

Fan & Fan (2008 [46]) demonstrated that under certain conditions, a t test for each HTD feature comparing differential expression between the non-responsive and responsive group can be used to reduce dimensionality from ultrahigh to a moderate scale below or on the order of the sample size n. This is a univariable Feature Selection (FS) method. The features could be ranked and selected by p-value, false discovery rate (Benjamini & Hochberg, 1995 [8]), or the absolute value of a t-statistic. The empirical Bayes t statistic (Smyth, 2004 [150]) performed well across a range of sample sizes (Jeffery, Higgins, & Culhane, 2006 [91]). Methods such as cross-validation can be used to select a cut-off value to determine features for use in the prediction rule.

However, as stated in Chapter 1, a test comparing gene expression means for the two classification groups may not prove useful in preserving features containing information regarding treatment subset prediction. A more appropriate method used in simulations in Freidlin et al. (2010 [55]) compares differential gene expression between enhanced and standard treatment in the responsive group to differential expression in the non-responsive group. The equation is in Chapter 1.

Dimension Reduction Impacts Classification Performance

In the Dudoit et al. (2002 [39]) study, DLDA performed well while FDA performed poorly when the same number of selected features were used for each method. However FDA had error rates comparable to DLDA when fewer features were selected in the DR step. The explanation was that estimates of the pooled within-class covariance matrix for FDA become unstable when the number of selected features is large. Since DLDA does not estimate the non-diagonal elements, it eliminates variance associated with these parameter estimates in exchange for some increase in bias.

Optimization of DR for classification depends on the classification method. The weighted voting method used in Freidlin & Simon's ASD model addresses this by including the DR tuning parameter as one in a set of inter-related tuning parameters. In this way DR is embedded in the

classification process. The set of tuning parameters chosen is the one which (using CV on the training set only) minimizes the p-value in a treatment arm comparison of predicted sensitive patients.

In selecting the list of plausible tuning parameters for this ASD implementation, it is important to take into account factors that influence optimal cutpoint selection. These include the sample size, the number of features, and the classification method. For example, if the classification method is LDA, the number of selected features p^* must be $n_D - 2$ or less (for two classes). Moreover, it has been reported that LDA is unstable if n_D is not at least 5 or 10 times p^* (Jain & Chandrasekaran, 1982 [90]).

Methods used after Initial Large Scale Feature Selection

Though univariable FS is often preferred for an initial large scale screening, one drawback is that it does not take into account correlation between features. Therefore it is possible that features selected in this manner will be highly correlated.

Principal Components Analysis

One solution is to use a univariable FS method first to reduce features to some number p^*, say less than or on the order of n_D. Then principal components analysis (PCA) can be used to extract a smaller number of uncorrelated features for use in treatment subset prediction in the validation set. The method of Fisherfaces (Belhumeur, Hespanha, & Kriegman 1997 [7]) used a similar approach.

PCA was originally proposed by Pearson over 100 years ago (1901 [124]) and further developed by Hotelling (1933 [83]). The principal components of p^* features are the p^* orthogonal directions which maximize variance between the features. The first principal component is the linear combination which maximizes the variance between the features, and the second PC is the linear combination orthogonal to the first PC which maximizes the remaining

variance, and so on. A number of principal components much smaller than m can usually be found which explains most of the variance of the original features. However as pointed out by Bickel & Li (2006 [5]), PCA breaks down when $p > n$: "…not only does the empirical covariance matrix become singular for $p > n$, but , …, if $\frac{p}{n} \to c,\ 0 < c < \infty,$ the empirical eigenvectors and eigenvalues are grossly inconsistent in terms of estimating the corresponding population quantities." Therefore PCA may not be an appropriate choice for an initial screening when $p > n$. To address this issue, penalized methods such as sparse principal components (Zou, Hastie, & Tibshirani, 2006 [185]) have been developed.

One difference between PCA and other DR methods is that PCA does not make use of association between explanatory and response variables. The next method is one way of addressing this issue.

Partial Least Squares

Partial least squares (PLS) "seeks directions that have high variance and have high correlation with the response" (Hastie, Tibshirani, & Friedman, 2009 [75]). It was proposed by Wold in the 1960's as an application for chemometrics for regression and DR (Geladi & Kowalski, 1986 [63]). Boulesteix (2007 [13]) compared applications of PLS in HTD. Nguyen & Rocke (2002 [118]) applied PLS to cancer classification. There are similar shortcomings in $p >> n$ problems for PLS as for PCA, and penalized methods such as Sparse PLS (Chun & Keleş, 2010 [27]) have been developed to help address this issue.

L1 and L2 Regularization Methods

There are other methods that could be used in this second step as well. For example L1R methods constrain the sum of the absolute value of the coefficients to be a constant. This results in shrinkage of regression coefficients and selection of at most n features depending on the magnitude of the penalty term(s). This enables some L1R methods to be used in a first step DR as well. Methods using L2R constrain the sum of the squares of the coefficients to be a constant

value. This has the effect of shrinking coefficients towards a constant value without setting them equal to that constant. In so doing degradation due to correlation is reduced. One example is the L2R logistic regression for gene-gene and gene-environment interactions introduced by Park & Hastie (2008 [120]). This method allows categorical factors such as treatment to be included in interactions. Higher order gene-gene or gene-treatment interactions can be included. The Adaptive LASSO (Zou, 2006 [184]) allows different penalty terms for each covariate. The elastic net (Zou & Hastie, 2005 [183]) is a mixture of L1R and L2R and encourages strongly correlated features to enter or leave the model together. The Group LASSO enables penalty terms to be selected differently for different features, or for selected groups of features (Yuan & Lin, 2006 [178]). Selection of groups makes it possible to select, for example, genes involved in a biological pathway. Jacob, Obozinski, & Vert (2009 [89]) extend Group LASSO to allow for overlap, as might happen with genes in more than one biological pathways. Table 6 provides more information on DR methods including software implementations.

Table 6: Parametric DR methods and software implementations.

Method	Description	R Packages	SAS Software
PCA	Principal components analysis – extracts a smaller number of uncorrelated features from a larger set of possibly highly correlated ones. Breaks down in p>>n situations.	pls	Proc princomp, proc factor
SPCA	Modification to PCA which enables use in p>>n data.	pcaPP, PMA, elasticnet	*
PLS	Partial least squares – like PCA, but also takes into account association between features and response variable (and therefore can also be used as classification method). Breaks down in p>>n situations.	pls, caret	Proc pls
SPLS	Modification to PLS that enables use in p>>n situations.	spls	*
L1R (LASSO)	Constrains sum of absolute value of coefficients to be a constant depending on penalty term. Shrinks parameter estimates towards 0, allowing some to be equal to 0; selects at most n features out of p.	penalized, glmpath, LiblineaR, elasticnet, lassoshooting	*
Adaptive LASSO	Allows different penalty terms for each feature.	lqa, parcor, lars	*
Group LASSO	Allows penalty terms to be different for each feature or for selected groups of features. Has been extended to account for overlap (Jacob, Obozinski and Vert, 2009).	grplasso, grpreg, SGL, standGL, gglasso	
L2R (Ridge Regression)	Constrains sum of the squares of parameter estimates to be a constant, depending on penalty term. Shrinks parameter estimates towards a common value or towards 0 without eliminating any of them. Can be used to stabilize covariance matrices. Shrinks coefficients of correlated predictors towards a common value. Ideal if there are many predictors all with non-zero coefficients.	LiblineaR, glmnet, lrm, penalized	*
Combined L1R and L2R	Constrains parameter estimates to be a sum of both the L1 and L2 constraints; also see Park and Hastie 2008.	elasticnet, glmnet	*

*For SAS code on penalized/regularized methods, one can search sites such as http://www.sas-programming.com/2010/09/regularized-discriminant-analysis.html and http://sasdiehard.blogspot.com/2011/03/fitting-logistic-regression-in-data.html for guidance.

4. Parametric Classification Models and Rules

Fisher's Discriminant Analysis

Fisher's idea was to find the compound distance which is the linear combination of covariates, say $z_i^* = \lambda_1 z_{i1} + ,..., + \lambda_{p^*} z_{ip^*}$ (or $z_i^* = \boldsymbol{\lambda}' \boldsymbol{z}_i$, where $\boldsymbol{\lambda}$ and $\boldsymbol{z}_i$ are the corresponding p^* - length vectors), which maximizes the between-class distance with respect to the within-class variance. The between-class distance is

$$\hat{\mu}_d^* = \hat{\mu}_1^* - \hat{\mu}_0^* = \boldsymbol{\lambda}' \left(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0 \right),$$

where $\hat{\boldsymbol{\mu}}_g \in \{\hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1\}$ denotes the p^*- length vectors of class specific feature means estimated from the training set; and where p^* is the number of selected features. Note that $\hat{\mu}_1^*$, $\hat{\mu}_0^*$, and $\hat{\mu}_d^*$ are scalars. The vector which maximizes the compound distance between classes $\hat{\mu}_d^*$ with respect to its pooled within-class variance can be found by differentiating and setting equal to 0 the quantity $\hat{\mu}_d^{*2} / s^*$ with respect to $\boldsymbol{\lambda}$, where the scalar $s^* = \boldsymbol{\lambda}' \boldsymbol{SS} \boldsymbol{\lambda}$, and $\boldsymbol{SS}$ is the pooled within-class sum of squares matrix for the original covariates. This leads to a system of p^* linear equations, where $p^* \le n_D - 2$ is the number of selected features. The solution is:

$$\boldsymbol{\lambda} = \boldsymbol{SS}^{-1} \left(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0 \right).$$

In the case of equal class sample sizes, a classification decision can be based on which class the compound distance for the subject is closest to:

$\hat{g}_i = \arg\min_g \left[\left(\dfrac{z_i^* - \hat{\mu}_g^*}{\hat{\sigma}^*} \right)^2 \right]$, where $\hat{\mu}_g^* \in \{\hat{\mu}_0^*, \hat{\mu}_1^*\}$ are class specific means and $\hat{\sigma}^*$ is the pooled within-class standard deviation of the compound measures estimated on the training set. Figure 6 gives a graphical depiction using the square-root of the squared distance above.

Alternatively, a cut-off point can be determined using CV. For example,

$$\hat{g}_i = I\left[\left\{\left(\frac{z_i^* - \hat{\mu}_0^*}{\hat{\sigma}^*}\right)^2 - \left(\frac{z_i^* - \hat{\mu}_1^*}{\hat{\sigma}^*}\right)^2\right\} > C\right].$$

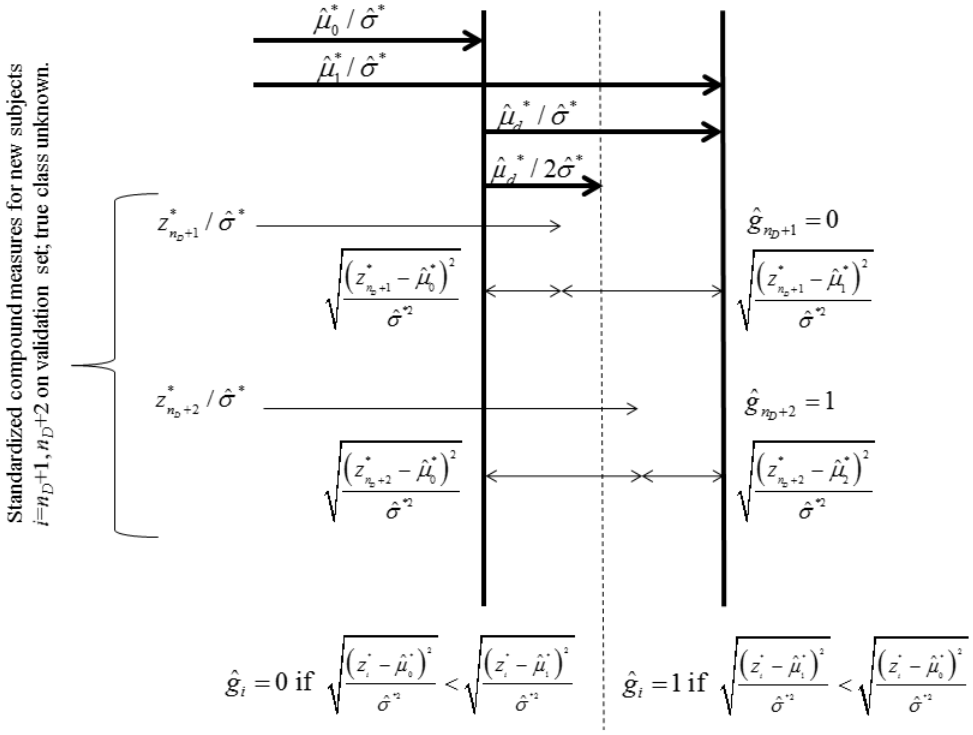

Figure 6: Use of compound measure in FDA to allocate new subjects; $\hat{\mu}_0^*$ and $\hat{\mu}_1^*$ are the average of the compound measures over all the subjects for class *g*=0 and *g*=1 on the training set; $\hat{\sigma}^{*2}$ is the within-class pooled variance of the compound measure, also calculated from the training set; $\hat{\mu}_0^* / \hat{\sigma}^*$ and $\hat{\mu}_1^* / \hat{\sigma}^*$ are the standardized class compound measures, and $\hat{\mu}_d^* / \hat{\sigma}^* = \hat{\mu}_1^* / \hat{\sigma}^* - \hat{\mu}_0^* / \hat{\sigma}^*$ is the difference between these two class compound measures. The standardized compound measure for each subject *i* in the validation set is $z_i^* / \hat{\sigma}^*$. Predicted class $\hat{g}_i$ is determined by which standardized class compound measure, $\hat{\mu}_g^* / \hat{\sigma}^*$, the standardized compound measure $z_i^* / \hat{\sigma}^*$ is closest to in terms of the standardized compound distance, $\frac{(z_i^* - \hat{\mu}_g^*)^2}{\hat{\sigma}^{*2}}$, or,equivalently, to its square root. The cut-off for class membership is shown by the dashed line half way in between the two classes (assuming equal sample sizes for classes). In this case, if $\sqrt{\frac{(z_i^* - \hat{\mu}_0^*)^2}{\hat{\sigma}^{*2}}} < \frac{\hat{\mu}_d}{2\hat{\sigma}^*}$, then $\hat{g}_i = 0$. Otherwise $\hat{g}_i = 1$.

Since no distributional assumptions are used in the derivation of the compound measure or distance, authors such as Rencher (1991 [131]) classified FDA as nonparametric or distribution-free. Moreover Ripley (1996 [134]) pointed out that "a t-distribution with a moderate number of degrees of freedom is often regarded as a better fit" than a normal distribution. This robust method is included in the R functions lda in package MASS (Venables & Ripley, 2002 [165]).

The Likelihood Ratio and Classification

Let $\boldsymbol{Z}$ denote the *n* by *p* gene expression profile matrix (the matrix over all features, and over both the training and the validation set) , and let $\boldsymbol{Z}_g$ denote the matrix of rows of $\boldsymbol{Z}$ consisting of $\forall i \in 1,...,n$ such that $g_i = g.$ Suppose $\boldsymbol{Z}_g$ follows a multivariate distribution with parameter vector $\boldsymbol{\theta} = \boldsymbol{\theta}_g.$ Each subject belongs to one and only one of *G* classes, but class membership for subjects in the validation set is not known. In the special case where $G = 2$, then an LRT can be used to assign subject i to one of the two classes. The hypotheses are:

$$H_1 : g = 1$$
$$H_0 : g = 0.$$

The *LR* for the i^{th} subject is then equal to the ratio of the product of the prior class probabilities π_g and class-dependent densities $f(\boldsymbol{z}_i \mid \boldsymbol{\theta} = \boldsymbol{\theta}_g)$ where here $\boldsymbol{z}_i$ is the *p*-length gene expression profile vector for subject *i*:

$$LR_i = \frac{\Pr(g = 1, \boldsymbol{z}_i)}{\Pr(g = 0, \boldsymbol{z}_i)} = \frac{\Pr(g = 1)\Pr(\boldsymbol{z}_i \mid g = 1)}{\Pr(g = 0)\Pr(\boldsymbol{z}_i \mid g = 0)} = \frac{\pi_1 f(\boldsymbol{z}_i \mid \boldsymbol{\theta} = \boldsymbol{\theta}_1)}{\pi_0 f(\boldsymbol{z}_i \mid \boldsymbol{\theta} = \boldsymbol{\theta}_0)}, \text{ or } LLR_i = \log(LR_i).$$

Note that this rule assumes knowledge of the true underlying class distributions. In this unusual scenario, no model building on a training set is necessary. The classification rule can then be expressed as $\hat{g}_i = I(LLR_i > 0),$ where $g \in \{0,1\}$. Now if $\boldsymbol{Z}$ follows a multivariate normal distribution, then, assuming a common covariance matrix $\boldsymbol{\Sigma}$ for the two classes:

$$LR_i = \left(\frac{\pi_1}{\pi_0}\right)\frac{f(\boldsymbol{z}_i \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma})}{f(\boldsymbol{z}_i \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma})} = \left(\frac{\pi_1}{\pi_0}\right)\left\{\frac{\exp\left[-\frac{1}{2}(\boldsymbol{z}_i - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}_i - \boldsymbol{\mu}_1)\right]}{\exp\left[-\frac{1}{2}(\boldsymbol{z}_i - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}_i - \boldsymbol{\mu}_0)\right]}\right\},$$

where $\boldsymbol{\mu}_g$ is the *p*-length vector of true means for the columns of $\boldsymbol{Z}_g$. Extension to $\{\boldsymbol{x}_i, \boldsymbol{z}_i\}$ is straightforward.

The log likelihood ratio is:

$$LLR_i = \log\left(\frac{\pi_1}{\pi_0}\right) + \left\{-\frac{1}{2}(\boldsymbol{z}_i - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}_i - \boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{z}_i - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}_i - \boldsymbol{\mu}_0)\right\}.$$

Note that $\sqrt{(\boldsymbol{z}_i - \boldsymbol{\mu}_g)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}_i - \boldsymbol{\mu}_g)}$ is the Mahalanobis distance.

Now, if the true parameter values are not known, then it is necessary to estimate them from a training set where class membership is known. Substituting ML estimates for true parameter values:

$$LLR_i = \log\left(\frac{\hat{\pi}_1}{\hat{\pi}_0}\right) + \left\{-\frac{1}{2}(\boldsymbol{z}_i - \hat{\boldsymbol{\mu}}_1)'\boldsymbol{S}^{-1}(\boldsymbol{z}_i - \hat{\boldsymbol{\mu}}_1) + \frac{1}{2}(\boldsymbol{z}_i - \hat{\boldsymbol{\mu}}_0)'\boldsymbol{S}^{-1}(\boldsymbol{z}_i - \hat{\boldsymbol{\mu}}_0)\right\},$$

where $\boldsymbol{S}$ is the pooled within-class sample covariance matrix from the training set, and $\hat{\boldsymbol{\mu}}_g$ is the *p**- length vector of ML estimates from the training set for $\boldsymbol{\mu}_g$. Here LLR_i for subject *i* in the validation set is used as the classification distance in place of LLR_i since true parameter estimates are not known. Note that when the true parameter values are known, there are no restrictions on the number of features that can be used. However the true values are almost always not known. In this case the number of selected features *p** must be less than or equal to $n_D - 2$. The classification rule is then: $\hat{g}_i = I(LLR_i > C)$, where $g \in \{0,1\}$. Note that a cut-off *C* may be used in place of 0 since there is no longer complete knowledge of the true underlying distributions of $\boldsymbol{Z}_g$. The cut-off may be chosen using CV.

Numeric Example

Take $\hat{\boldsymbol{\mu}}_0' = 0,1$, $\hat{\boldsymbol{\mu}}_1' = 1,2$, $\boldsymbol{S} = \begin{matrix} 2 & 1 \\ 1 & 2 \end{matrix}$, equal prior class probabilities, and $\boldsymbol{z}_i' = 2,2$. Then

$LLR_i = \log(1) + \{-(1/2)(2/3) + (1/2)(2)\} = 0 + 2/3 = 2/3.$ Then, if

$C = 0, \hat{g}_i = I(2/3 > 0) = 1.$

Relationship Between Linear Discriminant Analysis and Fisher's Discriminant Analysis

The distance $\sqrt{(\boldsymbol{Z}_i - \hat{\boldsymbol{\mu}}_0)' \boldsymbol{S}^{-1} (\boldsymbol{Z}_i - \hat{\boldsymbol{\mu}}_0)}$ is a standardized version of Fisher's compound distance DC_i and therefore LDA results in the same classification rule as FDA when ML estimates are substituted for true values.

Diagonal Linear Discriminant Analysis

For DLDA, and substituting ML estimates for true parameter values, the equation further reduces to:

$$LLR_i = \log\left(\frac{\hat{\pi}_1}{\hat{\pi}_0}\right) + \left[-\frac{1}{2}\sum_{k=1}^{p^*} (z_{ik} - \hat{\mu}_{1k})^2 / \hat{\sigma}_k^{\,2} + \frac{1}{2}\sum_{k=1}^{p^*} (z_{ik} - \hat{\mu}_{0k})^2 / \hat{\sigma}_k^{\,2} \right],$$

where $\hat{\sigma}_k^{\,2}$ is the estimate (from the training set) for the pooled within-class variance for the k^{th} selected gene, and $\hat{\mu}_{gk}$ is the k^{th} element of vector $\hat{\boldsymbol{\mu}}_g$. For DLDA, it is no longer necessary that p^* be less than or equal to $n_D - 2$, since means and variances can be estimated separately for each gene; however performance will still degrade unless p^* is much less than n_D.

The classification rule is the same as before – i.e. $\hat{g}_i = I(LLR_i > C),$ where $g \in \{0,1\}$. For a weighted voting method (Breiman, 1996 [16]), the summation sign would be eliminated and separate models would be used for each selected feature. A cut-off can be chosen for the LLR_{ik}'s using nested CV, and the subject assigned to class $g = 1$ if the number of genes having an LLR_{ik} exceeding this tuning parameter, say *R*, is equal to or greater than the value of another

tuning parameter, say H, also chosen using nested CV. This can be expressed as

$$\hat{g}_i = I\left[\left(\sum_{k=1}^{p^*} I(LLR_{ik} > R\right) \geq H\right].$$

Likelihood Ratio Test and Misclassification Costs

Different misclassification costs for each class $c(1-g,g) = c(\hat{g} = 1-g, g),\ g \in \{0,1\}$ can also be included in the LRT. In this case the classification rule is

$\hat{g}_i = I\left(\log \frac{\pi_1 c(0,1) f(\boldsymbol{z}_i \mid \boldsymbol{\theta} = \boldsymbol{\theta}_1)}{\pi_0 c(1,0) f(\boldsymbol{z}_i \mid \boldsymbol{\theta} = \boldsymbol{\theta}_0)} > 0\right)$. This is the expected cost of misclassification (ECM) rule (Johnson & Wichern, 2007 [92]). True parameters can be replaced with their maximum likelihood estimates as before. Note that, though a method such as CV could still be used here to assign subjects, care is needed to assign appropriate weights to account for misclassification costs and prior population probabilities. The distance used to make a classification decision in this case is then $\log\left(\frac{\hat{\pi}_1 c(0,1) f(\boldsymbol{z}_i \mid \boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_1)}{\hat{\pi}_0 c(1,0) f(\boldsymbol{z}_i \mid \boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_0)}\right)$. For simplicity, this distance will also be referred to as an LLR in this work.

Numeric Example

Take previous values for $\hat{\boldsymbol{\mu}}_g$ and $\boldsymbol{z}_i$, and take $\boldsymbol{\Sigma} = \boldsymbol{S}$ in the previous example: with $\frac{c(0,1)}{c(1,0)} = 2,$ and $\frac{\hat{\pi}_1}{\hat{\pi}_0}$=1, then $LLR_i = \log(1*2*2/3) = 0.288,$ and, if C is chosen as 0, $\hat{g}_i = I(0.288 > 0) = 1.$

A Note on Diagonal Linear Discriminant Analysis

The naïve Bayes DLDA, which compared favorably to methods such as Boosting and Random Forests in Dudoit et al.'s 2002 study [39], sets the non-diagonal elements of the within-class covariance matrix to 0, and therefore ignores correlation between features. Naïve Bayes methods have been observed to work well in many situations despite the fact that the assumption

is not generally true (Bickel & Levina, 2004 [9], Hastie et al., 2009 [75]). There is a trade-off between bias reduction that results from estimating more covariance terms, and the increased variance resulting from these estimates. In many situations the increased variance outweighs the benefits of bias reduction.

Diagonal Linear Discriminant Analysis and Nearest Shrunken Centroids

The NSC method is a penalized form of DLDA proposed by Tibshirani, Hastie, Narasimhan, & Chu (2002 [160]). It includes use of a tuning parameter which has the effect of shrinking distance of genes from class centroid $(\hat{\mu}_{gk})$ to overall centroid $(\hat{\mu}_{k})$. Genes which have a shrunken distance of zero for all classes are effectively filtered out of the classifier. In this way DR is naturally integrated into the classification process. The tuning parameter is selected based on nested CV. Increasing the tuning parameter results in fewer selected genes. It was found that both training error and prediction error were minimized with a tuning parameter near 4.34. Guo, Hastie, & Tibshirani (2007 [68]) introduced a related method called shrunken centroids RDA (SCRDA).

Modified Linear Discriminant Analysis

MLDA (Ledoit & Wolfe, 2004 [99]) reduces bias and improves stability of the LDA covariance matrix by shrinking the sample eigenvalues towards a common mean. Table 7 provides information on parametric classification methods, including regularized methods and software implementations.

Table 7: Some parametric classification methods and software implementations.

Method	Description	R Software	SAS Software
LCA	Classification using logistic regression.	glm(base), lrm(rms), polr(MASS)	Proc logistic, Enterprise Miner
Penalized LCA	Classification using penalized logistic regression (L1, L2, or both).	logistf(logistf), penalized(penalized), glmnet, elasticnet, also see Park and Hastie, 2008	Proc logistic
FDA/LDA	Classification method maximizing between-class squared distance w.r.t. within-class squared distance. Assumes common within-class covariance matrix.	lda(MASS)	Proc discrim
QDA	Same as FDA, except allows within-class covariance matrix to be different for each class.	qda(MASS)	Proc discrim
DLDA	Uses diagonal within class covariance matrix (non-diagonal elements set equal to 0).	diagDA(sfsmisc) stat.diag.da(WGCNA)	Proc discrim
QLDA	Same as DLDA, but allows unequal diagonal elements for each class.	diagDA(sfsmisc)	Proc discrim
Linear Regression Classification	Formulates FDA as regression – gives same classification rules when using appropriate numeric values for classes.	Any package for linear regression	Any procedure for linear regression
RDA	Shrinks unequal within-class covariance matrices assumed in QDA towards equal covariance matrices assumed in FDA; also shrinks individual class covariance matrices towards identity matrix multiplied by average of eigenvalues; amount of shrinkage in each direction determined by cross validation.	rda(klaR), rda(rda)	Proc discrim with sample code
PDA	Formulates FDA as a regression problem. Then penalized methods for regression can be used.	mda(mda), pdmclass	
Penalized DLDA	Uses diagonal within class covariance matrix, and shrinks diagonal elements towards common value or towards 0.	penalizedLDA (penalizedLDA), pamr	
Penalized DQDA	DQDA with diagonal elements shrunk toward 0 smoothed.	sdqda(sparsediscrim)	
MLDA	Shrink sample eigenvalues towards common mean using closed form solution.	Ledoit & Wolfe 2004, Xu,Brock, & Parrish 2009	
CS	Non-diagonal elements of covariance matrix assumed equal.	Not aware of any implementation	
Weighted Voting	Prediction using single gene models.	votingLinearPredictor(WGCNA)	

Linear Regression and Classification

Bishop (2006 [10]) and Duda, Hart, & Stork (2001 [38]) showed that linear regression, with appropriately chosen cutoffs for the two classes, has the same classification rule as FDA and LDA. This was also implied in Fisher's 1936 work. Robust regression methods such as Median Absolute Deviation (MAD) or M-regression could also be used, though in these cases the classification rule will not be the same as for FDA or LDA using ML estimates.

If using the linear regression form of LDA, the class variable G is used as the response variable, and gene expression Z and other covariates are used as explanatory variables. The classification rule for subject i in the validation set is is

$\hat{g}_i = \text{sign}(\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + ... + \hat{\beta}_m x_{im} + ... + \hat{\gamma}_\ell z_{il})$, where $\hat{g}_i \in \{-1,1\}$. It follows that the linear regression distance used for classification is $\hat{y}_i$. Interaction terms can also be included. The coefficients are estimated from the training set, using response variable $y \in \{-1,1\}$ for class $g \in \{-1,1\}$, assuming equal sample sizes. The response variable can also be adjusted for unequal sample sizes. See Bishop (2006 [10]) or Fisher (1936 [48]) for details. Nested CV can also be used to determine an appropriate cut-off *C*; i.e. $\hat{g}_i = \text{sign}(\hat{y}_i - C)$.

Numeric Example

Using nested CV, the tuning parameter set selects three genes on the validation set. A multi-gene linear regression is used on the validation set. Sample size for the two classes are equal. Estimated parameters from the training set are $\hat{\beta}_0 = -0.4,\ \hat{\beta}_1 = -0.5,\ \hat{\beta}_2 = -0.4,$ and $\hat{\beta}_3 = 1.2.$ Gene expression for the three genes for subject i in the validation set are

$z_{i1} = 1.1,\ z_{i2} = 0.8,$ and $z_{i3} = 1.3.$ Taking $C = 0,\ \hat{g}_i = \text{sign}(-0.4 + ... + 1.2 * 1.3) = 1.$

Linear Regression Classification with a Continuous Response as Outcome Variable

In some settings a continuous variable can be used as the outcome variable. This is more natural for methods such as linear regression. This approach was used in Rai, Cambon, Pan, Gargett & Chaires (2013 [129]) in an application involving differential scanning calorimetry plasma thermogram analysis. Residuals were used as a distance measure. Two models were compared, one assuming $g_i = 0$ and one assuming $g_i = 1$ for each subject i in the validation set. Parameters for the models were estimated from the training set, and feature values x_{ik} were taken from subjects in the validation set. Subjects in the validation set were allocated to the class whose model had the smallest residual distance. In this case the residual distance was the average

number of residuals whose absolute value exceeded a specified quantile: i.e.:

$\hat{g}_i = \text{sign}\{\bar{p}_0 - \bar{p}_1\}$, where $\bar{p}_g$ is the average proportion of residuals for subject i with absolute value exceeding a specified quantile, under the model assuming $g_i = g$. The specified quantile can be estimated from the training set. The distance used for classification in this method is then $\bar{p}_0 - \bar{p}_1$.

Another advantage of regression methods for classification is the wealth of related tools that have been developed. Splines (Friedman, 1991 [59]), additive models (Hastie & Tibshirani, 1987 [73]), regularized methods and quantile regression (Koenker & Bassett, 1978 [95]) are some of many examples. Importantly for treatment subset prediction, one can naturally include treatment interactions in regression models to make a classification decision. The classification method outlined next, LCA, is a regression method and can also make use of these approaches.

Logistic Classification

Many works such as Press & Wilson (1978 [127]) have compared LCA and LDA. It has been found that their performance is surprisingly similar under many scenarios. LDA will perform better when the normality assumption is approximately true (or when suitable transformations can be applied which results result in this assumption being approximately true), and LCA will have superior performance when this assumption is severely violated (Hastie et al., 2009 [75]).

Logistic regression takes the form

$\text{logit}(y|\boldsymbol{x}_i, \boldsymbol{z}_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots \beta_m x_{im} + \gamma_1 z_{i1} + \ldots \gamma_\ell z_{i\ell}$, where

$y \in \{0,1\}$, and $\text{logit}(p_i) = \log\left(\frac{\Pr(y=1|\boldsymbol{x}_i, \boldsymbol{z}_i)}{1-\Pr(y=1|\boldsymbol{x}_i, \boldsymbol{z}_i)}\right)$. Logistic regression is in the class of generalized linear models, and the logit is the link function. Covariates can be continuous or categorical, and this allows clinical as well as genomic features to be included in the model simultaneously. Logistic classification is simply logistic regression used for classification. For LCA, the categorical response is the class variable *g*. Appropriate cut-offs can be used to

determine classification. For example, $\hat{g}_i = I(\text{logit}(p_i) > C),\ g_i \in \{0,1\}$, where the constant C can be chosen as 0 or selected through nested CV. The distance for LCA is then the logit. Since the logit is the log of the odds, it is then natural to use the treatment arm OR to evaluate the difference in this distance between treatment arms for treatment subset prediction. This is the approach taken in Freidlin & Simon's version of ASD.

Single-Gene Logistic Classification-Treatment Subset Prediction

This version of ASD employs single gene logistic models to predict a subset of patients more likely to respond to treatment. In that method the treatment arm OR is used as a distance in a weighted voting procedure to determine sensitivity to treatment. The single gene logistic model for subject *i* and gene *k* is: $\text{logit}(p_i \mid z_{ik}, t_i) = \mu + \lambda_k t_i + \gamma_k t_i z_{ik}$, where t_i is the treatment arm indicator for subject i in the validation set, z_{ik} is gene expression for gene *k* and subject *i*, and λ_k and γ_{ik} are coefficients for treatment and treatment-gene interaction respectively. Main effects for gene expression are also included in the implementation of Scher, Nasso, Rubin & Simon (2011 [144]),]), though the model is different from the one shown here. When patient sensitivity prediction for the final validation set is undertaken, outcome results for the training set are known, as are gene expression results z_{ik} for subject i in the validation set, even though outcome results g_i are not. Gene-weighted voting estimates for sensitivity to treatment for subject i can then be obtained using the treatment arm OR. Subject *i* has not necessarily been assigned to a treatment arm, however the treatment arm OR compares both treatment assignment scenarios: i.e. $OR_{ik} = e^{\hat{\mu}+\hat{\lambda}_k+\hat{\gamma}_k z_{ik}} / e^{\hat{\mu}} = e^{\hat{\lambda}_k+\hat{\gamma}_k z_{ik}}$ using the parameter estimates on the training set and gene expression z_{ik} from the validation set. A value of OR_{ik} exceeding R results in a vote by gene k for subject i being sensitive to treatment. This can be expressed as:

$Sn_i = I\left\{\left[\sum_{k=1}^{p^*} I(OR_{ik} > R)\right] \geq H\right\}$; where Sn is an indicator variable for subject sensitivity, and $k = 1, ..., p^*$ are the genes selected from the DR step.

The assumed model for the weighted voting method is that some but not all sensitive genes for a subject sensitive to treatment are overexpressed. The weighted voting tuning parameters determine the fraction of genes and the amount of overexpression in order for the subject to be predicted sensitive. However a simplified approach may be to eliminate tuning parameter H by calculating a summary score over all the selected genes for a subject. Figure 1 is a detailed flowchart of the ASD method.

The assumed model for the weighted voting method is that some but not all sensitive genes for a subject sensitive to treatment are overexpressed. The weighted voting tuning parameters determine the fraction of genes and the amount of overexpression in order for the subject to be predicted sensitive. However a simplified approach may be to eliminate tuning parameter H by calculating a summary score over all the selected genes for a subject. Figure 1 is a detailed flowchart of the ASD method.

Multivariable Logistic Classification

The LCA method used as an example in Freidlin & Simon's work was a single gene logistic model. However multivariable multi-gene models can also be used for ASD, as in Scher et al. (2011 [144]); note that in this work a multivariable proportional hazards model was used. A multivariable LCA approach was not used in the Dudoit et al. 2002 work (which did not involve an ASD implementation) due to issues of class separation. In that case ML estimates do not converge. Use of a penalized ML method (Heinze & Schemper, 2002 [76]) is one way to address this, and this option is readily available in R or SAS software. An advantage of the L2R multivariable logistic method of Park & Hastie (2008 [120]) is that it enables incorporation of higher order gene-gene interactions. This method might therefore be preferred to LDA when it is

felt that higher-order interactions are not ignorable. Note that since higher order gene-gene interactions are involved, a prior DR method may still be needed in a typical HTD setting.

5. Discussion and Conclusions

Methods such as LDA and LCA are constructed using distance metrics. The example of the ASD method in Freidlin & Simon's work used the treatment arm *OR* as a distance in a weighted voting method to determine sensitivity to treatment. Other distance measures and other methods can be incorporated into treatment subset prediction. The *LLR* measures the distance of a subject from the border of the two classes. An LLR of 0 indicates the borderline between two classes. The sign and magnitude can be used to develop criteria for voting or allocating a subject to a given class. Sensitive subjects can then be predicted using methods which incorporate this distance. Using single gene densities, the equation for the difference in the LLR_{ik} between the two treatment arms for gene k and subject i in the validation set is:

$$DLLR_{ik} = LLR_{ik,t=1} - LLR_{ik,t=0} = \log \frac{\hat{\pi}_{11} f(z_{ik} \mid \hat{\theta}_{11k})}{\hat{\pi}_{01} f(z_{ik} \mid \hat{\theta}_{01k})} - \log \frac{\hat{\pi}_{10} f(z_{ik} \mid \hat{\theta}_{10k})}{\hat{\pi}_{00} f(z_{ik} \mid \hat{\theta}_{00k})},$$

where $f(z_{ik} \mid \hat{\theta}_{gtk})$ is the density for gene *k* with estimated parameter vector $\hat{\theta}_{gtk}$, and evaluated at z_{ik}. The parameter vector $\hat{\theta}_{gtk}$, is conditional on class *g*, treatment arm *t*, and gene expression for gene *k*. Estimated prior class probabilities $\hat{\pi}_{gt}$ are also conditional on class *g* and treatment arm assignment *t*, but not on gene *k*. The estimates for parameters and prior class probabilities are derived from the training set.

Numeric Example

From training set control arm there are 25 subjects who responded and 75 subjects who did not respond to treatment. On the training set treatment arm, 40 patients responded and 60 did not. Then $\hat{\pi}_{10} / \hat{\pi}_{00} = 0.333$ and $\hat{\pi}_{11} / \hat{\pi}_{01} = 0.667.$ Gene expression density parameters for gene k are estimated from the training set (after appropriate gene expression normalization). If Gaussian

densities are used, then densities for gene *k* conditional on class and treatment arm and evaluated at z_{ik} are $f(z_{ik} \mid \hat{\mu}_{gtk}, \hat{\sigma}^2_{gtk})$. Estimates of parameters conditional on response status *g* and treatment arm *t* can be derived from the training set used to estimate corresponding densities. Suppose for selected gene k we have

$(\hat{\mu}_{gtk}, \hat{\sigma}^2_{gtk}) \in \{(0, 0.04), (0.1, 0.04), (0.1, 0.04), (0.6, 0.16)\}$ for

$(g,t) \in \{(0,0),(1,0),(0,1),(1,1)\}$; gene expression $z_{ik} = 0.9$. Then

$$\frac{f(z_{ik} \mid \hat{\mu}_{11k}, \hat{\sigma}^2_{11k})}{f(z_{ik} \mid \hat{\mu}_{01k}, \hat{\sigma}^2_{01k})} = \frac{f(0.9 \mid 0.6, 0.16)}{f(0.9 \mid 0.1, 0.04)} = 1125,$$

$$\text{and } \frac{f(z_{ik} \mid \hat{\mu}_{10k}, \hat{\sigma}^2_{10k})}{f(z_{ik} \mid \hat{\mu}_{00k}, \hat{\sigma}^2_{00k})} = \frac{f(0.9 \mid 0.1, 0.04)}{f(0.9 \mid 0, 0.04)} = 8.373.$$

$$\text{Therfore } DLLR_{ik} = \log(1125 * 2/3) - \log(8.373 * 1/3) = 5.59.$$

Then, if $5.59 > R$, gene k would cast a vote for sensitivity for patient i in the validation set. The parameter set $\{\eta, R, H\}$ could, as before, be selected using nested CV or the nested bootstrap from a list of prospectively chosen turning parameter sets.

Alternatives to Weighted Voting

Distances such as these can be used in an ASD treatment subset prediction method incorporating tuning parameters very similar to η, R, and H. Alternatively, if the weighted voting assumption that some but not all selected genes are overexpressed is deemed not appropriate, tuning parameter H can be eliminated, and the cut-off *R* can be used on the average $DLLR_{ik}$ of the selected genes for a subject. If the naïve Bayes assumption is severely violated, a multivariable method such as LDA could be used to calculate $DLLR_i$ over all the selected features for subject i in the validation set. This equation then becomes:

$$DLLR_i = \log \frac{\hat{\pi}_{11} f(\boldsymbol{z}_i \mid \hat{\boldsymbol{\mu}}_{11,} \boldsymbol{S}_{.1})}{\hat{\pi}_{01} f(\boldsymbol{z}_i \mid \hat{\boldsymbol{\mu}}_{01,} \boldsymbol{S}_{.1})} - \log \frac{\hat{\pi}_{10} f(\boldsymbol{z}_i \mid \hat{\boldsymbol{\mu}}_{10,} \boldsymbol{S}_{.0})}{\hat{\pi}_{00} f(\boldsymbol{z}_i \mid \hat{\boldsymbol{\mu}}_{00,} \boldsymbol{S}_{.0})},$$

where $\hat{\pi}_{gt}$ indicates estimated prior class probabilities for class *g* and treatment arm *t*, $\hat{\mu}_{gt}$ is the estimated vector of means for selected features conditional on class *g* and treatment arm *t*, $\boldsymbol{S}_{.t}$ denotes the pooled sample covariance matrix for treatment arm *t*, and $\boldsymbol{z}_i$ is the vector of selected features for subject i in the validation set. As before, all estimates are from the training set. Tuning parameter H is also eliminated in this approach. Values of $DLLR_i$ near 0 are indicative of no treatment interaction for subject i.

The Posterior Odds Ratio and Treatment Subset Prediction

It may also be more intuitive to use the estimated or plug-in posterior *OR*:

$OR_{i\ post} = \dfrac{Pr_{post,i11} / (1 - Pr_{post,i11})}{Pr_{post,i10} / (1 - Pr_{post,i10})}$, where the estimated posterior probability

$$Pr_{post,i1t} = \frac{\hat{\pi}_{1t} f(\boldsymbol{z}_i \mid \hat{\boldsymbol{\theta}}_{1t})}{\sum_{g=0}^{1} \hat{\pi}_{gt} f(\boldsymbol{z}_i \mid \hat{\boldsymbol{\theta}}_{gt})}.$$

Also single gene posterior odds ratios $OR_{ik\ post}$ can also be used in the same way as the treatment OR_{ik} in an ASD method. That is, tuning parameter R is used to select a cutoff for the $OR_{ik\ post}$, and parameter H selects the number of genes needed to exceed R in order for the subject to be predicted sensitive. Parametric methods using the above approaches are described below.

Numeric Example for Single-Gene Posterior Odds Ratio

Using the previous numeric example,

$$\frac{\hat{\pi}_{11} f(z_{ik} \mid \hat{\mu}_{11k}, \hat{\sigma}^2_{11k})}{\hat{\pi}_{01} f(z_{ik} \mid \hat{\mu}_{01k}, \hat{\sigma}^2_{01k}) + \hat{\pi}_{11} f(z_{ik} \mid \hat{\mu}_{11k}, \hat{\sigma}^2_{11k})} = 0.9987 \text{ and}$$

$$\frac{\hat{\pi}_{10} f(z_{ik} \mid \hat{\mu}_{10k}, \hat{\sigma}^2_{10k})}{\hat{\pi}_{00} f(z_{ik} \mid \hat{\mu}_{00k}, \hat{\sigma}^2_{00k}) + \hat{\pi}_{10} f(z_{ik} \mid \hat{\mu}_{10k}, \hat{\sigma}^2_{10k})} = 0.7362; \text{ then}$$

$$OR_{ik\ post} = \frac{0.9987/(1-0.9987)}{0.7362/(1-0.7362)} = 89.6, \text{ or } \log(OR_{ik}) = 4.50.$$

If OR_{ik} exceeds tuning parameter R (obtained through nested CV), then the gene k casts a vote in favor sensitivity for subject i.

Linear Discriminant Analysis and Treatment Subset Prediction

The compound distance measure derived in FDA measures how close a subject is to one class or the other. This approach also has the same classification rule as LDA when ML estimates are substituted for true values. For classification purposes, a cutpoint is usually determined based on CV and the subject is allocated based on this decision. However for treatment subset prediction, the difference in this distance across treatment arms can be evaluated. The equation becomes

$$DLLR_i = \log \frac{\pi_{11} \exp\left\{-\frac{1}{2} m_{i11}\right\}}{\pi_{01} \exp\left\{-\frac{1}{2} m_{i01}\right\}} - \log \frac{\pi_{10} \exp\left\{-\frac{1}{2} m_{i10}\right\}}{\pi_{00} \exp\left\{-\frac{1}{2} m_{i00}\right\}},$$

where m_{igt} is the square of the Mahalanobis distance; $m_{igt} = (\boldsymbol{z}_i - \boldsymbol{\mu}_{gt})' \boldsymbol{\Sigma}_{.t}^{-1} (\boldsymbol{z}_i - \boldsymbol{\mu}_{gt})$ for subject i in the validation set, and using parameters for class g, and treatment arm t. This necessitates computation of *m* for each treatment arm as well as each class. A common within-class covariance matrix is assumed for a given treatment arm for LDA. However the covariance matrix might be expected to vary across the two treatment arms. In this case a different within-class pooled covariance matrix could be used across treatment arms ($\boldsymbol{\Sigma}_{.t}$), as shown in the equation above. As before, estimates usually need to be used in place of unknown true parameter values, and prior class probabilities and densities are estimated from the training data. Further improvements can often be obtained by using L2R versions for the covariance matrix such as those described in Ledoit & Wolfe (2004 [99]) and implemented in Xu et al. (2009 [177]), and L1R or L2R versions of LDA, which also have an impact on the covariance matrix. Extensions of the above method to DLDA, weighted voting, and posterior OR's are straightforward.

For QDA and DQDA (Diagonal QDA), to account for the different covariance matrices between classes, the quantities $\exp\left\{-\frac{1}{2}m_{igt}\right\}$ in the above equation must be replaced with

$$\frac{1}{(2\pi)^{p^*/2}\,|\,\boldsymbol{\Sigma}_{gt}\,|^{1/2}}\exp\left\{-\frac{1}{2}\left[(\boldsymbol{z}_i-\boldsymbol{\mu}_{gt})'\boldsymbol{\Sigma}_{gt}^{-1}(\boldsymbol{z}_i-\boldsymbol{\mu}_{gt})\right]\right\}.$$

QDA is more sensitive to departures from normality than LDA, and it is to be expected that instability of the covariance matrix will be even more of an issue with QDA than with LDA, since separate covariance matrices are estimated for each class. There are L2R versions of QDA and DQDA (Table 3) that help address these issues.

For treatment subset prediction in an application involving HTD, a separate DR step is needed before application of unpenalized versions of LDA or QDA. For guidance see Fan & Lv (2010 [16]). For the unpenalized versions of LDA, DR would need to reduce the number of selected feature to be less than treatment arm-specific sample sizes on training set. Specifically: $p^* \le \min\left(n_{D,t=1}-2,\ n_{D,t=0}-2\right)$. Further reduction would be needed, as described earlier, to avoid instability of the covariance matrix. The remaining features are then used to predict a subset of sensitive patients. To select the tuning parameter to be used in the validation set for $DLLR_i$, a nested CV approach can be used on the training set only.

Linear Discriminant Analysis Regression and Treatment Subset Prediction

As stated earlier, linear regression, with appropriately chosen cutoffs for two classes, is equivalent to LDA. However one advantage of using linear regression in place of LDA in an ASD setting is that treatment-gene interactions can be naturally incorporated into the regression model. For example, using an ASD approach similar to Freidlin & Simon: $E(g)=\beta_{0k}+\lambda_k t_i+\gamma_k t_i z_{ik}$, where $g\in\{-1,1\}$. In place of the treatment arm OR in the LCA approach to ASD, this equation has a treatment arm estimate $\hat{\lambda}_k+\hat{\gamma}_k z_{ik}$. Again, for unequal class sizes, appropriate adjustments can easily be made to the outcome variable.

The equivalence of linear regression to LDA extends to multivariable regression as well. In that case treatment gene interaction terms and main effects would be needed for each (selected) gene. Furthermore, taking advantage of the relationship between LDA and regression, use of robust regression methods such as M-estimation (Venables & Ripley, 2002 [165]) and MAD could be explored in a treatment subset method.

Use of Residuals and Treatment Subset Prediction

With many regression methods it is more natural to use a continuous variable as the response. The approach used in Rai et al., (2013 [129]) described in Section 4 can be extended to treatment subset prediction. The classification distance used in that method to classify a subject as diseased $(g=1)$ or disease free $(g=0)$ is $(\bar{p}_{i0}-\bar{p}_{i1})$. A positive distance for subject i is more indicative of disease, and a negative distance is more indicative of a disease-free status. To extend this distance to treatment subset prediction, the models are built separately on both the training set treatment arm and the training set control arm. The 95% quantiles are also determined from these two arms. Then the difference in these distances between the two treatment arms is calculated: $(\bar{p}_{i01}-\bar{p}_{i11})-(\bar{p}_{i00}-\bar{p}_{i10})$, where $\bar{p}_{igt}$ is the average number of residuals exceeding a specified quantile under the model assuming $g=g$ and $t=t$. In the differential scanning calorimetry setting, there is one equation for each subject i, so a tuning parameter similar to R can be used as a cut-off to predict patient sensitivity to treatment, without additional need for tuning parameter H.

6. Challenges and Future Directions

It is known that no single classification method can be optimal for all situations (Wolpert, 1997 [176]), so it is advantageous to have different methods for different data structures. In particular, the DR and classification method should be selected to be consistent with the goal of the study and with the distribution of the features. For example, if the goal is treatment subset prediction, then a DR method that selects features based only on treatment main effects is not optimal. If the data is highly skewed and a transformation such as those found in Box & Cox

(1964 [14]) or Parrish et al. (2009 [121]) cannot be found which results in features which are at least approximately normal, then both a DR and classification method are needed which are robust to departures from normality. For example the use of logistic regression for both DR and classification might be considered in this case (though LDA or DLDA may still perform well when the normality assumption is not severely violated). Table 8 provides some guidance regarding selection of DR and classification methods for difference scenarios and situations.

Table 8: Select parametric dimension reduction and classification methods – situation and conditions.

Method	Situation	Condition
Dimension Reduction		
PLS, PCA	Used in presence of high multicollinearity/correlation between features such as in face recognition or chemometrics. PLS can also be used as a classification method.	$p^*>n_D$, $p^*<n_D$
SPLS or SPCA	Same as above.	$p>> n_D$
t test	Used to screen for treatment main effect, but not treatment-gene interactions; allows moderate deviation from normality (conditional on class) possibly after appropriate transformations; can allow for different variances between classes.	$p>> n_D$
ANOVA	Used to select features based on treatment-feature interactions, or to select features based on more than two groups/classes; assumes constant variance between groups/classes.	$p>> n_D$
Univariable logistic regression	Used to screen for main effects or treatment-feature interactions; no requirement that features follow a specific distribution.	$p>> n_D$
L1R methods	Best applied when only small subset of a large number of covariates/features is active.	$p^*> n_D$, $p>> n_D$
L2R Methods	Best used when there are a large subset of active covariates/features –can be used to identify gene-treatment interactions in presence of gene-gene interactions.	$p^*> n_D$
Classification		
FDA or linear regression	Best performance when deviation from multivariate normality not severe for selected features; assumes common within class covariance matrix.	Recommend $p^*< n_D /5$
Multivariable LCA	Robust to departures from distributional assumptions; often used as standard.	Recommend $p^*< n_D /5$
QDA	Best performance when deviation from multivariate normality not severe for selected features; assumes different covariance matrices between classes. Not robust when assumptions are violated.	Recommend $p^*< n_D /5$
DLDA	Assumes independence between features, but naïve Bayes methods can work well with some correlation; assumes constant variance between classes for each feature. Allows moderate deviation from normality	$p^*< n_D$
DQDA	Assumes independence between features, but naïve Bayes methods can work well with some correlation; assumes different variance between classes for each feature. Allows moderate deviation from normality.	$p^*< n_D$
MLDA	Best used when there are a large number of active features and non-diagonal elements of covariance matrix are not ignorable (naïve Bayes approach no longer optimal).	$p^*> n_D$
Weighted voting	Single gene models "vote" to predict class; in class of naïve Bayes method; can tolerate some correlation between features (as high as 0.6 in Freidlin & Simon 2005 work).	$p^*< n_D$

p - total number of features; p^* - number of selected features; n_D sample size of training set; note - DR methods such as PLS, logistic regression, and L1R/L2R methods can also be used for classification.

There are many challenges remaining:

1. How does one choose between different methods? There are established tests for normality, but some methods such as LDA may still perform well when this assumption is not severely violated. Also transformations may be used to make expression values less skewed and/or more approximately normally distributed. Are ensemble classifiers that incorporate several different classification methods an alternative approach?
2. The assumption inherent in weighted voting is that some but not all selected features in a signature are overexpressed. When is this assumption appropriate? If not appropriate, simplification can be achieved by eliminating parameter H and calculating an average score over all selected features. Further, when are multivariable models preferred over single feature models?
3. Further work needs to be done on evaluating methods which incorporate higher-order gene-gene and gene-treatment interactions for treatment subset prediction.
4. Several works cite the importance of biological pathways in treatment subset prediction. How can pathway analysis be incorporated more directly into treatment subset prediction. Do treatment interactions occur at the gene level or at the pathway level? How can treatment subset prediction be incorporated into studies which include different types of genomic information, such as miRNA's, mRNA's, and SNP's.
5. The purpose of the ASD methods described in this work is not to establish a set of predictive biomarkers but to predict sensitive patients. However the method will not work unless predictive genes are selected, so there is information that could be used. Nested CV on the training set can result in different genes being selected. At the same time, there is information that could be used to help move the process forward, such as the frequency of genes selected for each bootstrap or nested CV. Methods which mine ASD results to help establish predictive biomarkers would seem useful.

ESTIMATING DESIGN PARAMETERS IN THE PRESENCE OF GENE AND GENE -TREATMENT INTERACTION

1. Introduction

The effect or association of specific genes or genomic signatures on response rates to types of treatment are well documented. One example involves triple-negative breast cancer (TNBC), which is breast cancer in absence of staining for the estrogen receptor (ER) progesterone receptor (PR) and HER2 (Irvin & Carey, 2008 [87]). "Neoadjuvant chemotherapy studies have consistently reported higher response rates in TNBC than non-TNBC … The pCR [pathologic complete response] rate in the 23% of patients with TNBC was double that of the non-TNBC subset" (Isakoff , 2010 [88]).

As well, treatments specific to melanoma and applicable to a subset of patients have more recently been coming onto market. For example, subsequent to the findings in Viros et al. (2008 [166]), a treatment Vemurafenib, for a subset of patients having melanoma with BRAF V600e mutations (Chapman et al., 2011 [24]) was approved by the FDA. In the year prior, Ipilimumab was approved by the FDA for treatment of metastatic melanoma (Hodi et al., 2010 [79]). Saenger & Wochok (2009 [139]) had previously shown that heterogeneity is present in patient response to Ipilimumab. Methods are being developed for clinical studies which use genomic information to find a subset of patients which respond differently to treatment. The Adaptive Signature Design (ASD) in Freidlin & Simon (2005 [54]) is used to find a subset of patients responding differently to treatment in phase III clinical studies in cases where a genomic signature has not yet been developed.

Purpose of Study

The ASD model consists of two tests- one to assess overall treatment effect, and another to evaluate treatment effect restricted to a subset of patients prospectively predicted to be sensitive to treatment. Type 1 error controlled by allocating it between these two tests. In this work a method is given for calculating sample size of training set to attain a given power for the subset test. This method takes into account gene overlap between classes.

Organization of Study

Section 2 covers notation and definitions. Section 3 outlines background for the adaptive signature design. Section 4 describes an adaptive signature design similar to Freidlin and Simon. Power for the Adaptive Signature Design is in Section 5, and limitations are in Section 6.

2. Notation and Definitions

The sample sizes of subjects for Stage 1 and Stage 2 are n_1 and n_2 respectively; the total sample size is $n = n_1 + n_2$. Stage 1 serves as the training set where the prediction model is built (using nested cross validation), and Stage 2 serve as the final validation set where patient sensitivity is predicted.

A subject that is sensitive to treatment has a greater probability of response on the treatment arm than on the control arm. Response-to-treatment status is denoted by random variable Y, which takes on values 0 for no response, and 1 for response to treatment. The probability that any new subject accrued into the study is sensitive is p_S, and $p_{R_{ST}}$ is the probability of response given $S_i = s$ and $T_i = t$, where $S \in \{0,1\}$ and $T \in \{0,1\}$ are random variables for patient sensitivity status and patient treatment arm status respectively. The number of evaluated genes is p. Within these p genes it is assumed that there is a set of predictive genes which can be used to predict sensitivity status of subjects. Without loss of generality, it is assumed that the first m genes, $k = 1,...,m$, are the predictive genes. The p-length vector of gene expression for subject i

is denoted by random variable $\mathbf{Z}_i$, and the k^{th} element of that vector (gene expression for the k^{th} gene for subject *i*) is Z_{ik}. Fixed realization of these quantities are denoted by lower case letters.

3. Background for the Adaptive Signature Design

1) Gene expression profile of predictive genes $k = 1,...,m$ follows a multivariate normal distribution which depends on sensitivity status of subject. The distribution of predictive genes is multivariate normally distributed with mean vector $\boldsymbol{\mu}_1$ and covariance matrix $\boldsymbol{\Sigma}_1$ for sensitive patients, and mean vector $\boldsymbol{\mu}_0$ and covariance matrix $\boldsymbol{\Sigma}_0$ for nonsensitive patients.
    a. The sample size planning method proposed constrains these two covariance matrices to be equal.
2) The distribution of nonpredictive genes $k = m+1,..., p$ also follows a multivariate normal distribution, but the means do not depend on sensitivity status- the distribution is the same for all subjects.
3) The distribution of predictive genes is independent of treatment assignment; the results of the assay are not used to assign subjects to a specific treatment arm:
$$\Pr(Z_{ik}, T_i) = \Pr(Z_{ik})\Pr(T_i),\ \forall i, i = 1,...,n; k = 1,...,m.$$
4) The distribution of expression of predictive genes for sensitive patients is not independent of response: $\Pr(Z_{ik}, Y_i) \neq \Pr(Z_{ik})\Pr(Y_i); \forall i$ s.t. $s_i = 1, k = 1,...,m.$
5) The distribution of expression of predictive genes for nonsensitive patients is independent of response: $\Pr(Z_{ik}, Y_i) = \Pr(Z_{ik})\Pr(Y_i); \forall i$ s.t. $s_i = 0, k = 1,...,m;$
6) For the non-predictive genes, expression is independent of response; i.e.:
$$\Pr(Z_{ik}, Y_i) = \Pr(Z_{ik})\Pr(Y_i); i = 1,...,n;\ k = m+1,..., p.$$
7) The nonpredictive genes are constrained to have the same mean regardless of sensitivity status of a subject. In the ASD model proposed by Freidlin & Simon, they have the same

mean as the predictive genes for non-sensitive subjects, but have different variance which is denoted here by $\sigma_{ns}^{\ 2}$: i.e.:

$$Z_{ik} \sim N(\mu_0, \sigma_{ns}^{\ 2}), \forall k,\ k = m+1, ..., p.$$

8) There may also be positive correlation among predictive genes, and correlation among nonpredictive genes.

4. An Adaptive Signature Design Model Similar to Freidlin and Simon

In the ASD model, parameters are estimated over gene expression for both sensitive and non-sensitive subjects. The purpose is to predict patient sensitivity, so patient sensitivity cannot be treated as known in the model. Therefore the model parameters must be estimated on the training set without regard to subject sensitivity status. However gene expression z_{ik} is known, and for predictive genes $k = 1, ..., m$, z_{ik} is conditional on the unknown value of s_i. The model relating expected value of gene expression for predictive genes $E(Z_{ik}), k = 1, ..., m$, patient sensitivity status and treatment arm status to response probability p_{R_i} can be written:

$$\begin{aligned} &\text{logit}(p_{R_{STi}}) = \text{logit}\left(p_{R_i} \mid S_i, T_i, E(\boldsymbol{Z}_i)\right) = \log\left(\frac{p_{R_i} \mid S_i, T_i}{1 - p_{R_i} \mid S_i, T_i}\right) = \beta_0 + \beta_1 T_i \\ &+ \sum_{k=1}^{m} \begin{bmatrix} (1-S_i)\{\beta_{2k} E(Z_{ik} \mid S_i = 0) + \beta_{12k} T_i E(Z_{ik} \mid S_i = 0)\} \\ + S_i\{\beta_{2k} E(Z_{ik} \mid S_i = 1) + \beta_{12k} T_i E(Z_{ik} \mid S_i = 1)\} \end{bmatrix}. \end{aligned} \tag{4.1}$$

Now, if expected value of expression for predictive gene depends only on sensitivity status of subject, (i.e.- $E(Z_{ik} \mid S_i = s) = \mu_{sk}$) then subscript *i* can be dropped since probability of response no longer depends on *i*, and for a sensitive subject on the treatment arm:

$$\text{logit}(p_{R_{11}}) = \text{logit}(p_R \mid S = 1, T = 1) = \beta_0 + \beta_1 + \sum_{k=1}^{m} \left(\beta_{2j} \mu_{1k} + \beta_{12k} \mu_{1k}\right). \tag{4.2a}$$

For non-sensitive subjects on the treatment arm:

$$\text{logit}(p_{R_{01}}) = \text{logit}(p_R \mid S = 0, T = 1) = \beta_0 + \beta_1 + \sum_{k=1}^{m} (\beta_{2k} \mu_{0k} + \beta_{12k} \mu_{0k}). \tag{4.2b}$$

For a sensitive subject on the control arm:

$$\text{logit}(p_{R_{10}}) = \text{logit}(p_R \mid S=1, T=0) = \beta_0 + \sum_{k=1}^{m} \beta_{2k}\mu_{1k}. \quad (4.2c)$$

For a nonsensitive subject on the control arm:

$$\text{logit}(p_{R_{00}}) = \text{logit}(p_R \mid S=0, T=0) = \beta_0 + \sum_{k=1}^{m} \beta_{2k}\mu_{0k}. \quad (4.2d)$$

where β_1 is coefficient for the treatment main effect over all the subjects, β_{2k} is the coefficient for gene expression main effect for the k^{th} predictive gene, and β_{12k} is the treatment-expression interaction effect for the k^{th} predictive gene.

4.1 Predictive Value of Mean Expression

Note that this model assumes that it is the expected or mean value of predictive gene expression for a subject that is predictive of probability of response, but not the variation around the mean. The reasoning behind this is that short term variation in gene expression around the mean can be due to technical variation and short term (e.g. day to day) biological variation, such as variation due to instability in mRNA. These would not be expected to influence p_R. (Over time, mean expression may change, and this change may influence probability of response. This situation is not considered in this work). However, even though in this model the assumption is that it is the mean of expression values which are predictive, these values are not known in practice, so the gene expressions z_{ik} are used in their place for prediction purposes. Sensitivity status for a subject is also not known. The purpose of ASD is to predict sensitivity status. The ASD model, using maximum likelihood estimates from the training set substituted for true unknown parameter values, using selected genes in place of the true unknown predictive genes, is:

$$\text{logit}(\hat{p}_{R_{ti}}) = \text{logit}(\hat{p}_R \mid t_i, z_{ik^*}) = \hat{\beta}_0 + \hat{\beta}_1 t_i + \sum_{k^*=1}^{m^*} \left(\hat{\beta}_{2k} z_{ik^*} + t_i \hat{\beta}_{12k} z_{ik^*} \right),$$

where $k^* = 1, ..., m^*$ are the m^* genes predicted to be sensitive (using the DR step), and where information to predict s_i must be derived from the z_{ik^*} expression values in place of the μ_{sk}, as well as the estimated model coefficients, and their relationship with t_i. The following observations are made:

1) High overlap is obviously desired between the m^* genes predicted to be sensitive, and the m genes that are actually predictive. If there is no overlap, then the classification accuracy will be no greater than that expected due to chance alone.

2) Subscript s is not used in estimated logit $\hat{p}_{R_{ti}}$. Even though expected values of gene expressions for predictive genes are dependent on sensitivity status, gene expression values are used regardless of sensitivity status, since sensitivity status is being predicted and is unknown.

3) The assumption that probability of response is dependent only on treatment arm and sensitivity status and not on subject is made for purposes of generalization for sample size and power analysis. For example in the Dobbin & Simon work (2007 [36]), sample size calculations for probability of correct classification $(PCC(n_1))$ are based on a common effect size for all predictive genes. The effect size is the difference between means of gene expression for predictive genes for sensitive patients and means of gene expression for nonsensitive patients, divided by the standard deviation. For purposes of sample size calculation, Dobbin and Simon argue that the variance estimate used to calculate effect size can be based on the 90 percentile of gene variances.

4) Even though the true logit does not depend on i, the estimated logit does depend on i, since the realized gene expression values must be used in place of true expression means. In order to predict sensitivity for subject i in the validation set one could use the treatment arm odds ratio, as is done in F&S. Subject i has not necessarily been assigned to a treatment arm, but

the treatment arm odds ratio is a measure of the difference in probability of response over the two treatment arms, which in turn is a measure of sensitivity of the subject to treatment.

$$OR_i = \exp\left[\text{logit}(\hat{p}_{R_{1i}}) - \text{logit}(\hat{p}_{R_{0i}})\right] = \exp\left(\hat{\beta}_1 + \sum_{k^*=1}^{m^*} \hat{\beta}_{12k^*} z_{ik^*}\right).$$

Simulations by the authors have also shown that single gene posterior odds ratios based on, for example posterior odds ratios from parametric or nonparametric densities (Cambon, Baumgartner, Brock, Cooper, & Rai 2015 [20]), hold up reasonably well to single gene logistic models, even when model assumptions for LDA are violated.

4.2 Probability of Response

Now probability of response conditional on treatment arm and sensitivity status can be expressed as: $E(Y \mid S = s, T = t) = E(Y_{st}) = p_{R_{st}} = \dfrac{\exp\left(\text{logit}(p_{R_{st}})\right)}{1 + \exp\left(\text{logit}(p_{R_{st}})\right)}$. (4.3)

For example, for sensitive patients on the treatment arm, the probability of response is:

$$E(Y_{11}) = p_{R_{11}} = \frac{\exp\left(\text{logit}(p_{R_{11}})\right)}{1 + \exp\left(\text{logit}(p_{R_{11}})\right)}, \text{ and } E(Y_{01}) = p_{R_{01}} = \frac{\exp\left(\text{logit}(p_{R_{01}})\right)}{1 + \exp\left(\text{logit}(p_{R_{01}})\right)},$$

and similarly for $E(\mathrm{Y}_{10})$ and $E(\mathrm{Y}_{00})$, and where equations for $\text{logit}(p_{R_{st}})$ are given in equations 4.2a-4.2d.

The probability of response on the treatment arm over both sensitive and insensitive subjects is then:

$$E(Y \mid T = 1) = p_{R_E} = (1 - p_S)\, p_{R_{01}} + p_S\, p_{R_{11}},$$

where p_S is the probability that any patient is sensitive to treatment.

The expected probability of response over all subjects on the control arm is then:

$$E(Y \mid T = 0) = p_{R_C} = (1 - p_S)\, p_{R_{00}} + p_S\, (p_{R_{10}}).$$

The marginal probability of response over both treatment arms is:

$$p_R = E(Y) = \sum_{t=0}^{1}\sum_{s=0}^{1}(p_{R_{st}} \mid S=s, T=t)\Pr(S=s)\Pr(T=t)$$
$$= p_{R_{00}}(1-p_S)\Pr(T=0) + p_{R_{10}}p_S\Pr(T=0) + p_{R_{01}}(1-p_S)\Pr(T=1) + p_{R_{11}}p_S\Pr(T=1).$$

If subjects are randomly assigned to either treatment arm with probability 0.5, then

$$p_R = 0.5\left[(1-p_S)(p_{R_{00}} + p_{R_{01}}) + p_S(p_{R_{10}} + p_{R_{11}})\right] = (p_{R_E} + p_{R_C})/2.$$

5. Power for the Adaptive Signature Design

5.1 Power for the Overall Treatment Effect

For a study with *n* patients, the power of a two-sided α -level test to detect a difference in response between two treatment arms with equal sample size is approximately:

$$\text{Power} = 1-\beta = \Phi\left\{\frac{(p_{R_E} - p_{R_C}) - z_{1-\alpha_1}\left(\bar{p}(1-\bar{p})\frac{4}{n}\right)^{1/2}}{\left(p_{R_E}(1-p_{R_E})\frac{2}{n} + p_{R_C}(1-p_{R_C})\frac{2}{n}\right)^{1/2}}\right\} \quad (5.1)$$

(Freidlin & Simon, 2005 [54]) where β is the type II error, $\Phi(\cdot)$ is the cumulative distribution function for the standard normal distribution, p_{R_E} and p_{R_C} are probability of response in enhanced and standard treatment arms respectively, α_1 is the portion of the Type 1 error allocated to the overall test, $z_{1-\alpha}$ is the $(1-\alpha)$ percentile of the standard normal distribution, and $\bar{p} = \frac{p_{R_E} + p_{R_C}}{2}$. For equal probability of treatment assignment 0.5 to each group, $\bar{p} = p_R$. The formula is valid for $n_1\bar{p} \geq 5$. Power calculations based on the continuity corrected arcsine transformation (Dobson & Gebski, 1986 [37]) can offer more accurate power estimations especially for small sample sizes and for proportions close to 0.

5.2 Power for Treatment Effect in Predicted Subset of Patients

Let p_{sens} and p_{spec} denote the sensitivity and specificity of the sensitivity status prediction method. The purpose of ASD is to predict sensitivity status for stage 2 patients, using a model

built on the training set (stage 1 patients), and gene expression from the stage 2 patients reserved for prediction. The probability that a selected patient is sensitive (the positive predictive value *PPV*) is

$$PPV = \frac{(p_S)(p_{sens})}{(p_S)(p_{sens}) + (1-p_S)(1-p_{spec})} \ . \tag{5.2}$$

The expected probability of response for a subject on the treatment arm in the selected subset is $p_{R_{+E}} = PPV(p_{R_{11}}) + (1-PPV)(p_{R_{01}})$; for a subject on the control arm the expected probability of response is $p_{R_{+C}} = PPV(p_{R_{10}}) + (1-PPV)(p_{R_{00}})$; if there is no gene main effect, then $p_{R_{10}} = p_{R_{00}}$, and $p_{+C} = p_C$. In either case, the expected sample size of the predicted subset of patients in stage 2 is

$$n_{+2} = n_2\left[\, p_s p_{sens} + (1-p_s)(1-p_{spec})\right],$$

and the power of the subset comparison using Stage 2 patients predicted to be sensitive to treatment is (Freidlin & Simon, 2005 [54]):

$$1-\beta_+ = \Phi\left\{ \frac{(p_{R_{+E}} - p_{R_{+C}}) - z_{1-\alpha_2}\left(\bar{p}_+(1-\bar{p}_+)\frac{4}{n_{+2}}\right)^{1/2}}{\left(p_{R_{+E}}(1-p_{R_{+E}})\frac{2}{n_{+2}} + p_{R_{+C}}(1-p_{R_{+C}})\frac{2}{n_{+2}}\right)^{1/2}} \right\}. \tag{5.3}$$

where α_2 is the portion of the Type I error allocated to the test for the predicted subset. Note that while equation 5.1 is for all subjects in the study, the power for subset prediction (5.3) is for the subset of the stage 2 patients predicted to be sensitive to treatment. However it is a function of the sample size of the training set through p_{sens} and p_{spec}, which are used to calculate *PPV*, which in turn is used to calculate $p_{R_{+E}}$ and $p_{R_{+C}}$ in equation 5.3. Also note that this equation does not take into account how much the power of the subset test increases the power of the ASD method over

the power for the overall treatment effect alone. In the ASD method, significance is declared if there is either a significant overall effect or a significant subset effect.

5.2.1 Sensitivity and Specificity of Prediction Method

Power of the subset comparison is a function of the sensitivity p_{sens} and specificity p_{spec}, of the prediction method as well as p_S, the probability of a patient being sensitive to treatment, and n_2, the sample size for Stage 2. As p_{sens} and p_{spec} increase, *PPV* and $p_{R_{+E}}$ also increase, thus increasing effect size in the subset. Sensitivity and specificity of the ASD method are in turn a function of the sample size for stage 1 (n_1), the magnitude of the difference in differential expression between treatment arms for predictive genes (for the DR step), and the amount of overlap in multivariate distributions of gene expression profile of selected genes between sensitive and nonsensitive subjects. To the extent there is overlapping space between the sensitivity-status specific multivariate distributions of the predictive genes, then no matter how large the sample size n_1, the predictor will not be able to achieve perfect classification accuracy, and p_{sens} and p_{spec} will not approach 1 even as n_1 becomes very large.

5.2.2 Sample Size Planning for the Simple Two-Class Problem in a High-Dimensional Setting

For the two-class problem, Dobbin & Simon (2007 [36]) outline a method to calculate $PCC(n_1)$ (probability of correct classification given a training set sample size of n_1) in a high dimensional setting, taking into account distribution of gene expression for the two classes. It is assumed a small proportion of genes m/p is predictive of class status. If $PCC(n_1)$ is calculated for each class, then this gives estimates for sensitivity and specificity for the classification method. If a way can be found to apply or adapt this method to the ASD setting, then p_{sens} and p_{spec} could be calculated as a function of n_1 and used to derive quantities in (5.3) to estimate power for the

subset of patients in Stage 2 predicted to be sensitive to treatment. Their method (referred to as D&S) takes into account variation in both the DR and the classification step.

Now the normal approximation sample size formula, applied to the training set, (and assuming gene expression for one differentially expressed gene) is:

$$n_1 \approx 4\frac{\sigma^2}{(2\delta)^2}(t_{n_1-2,1-\alpha/2} + t_{n_1-2,1-\beta})^2, \tag{5.4}$$

where 2δ is the difference between class means, σ^2 is the within-class variance, $1-\alpha$ is the specificity associated with correctly identifying a gene that is not differentially expressed, $t_{n_1-2,1-\alpha/2}$ is the quantile function of the central *t*-distribution with $n_1 - 2$ degrees of freedom at probability $1-\alpha/2$ – i.e. $t_{n_1-2,1-\alpha/2} = T_{n-2}^{-1}(1-\alpha/2)$, where $T_{n-2}^{-1}(\cdot)$ is the inverse cumulative distribution function for a central *t*-distribution with $n-2$ degrees of freedom.

Dobbin & Simon (2007 [36]) derived the following approximate formula for the power for equation (5.4):

$$1-\beta \approx T_{n_1-2}\left(\frac{\delta}{\sigma}\sqrt{n_1} - t_{n_1-2,1-\alpha/2}\right), \tag{5.5a}$$

where $T_{n-2}(\cdot)$ is the cumulative distribution function for a central *t*-distribution with $n-2$ degrees of freedom. The formula derived by Chow, Shao, & Wang (2002 [26]) takes into account the fact that power is under the alternative hypothesis and therefore uses the cumulative distribution function for the noncentral *t*-distribution:

$$1-\beta \approx 1-T_{n_1-2}\left(t_{n_{01}+n_{11}-2,1-\alpha/2}\left|\frac{2\delta}{\sigma\sqrt{\frac{1}{n_{01}}+\frac{1}{n_{11}}}}\right.\right), \tag{5.5b}$$

where $t_{n_{01}+n_{11}-2,1-\alpha/2}$ is the quantile function for a central *t*-distribution with degrees of freedom $n_{01}+n_{11}$ (where n_{01} and n_{11} are class-specific sample sizes on the training set - $n_{01}+n_{11}=n_1$),

and evaluated at quantile $1-\alpha/2,$ and $T_{n-2}(\bullet|\,\theta)$ is the cumulative distribution function of the noncentral *t*-distribution with n_1-2 degrees of freedom and noncentrality parameter $\theta.$

Now (5.5a) and (5.5b) are expressions for power to detect a difference between two groups using one differentially expressed gene. Dobbin & Simon also derived an expression for a linear classifier with *m* differentially expressed genes with equal effect sizes $2\delta/\sigma$ and $p-m$ nondifferentially expressed genes (details in Appendix and in D&S). When prior class probabilities are equal,

$$PCC(n_1) \geq \Phi\left(\frac{\delta}{\sigma\sqrt{\lambda_1}}\frac{m(1-\beta)}{\sqrt{m(1-\beta)+(p-m)\alpha}}\right), \tag{5.6a}$$

where $1-\beta$ is sensitivity or power for correctly selecting any differentially expressed gene, λ_1 is the largest eigenvalue of the correlation matrix of the genes (see Schott, 2005 [146]), and $1-\alpha$ is specificity for correctly identifying a non-differentially expressed gene. Each differentially expressed gene has difference in class means $2\delta,$ and both predictive and nonpredictive genes have within-class standard deviation $\sigma.$

This lower bound for $PCC(n_1)$ in 5.6a takes correlation between genes into account, using properties of extremal eigenvalues (Schott, 2005 [146]). However in simulations and applications involving real data sets in D&S, it was found that this approach tended to be overly conservative, and that an equation for $PCC(n_1)$ assuming gene expression independence was either accurate (accurately estimated the required sample size) or conservative (the sample size was higher than required). The equation assuming gene expression independence takes advantage of the fact that under that assumption, $\sqrt{\lambda_1}=1$. The equation then becomes:

$$PCC(n_1) \approx \Phi\left(\frac{\delta}{\sigma}\frac{m(1-\beta)}{\sqrt{m(1-\beta)+(p-m)\alpha}}\right), \tag{5.6b}$$

Now in many settings (including the ASD setting), the populations of sensitive and nonsensitive patients are unlikely to be equal. In this case, if equation 5.6a is applied to class prediction, the equation becomes:

$$PCC(n_1) \geq p_S \Phi\left(\frac{\delta}{\sigma\sqrt{\lambda_1}} \frac{m(1-\beta)-k}{\sqrt{m(1-\beta)+(p-m)\alpha}} \right) + (1-p_S)\Phi\left(\frac{\delta}{\sigma\sqrt{\lambda_1}} \frac{m(1-\beta)+k}{\sqrt{m(1-\beta)+(p-m)\alpha}} \right) \quad (5.6c)$$

where $k = \frac{1}{2}\log\frac{1-p_S}{p_S}$. As before, if gene independence is assumed, then $\sqrt{\lambda_1} = 1$ and the two sides of the equation are approximately equal. Dobbin and Simon also showed how to modify 5.6a to control *PCC* in each class (particularly the smaller class, which will have a lower *PCC*).

This formula for a linear classifier is conservative and does not assume Bayes rule, but under this approach, as n_1 becomes large, $PCC(n_1)$ does approach $PCC(\infty)$, the probability of correct classification assuming the Bayes rule is the normal classifier and assuming independence between genes.

Then D&S used (5.5a) to eliminate $1-\beta$ in (5.6a). The equation for the simple two-class problem (which is not equivalent to the ASD setting), assuming equal prior class probabilities and based on the normal approximation sample size formula and assuming gene independence is then:

$$PCC(n_1,\alpha) \approx \Phi\left(\frac{\frac{m\delta}{\sigma} T_{n_1-2}\left(\frac{\delta}{\sigma}\sqrt{n_1} - t_{n_1-2,1-\alpha/2} \right)}{\left[mT_{n_1-2}\left(\frac{\delta}{\sigma}\sqrt{n_1} - t_{n_1-2,1-\alpha/2} \right) + \alpha(p-m) \right]^{1/2}} \right), \quad (5.7)$$

where $PCC(n_1,\alpha)$ is the probability of correct classification given sample size n_1 and given α. For unequal prior class probabilities, 5.5a could be substituted into 5.6c instead. Note that equation 5.5b could also be substituted into equation 5.6a or 5.6c in which case the cumulative distribution function for the noncentral *t*-distribution is used. For a given n_1, (5.7) can be used to

find an optimal α for gene selection that will maximize $PCC(n_1)$. It can also by extension be used to estimate *PCC* for different n_1.

Now the value of the quantile function for the *t*-distribution ($t_{n_1-2,1-\alpha/2}$) included in (5.5a) and (5.5b) is directly related to a DR method often used in the two-class setting. This DR method is based on class-specific parameters only and is, for each gene *k*:

$$t_{Dk} = \frac{(\hat{\mu}_{1k} - \hat{\mu}_{0k})}{\sqrt{\hat{\sigma}_{k_g}{}^2 (1/n_1 + 1/n_0)}},\ k = 1,..., p, \tag{5.8}$$

where $\hat{\sigma}_{k_g}{}^2$ is the variance pooled over response status group for gene *k* on the training set, $\hat{\mu}_{gk}$ are class -specific means for gene *k* on the training set, and n_g , $g \in \{0,1\}$, is the sample size of subjects specific to class *g* on the training set*;* i.e. - $n_0 + n_1 = n_D$. In this DR method, t_{Dk} is used to evaluate gene *k* based on a *t*-statistic with $n_1 - 2$ degrees of freedom. If the absolute value of the *t*-statistics for gene *k* exceeds $t_{n_1-2,1-\alpha/2}$, then the gene is selected. Next note that the cumulative distribution function in (5.5), which is used to approximate sensitivity for gene selection, decreases with increasing $t_{n_1-2,1-\alpha/2}$, and increases with increasing values of $\frac{\delta}{\sigma}\sqrt{n_1}$ with α held constant. In fact, if effect size and n_1 are sufficiently large, so that α can be kept very small to minimize the quantity $\alpha(p-m)$ in the denominator (also assuming *p*-*m* is not too large), then $PCC(n_1, \alpha)$ will approach $\Phi\left(\frac{m^{1/2}\delta}{\sigma}\right)$, which is the probability of correct classification assuming the Bayes rule is known and that it is the normal classifier, and assuming gene independence.

Note that (5.5a) and (5.5b) are expressions for power or sensitivity of gene selection, and therefore the equations are a function of the DR step. From the same equation it can be seen that the sensitivity of the DR step is a function of effect size and sample size n_1. Reducing α reduces

the sensitivity but increases specificity $1-\alpha$. Note that the DR step can take advantage of increasing values of n_1 so that, with $\sqrt{n_1}$ large enough, high sensitivity (and specificity) for gene selection can be obtained even with smaller values of $\frac{\delta}{\sigma}$. However, even if the DR step perfectly separates out the predictive genes, the probability of correct classification will be limited by the effect size $\frac{2\delta}{\sigma}$, which is independent of n_1. Subjects are classified one at a time, not as a group. No matter how large n_1, the sample size for each subject *i* in the validation set is still 1, and classification decisions are based on the $\mathbf{z}_i$ vector for the subject and the difference in sensitivity-status specific distributions of the different $\mathbf{z}_i$.

5.2.3 Dimension Reduction and the Adaptive Signature Design

The two main differences between settings for ASD and D&S are related to distributions for sensitivity status for ASD, and difference in probability of response of sensitive patients between treatment arms. In the ASD setting, expression distribution of predictive genes depends on sensitivity status of patient, which is unknown, but is being predicted. Sensitivity status is measured by the extent to which probability of response for a patient is greater in the treatment arm than in the control arm. The DR step can be motivated by the fact that, assuming there are sensitive patients enrolled in the study, then expression for predictive genes should have a greater difference in response status-specific distributions in the treatment arm than in the control arm, whereas nonpredictive genes should show little or no difference. The DR method used in simulations in Freidlin, Jiang, & Simon (2010 [36]) compares differential expression between responders and non-responders in the treatment group to differential expression in the control group. The equation is:

$$t_{D_D k} = \frac{(\hat{\mu}_{11k} - \hat{\mu}_{01k}) - (\hat{\mu}_{10k} - \hat{\mu}_{00k})}{\sqrt{\hat{\sigma}_{k_{gt}}{}^2 (1/n_{11} + 1/n_{01} + 1/n_{10} + 1/n_{00})}},\ k = 1, \ldots, p, \tag{5.9}$$

where $\hat{\sigma}_{k_{gt}}{}^2$ is the variance pooled over each combination of treatment-specific and response status-specific groups for gene *k* on the training set, $\hat{\mu}_{gtk}$ are class and treatment arm-specific means for gene *k* on the training set $(g \in \{0,1\}, t \in \{0,1\})$, and n_{gt} is the sample size of subjects specific to class *g* and treatment arm *t* on the training set; i.e. - $n_{00} + n_{01} + n_{10} + n_{11} = n_1$. Note that in a DR step for gene selection $t_{D_D k}$ can be evaluated as a *t*-statistic with $n_1 - 4$ degrees of freedom, assuming distributions of means are approximately normally distributed. The selection criteria could be based on the absolute value of the *t*-statistic: if $|t_{D_D k}| > t_{n_1-4,1-\alpha^*/2}$, then gene *k* is selected.

The expression for power assuming one gene with a difference in differential expression between the two treatment arms can be derived as follows:

$$\left| \frac{(\hat{\mu}_{11k} - \hat{\mu}_{01k}) - (\hat{\mu}_{10k} - \hat{\mu}_{00k})}{\sqrt{\hat{\sigma}_{k_{gt}}{}^2 (1/n_{11} + 1/n_{01} + 1/n_{10} + 1/n_{00})}} \right| > t_{1-\alpha/2, n_{11}+n_{01}+n_{10}+n_{00}-4}$$

Under the alternative hypothesis that $\delta \neq 0$, after ignoring a small term of value $\leq \alpha/2$, it follows that:

$$1-\beta \approx 1 - T_{n_{01}+n_{11}+n_{10}+n_{00}-4}\left(t_{n_{01}+n_{11}+n_{10}+n_{00}-4,1-\alpha/2} \left| \frac{2\delta^*}{\sigma^* \sqrt{\frac{1}{n_{01}} + \frac{1}{n_{11}} + \frac{1}{n_{10}} + \frac{1}{n_{00}}}} \right. \right), \qquad (5.10)$$

where here $2\delta^* = (\mu_{11k} - \mu_{01k}) - (\mu_{10k} - \mu_{00k})$ is the mean difference in differential expression between the treatment and control arms, σ^{*2} is the variance pooled over response status and treatment arm, the n_{gt} are the class and treatment arm-specific sample sizes on the training set, $t_{n_{01}+n_{11}+n_{10}+n_{00}-4,1-\alpha/2}$ is the quantile function for the central *t*-distribution with degrees of freedom $n_{01} + n_{11} + n_{10} + n_{00} - 4$, and $T_{n_{01}+n_{11}+n_{10}+n_{00}-4}(\bullet|\theta)$ is the cumulative distribution function of the

noncentral *t*-distribution with $n_{01}+n_{11}+n_{10}+n_{00}-4$ degrees of freedom and noncentrality parameter θ.

Now in order to calculate $2\delta^*$ and σ^{*2} we need expectation of (predictive) gene expression given sensitivity status $E(Z|\mathrm{S}=s)$, the proportion of sensitive patients p_S, the proportion of sensitive patients on each treatment arm who respond $p_{R_{S=1|T=t}}$ and the proportion of nonsensitive patients who respond $p_{R_{S=0|T=t}}$. The equations for response status and treatment arm specific gene expression means (μ_{gt}) for predictive genes, are:

$$E(Z|Y=1,T=1)=\mu_{11}=\frac{p_S p_{R_{S=1,T=1}} E(Z|S=1)}{p_S p_{R_{S=1,T=1}}+(1-p_S)p_{R_{S=0,T=1}}}+\frac{(1-p_S)p_{R_{S=0,T=1}} E(Z|S=0)}{p_S p_{R_{S=1,T=1}}+(1-p_S)p_{R_{S=0,T=1}}}. \quad (5.11a)$$

$$E(Z|Y=0,T=1)=\mu_{01}$$

$$=\frac{p_S(1-p_{R_{S=1,T=1}})E(Z|S=1)}{p_S(1-p_{R_{S=1,T=1}})+(1-p_S)(1-p_{R_{S=0,T=1}})}+\frac{(1-p_S)(1-p_{R_{S=0,T=1}})E(Z|S=0)}{p_S(1-p_{R_{S=1,T=1}})+(1-p_S)(1-p_{R_{S=0,T=1}})}. \quad (5.11b)$$

$$E(Z|Y=1,T=0)=\mu_{10}=\frac{p_S p_{R_{S=1,T=0}} E(Z|S=1)}{p_S p_{R_{S=1,T=0}}+(1-p_S)p_{R_{S=0,T=0}}}+\frac{(1-p_S)p_{R_{S=0,T=0}} E(Z|S=0)}{p_S p_{R_{S=1,T=0}}+(1-p_S)p_{R_{S=0,T=0}}}. \quad (5.11.c)$$

$$E(Z|Y=0,T=0)=\mu_{00}$$

$$=\frac{p_S(1-p_{R_{S=1,T=0}})E(Z|S=1)}{p_S(1-p_{R_{S=1,T=0}})+(1-p_S)(1-p_{R_{S=0,T=0}})}+\frac{(1-p_S)(1-p_{R_{S=0,T=0}})E(Z|S=0)}{p_S(1-p_{R_{S=1,T=0}})+(1-p_S)(1-p_{R_{S=0,T=0}})}. \quad (5.11d)$$

The variances for response status and treatment arm specific gene expression (σ_{gt}^2) , for predictive genes, are:

$$\mathrm{Var}(Z \mid Y=1, T=1)$$

$$= \sigma^2{}_{11} = \frac{p_S\, p_{R_{S=1,T=1}}(\mu^2_{S=1}+\sigma^2_{S=1})}{p_S\, p_{R_{S=1,T=1}}+(1-p_S)\, p_{R_{S=0,T=1}}} + \frac{(1-p_S)\, p_{R_{S=0,T=1}}(\mu^2_{S=0}+\sigma^2_{S=0})}{p_S\, p_{R_{S=1,T=1}}+(1-p_S)\, p_{R_{S=0,T=1}}} - \mu_{11}{}^2. \qquad (5.12a)$$

$$Var(Z \mid Y=0, T=1) = \sigma^2_{01}$$

$$= \frac{p_S(1-p_{R_{S=1,T=1}})(\mu^2_{S=1}+\sigma^2_{S=1})}{p_S(1-p_{R_{S=1,T=1}})+(1-p_S)(1-p_{R_{S=0,T=1}})} + \frac{(1-p_S)(1-p_{R_{S=0,T=1}})(\mu^2_{S=0}+\sigma^2_{S=0})}{p_S(1-p_{R_{S=1,T=1}})+(1-p_S)(1-p_{R_{S=0,T=1}})}$$

$$-\mu^2_{01}. \qquad (5.12b)$$

$$\mathrm{Var}(Z \mid Y=1, T=0) = \sigma^2_{10}$$

$$= \frac{p_S\, p_{R_{S=1,T=0}}(\mu^2_{S=1}+\sigma^2_{S=1})}{p_S\, p_{R_{S=1,T=0}}+(1-p_S)\, p_{R_{S=0,T=0}}} + \frac{(1-p_S)\, p_{R_{S=0,T=0}}(\mu^2_{S=0}+\sigma^2_{S=0})}{p_S\, p_{R_{S=1,T=0}}+(1-p_S)\, p_{R_{S=0,T=0}}} - \mu^2_{10}. \qquad (5.12c)$$

$$\mathrm{Var}(Z \mid Y=0, T=0) = \sigma^2_{00}$$

$$= \frac{p_S(1-p_{R_{S=1,T=0}})(\mu^2_{S=1}+\sigma^2_{S=1})}{p_S(1-p_{R_{S=1,T=0}})+(1-p_S)(1-p_{R_{S=0,T=0}})} + \frac{(1-p_S)(1-p_{R_{S=0,T=0}})(\mu^2_{S=0}+\sigma^2_{S=0})}{p_S(1-p_{R_{S=1,T=0}})+(1-p_S)(1-p_{R_{S=0,T=0}})}$$

$$-\mu^2_{00}. \qquad (5.12d)$$

Now note that a pooled standard deviation for σ^* assumes that variances in each of the four groups are equal. However based on equations 5.12a through 5.12d, the variances cannot be expected to be equal. Sample size calculations for unequal variances using the noncentral *t*-distribution have been presented in Harrison and Brady (2004 [72]). These use adjustments for degrees of freedom based on Satterthwaite (1946 [142]) or Welch (1947 [172]). These methods can also be used to modify degrees of freedom ν for the *t*-statistic ($t_{n_{01}+n_{11}+n_{10}+n_{00}-4,1-\alpha/2}$) in 5.10

to account for the unequal variances. For example based on Satterthwaite's formula (1941 [143])

$$\nu = \frac{\left(\sigma_{11}^2 / n_{11} + \sigma_{01}^2 / n_{01} + \sigma_{10}^2 / n_{10} + \sigma_{00}^2 / n_{00}\right)^2}{\frac{\left(\sigma_{11}^2 / n_{11}\right)^2}{n_{11}-1} + \frac{\left(\sigma_{10}^2 / n_{10}\right)^2}{n_{10}-1} + \frac{\left(\sigma_{01}^2 / n_{01}\right)^2}{n_{01}-1} + \frac{\left(\sigma_{00}^2 / n_{00}\right)^2}{n_{00}-1}}.$$

The formula for power (or sensitivity to correctly identify any gene with a significant difference in differential expression between the two arms) in 5.10 is then modified for unequal variances as follows:

$$1-\beta \approx 1-T_\nu\left(t_{\nu,1-\alpha/2} \left| \frac{2\delta^*}{\sqrt{\frac{\sigma_{01}^2}{n_{01}} + \frac{\sigma_{11}^2}{n_{11}} + \frac{\sigma_{10}^2}{n_{10}} + \frac{\sigma_{00}^2}{n_{00}}}} \right.\right). \tag{5.13}$$

Again, for each treatment group, these are calculated from the proportion of sensitive patients, the proportion of sensitive subjects expected to respond and the proportion of nonsensitive subjects expected to respond. The expected sample sizes n_{gt} given a training set of size n_1 and equal allocation of subjects to the two treatment arms are:

$$n_{11} = 0.5n_1\{p_S p_{R_{S=1,T=1}} + (1-p_S)p_{R_{S=0,T=1}}\}$$

$$n_{01} = 0.5n_1\{p_S(1-p_{R_{S=1,T=1}}) + (1-p_S)(1-p_{R_{S=0,T=1}})\}$$

$$n_{10} = 0.5n_1\{p_S p_{R_{S=1,T=0}} + (1-p_S)p_{R_{S=0,T=0}}\}$$

$$n_{00} = 0.5n_1\{p_S(1-p_{R_{0=1,T=0}}) + (1-p_S)(1-p_{R_{S=0,T=0}})\}$$

Note that, on the control arm, if there is no difference in probability of response between sensitive and nonsensitive patients, then $p_{R_{T=0}} = p_{R_{S=1,T=0}} = p_{R_{S=0,T=0}}$, and

$n_{10} = 0.5p_{R_{T=0}}n_1$ and $n_{00} = 0.5(1-p_{R_{T=0}})n_1$. There will also be simplifications for equations for σ_{gt}^2 and μ_{gt}.

5.2.4 Probability of Correct Classification and the Adaptive Signature Design

The previous section has shown how to modify $PCC(n_1)$ to take into account the DR step in the ASD setting (equation 5.13) as opposed to the simple two-class problem. Now, the extent that the DR step correctly selects genes with a difference in differential expression between the two arms, the simple linear classifier proposed by D&S and applied to the ASD setting is left with the selected genes from the vector $\mathbf{z}_i$ to make a classification decision for sensitivity status for each subject. Remember that $2\delta/\sigma$ has been previously defined as the effect size for predictive genes, between sensitive and nonsensitive subjects.

Using this approach, assuming equal class sizes and gene independence, equation 5.6b can be modified as follows:

$$PCC(n_1) \approx \Phi\left(\frac{\frac{\delta}{\sigma} m \left\{ 1 - T_\nu \left(t_{\nu,1-\alpha/2} \middle| \frac{2\delta^*}{\sqrt{\frac{\sigma_{01}^2}{n_{01}} + \frac{\sigma_{11}^2}{n_{11}} + \frac{\sigma_{10}^2}{n_{10}} + \frac{\sigma_{00}^2}{n_{00}}}} \right) \right\}}{\sqrt{m \left\{ 1 - T_\nu \left(t_{\nu,1-\alpha/2} \middle| \frac{2\delta^*}{\sqrt{\frac{\sigma_{01}^2}{n_{01}} + \frac{\sigma_{11}^2}{n_{11}} + \frac{\sigma_{10}^2}{n_{10}} + \frac{\sigma_{00}^2}{n_{00}}}} \right) \right\} + (p-m)\alpha}} \right), \tag{5.14}$$

If class sizes are not equal, as is likely for sensitivity status in the ASD setting, the method by D&S to adapt this equation to control *PCC* in the rarer class can be used. Alternatively we can modify equation 5.6c as follows (again setting $\sqrt{\lambda} = 1$ for gene independence):

$$PCC(n_1) \approx p_S \Phi \left(\frac{\frac{\delta}{\sigma} m \left\{ 1 - T_\nu \left(t_{\nu,1-\alpha/2} \left| \frac{2\delta^*}{\sqrt{\frac{\sigma_{01}^2}{n_{01}} + \frac{\sigma_{11}^2}{n_{11}} + \frac{\sigma_{10}^2}{n_{10}} + \frac{\sigma_{00}^2}{n_{00}}}} \right. \right) \right\} - k}{\sqrt{m \left\{ 1 - T_\nu \left(t_{\nu,1-\alpha/2} \left| \frac{2\delta^*}{\sqrt{\frac{\sigma_{01}^2}{n_{01}} + \frac{\sigma_{11}^2}{n_{11}} + \frac{\sigma_{10}^2}{n_{10}} + \frac{\sigma_{00}^2}{n_{00}}}} \right. \right) \right\} + (p-m)\alpha}} \right)$$

$$+ (1 - p_S) \Phi \left(\frac{\frac{\delta}{\sigma} m \left\{ 1 - T_\nu \left(t_{\nu,1-\alpha/2} \left| \frac{2\delta^*}{\sqrt{\frac{\sigma_{01}^2}{n_{01}} + \frac{\sigma_{11}^2}{n_{11}} + \frac{\sigma_{10}^2}{n_{10}} + \frac{\sigma_{00}^2}{n_{00}}}} \right. \right) \right\} + k}{\sqrt{m \left\{ 1 - T_\nu \left(t_{\nu,1-\alpha/2} \left| \frac{2\delta^*}{\sqrt{\frac{\sigma_{01}^2}{n_{01}} + \frac{\sigma_{11}^2}{n_{11}} + \frac{\sigma_{10}^2}{n_{10}} + \frac{\sigma_{00}^2}{n_{00}}}} \right. \right) \right\} + (p-m)\alpha}} \right).$$

6. Limitations and Future Work

There are some differences between the classification of sensitivity status in an ASD setting and classification in the simple two-class problem outlined in D&S. For example, sensitivity status does not become known on the training set. Instead a CV method is used which chooses a set of tuning parameters which optimize sensitivity status prediction. Simulations need to be performed to evaluate what effect this has on sample size and power estimation. Also the use of the Student's *t*-distribution assumes that the sample means $\hat{\mu}_{gt}$ are at least approximately normally distributed. This assumption is likely to break down with small sample sizes n_{gt}, because of the mixture of normal distributions. A nonparametric approach based on ranks along the lines of those described in Conover and Iman (1981 [29]) or power and sample size methods

for logistic regression methods on Shieh (2000 [149]) and Self, Mauritsen, & Ohara (1992 [148]) may be avenues for exploration.

Still another approach may be to use an Expectation Maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977 [34]) to attempt to retrieve the distributions based on sensitivity status, which is a latent variable. An EM approach for Gaussian mixtures involving latent variables is outlined in Bishop (2006 [10]). A Bayesian method is also outlined which addresses "significant limitations in the maximum likelihood approach". If an EM or Bayesian method were to be adopted, the challenge would still remain of incorporating a power and sample size method for the approach.

Finally, another limitation is that the method described in this work assumes equal within-class covariance matrices for each class. The works of Freidlin & Simon assume unequal covariance matrices for each sensitivity status class.

Appendix

Linear Classifier for PCC(n)

Linear classifier with weights 1 for genes that are selected and 0 for those that are not, assuming centering of gene expression class means around 0, and assuming equal prior probabilities: i.e. the linear classifier makes the following classification decision: $\hat{g}_i = I\left(\mathbf{w}'\mathbf{z}_i > 0\right)$. Here $\mathbf{w}$ is a vector of weights of 0's and 1's. This is consistent with a DR step for gene selection. In the case of unequal prior probabilities, the equation is

$$\hat{g}_i = I\left(\mathbf{w}'\mathbf{z}_i > \frac{1}{2}\log\frac{1-\pi_1}{\pi_1}\right).$$

The equation for probability of correct classification for this linear classifier assumes *m* differentially expressed genes, all with same effect size $2\delta/\sigma$, as well as $p-m$ nondifferentially expressed genes. More details can be found in Dobbin & Simon (2007)

Note on Equation for Probability of Correct Classification

A method is outlined in D&S for deriving $PCC(\infty)$, which is the probability of correct classification given that the Bayes rule is known and that it is the normal classifier (Linear Discriminant Analysis or LDA). The equation for $PCC(\infty)$ can be derived from LDA by setting elements of the $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_0$ vectors of expected gene expression means to 1 and -1 respectively for predictive genes and sensitive , and both to 0 for nonpredictive genes. The pooled correlation matrix $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_0$ can be diagonal for both predictive and nonpredictive genes. Correlation between genes can also be taken into account. In this setting, it can be shown that the normal classifier is itself normally distributed since it is a linear combination of the elements of the multivariate normally distributed vector $\mathbf{z}_i$ vector times a constant. This is important because the classifier is essentially a test statistic that makes classification decisions at the unit level - for each subject *i*. Therefore asymptotic normality cannot be invoked, as is often done for test statistics when testing for difference between two groups. However one limitation of using this method in the ASD context is that although gene expression is assumed to be multivariate normally distributed, the relationship with expected value of gene expressions and response is assumed to follow a logistic model. So we can expect $PCC(\infty)$ assuming LDA is the Bayes rule to be optimistic if the logistic model is the true model, since LDA has superior performance to logistic regression when assumptions for LDA are true. However what we are really after is $PCC(n_1)$, which is the probability of correct classification given the sample size on the training set. The expression for $PCC(n_1)$ derived in Dobbin & Simon (2007 [36]) does not assume LDA, and the sample sizes given are shown to be conservative in many cases.

Bayes rule assuming the normal classifier

It is assumed that, at least after log normalization, gene expression between classes is multivariate normally distributed, and that variance (or covariance matrix) for the classes are the

same. In addition it is assumed that gene expression is log normalized and standardized in such a way that the mean vector for the two difference classes is centered around 0. The reason for this is that, for linear discriminant analysis, quadratic terms drop out and, if gene expression vector $\boldsymbol{z}_i$ is multivariate normally distributed, then the classifier itself is normally distributed since it is a linear combination of the multivariate normally distributed variable $\boldsymbol{z}_i$. Specifically

$$LLR_i = \log\left(\frac{\pi_1}{\pi_0}\right) + \left\{-\frac{1}{2}(\boldsymbol{z}_i - \boldsymbol{\mu_1})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}_i - \boldsymbol{\mu_1}) + \frac{1}{2}(\boldsymbol{z}_i - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}_i - \boldsymbol{\mu}_0)\right\}$$
$$= \log\left(\frac{\pi_1}{\pi_0}\right) + \left\{(\boldsymbol{\mu_1} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}\boldsymbol{z}_i - \frac{1}{2}(\boldsymbol{\mu_1} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu_1} + \boldsymbol{\mu}_0)\right\}.$$

And if the class mean vectors are centered around 0, this becomes

$LLR_i = \log\left(\frac{\pi_1}{\pi_0}\right) + \left\{(\boldsymbol{\mu_1} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}\boldsymbol{z}_i\right\} = \log\left(\frac{\pi_1}{\pi_0}\right) + \left\{\boldsymbol{a}'\boldsymbol{z}_i\right\}$. The classifier itself is then normally distributed: if $\boldsymbol{\mu}_1 = 1$ and $\boldsymbol{\mu}_0 = -1$ then

$$LLR_i = \log\left(\frac{\pi_1}{\pi_0}\right) + \left(2\boldsymbol{\mu_1}'\boldsymbol{\Sigma}^{-1}\boldsymbol{z}_i\right), \text{ or } \log\left(\frac{1-\pi_1}{\pi_1}\right) = \left(2\boldsymbol{\mu_1}'\boldsymbol{\Sigma}^{-1}\boldsymbol{z}_i\right).$$

PROPERTIES OF ADAPTIVE CLINICAL TRIAL SIGNATURE DESIGN IN THE PRESENCE OF GENE AND GENE –TREATMENT INTERACTION

Abstract

In this work properties of the adaptive signature design are investigated through simulation. The scenarios include presence of gene expression-treatment interaction effect only, presence of both gene expression main effect and expression-treatment interaction, and presence of expression, treatment, and expression-treatment interaction. Classification methods are examined which both include and exclude gene expression main effect. It was found that, under the scenarios considered, the models which exclude expression main effect while including treatment main effect and expression-treatment interaction often had superior performance to models which included expression main effect.

Key Words: classification; machine learning; dimension reduction; interaction; melanoma; clinical study.

1. Introduction

Freidlin & Simon (2005 [54]) introduced the adaptive signature design (ASD) to predict a subset of patients more sensitive to treatment. A flowchart of this method is shown in Figure 1. The first stage of the two-stage design is used to develop a predictive model prospectively. When the predictive model is built, response and gene expression data is available for Stage 1 patients. Response information becomes available after predictions are made. The classification rule is then applied to the Stage 2 patients to predictive a subset of patients more sensitive to treatment. Gene expression for Stage 2 patients is available for prediction purposes, but outcome results are not yet available (or if they are they are not used).

When results for both Stage 1 and Stage 2 are available, an overall treatment-control comparison test is conducted over all the patients. In addition a treatment-control comparison is conducted on Stage 2 only. The treatment is considered significant if either test is significant. Type 1 error is controlled by constraining the sum of the two alpha levels for these two tests. Freidlin, Jiang, & Simon (2010 [55]) extend this approach using cross validation to apply the prediction model over all the patients (both Stage 1 and Stage 2). Many classification methods can be employed in this approach. The specific example used in the Freidlin & Simon work was a weighted voting single-gene logistic regression model with treatment main effect and gene expression-treatment interaction. The simulations included scenarios in which the probability of response for sensitive subjects was the same as that for nonsensitive subjects except in the treatment arm. This equates to scenarios involving only treatment- expression interaction, but no main effects for expression or treatment.

Purpose of Study

In Cambon, Baumgartner, Brock, Cooper, & Rai (2015a [19]) and Cambon, Baumgartner, Brock, Cooper, & Rai (2015b, [20]) methods were proposed for modifying or extending nonparametric and parametric classifiers for use in an ASD setting. In this work empirical power of these methods is evaluated under scenarios which include treatment expression interaction

only, expression main effect and treatment interaction, and both treatment and expression main effects together with interaction.

Organization of Study

Section 2 covers the background for the ASD model. Section 3 outlines extensions and modifications of ASD method, and Section 4 describes the simulation study. Results are in Section 5, and discussion is found in Section 6.

2. Background of ASD Model

The sample sizes of subjects for Stage 1 and Stage 2 are n_1 and n_2 respectively, and the total sample size is $n = n_1 + n_2$. Stage 1 serves as the training set where the prediction models is built (using nested cross validation), and Stage 2 patients serve as the final validation set where patient sensitivity is predicted.

A subject that is sensitive to treatment has a greater probability of response on the treatment arm than on the control arm. The probability that any new subject accrued into the study is sensitive is p_S, and is the probability of response given $S_i = s$ and $T_i = t$, where $S_i \in \{0,1\}$ and $T_i \in \{0,1\}$ are random variables for patient sensitivity status and patient treatment arm status respectively, and can take on values of 0 and 1. The number of evaluated genes is *p*. Within these *p* genes it is assumed that there is a set of predictive genes which can be used to predict sensitivity status of subjects. Without loss of generality, it is assumed that the first *m* genes, $k = 1,...,m$, are the predictive genes. The *p*-length vector of gene expression for subject *i* is denoted by random variable $\mathbf{Z}_i$, and the k^{th} element of that vector (gene expression for the k^{th} gene for subject *i*) is Z_{ik}. Fixed realization of these quantities are denoted by lower case letters.

The model relating expected value of gene expression for predictive genes $E(Z_{ik}), k = 1,...,m$, patient sensitivity and treatment arm status to response probability p_{R_i} can be written:

$$\text{logit}(p_{R_{STi}}) = \text{logit}\left(p_{R_i} \mid S_i, T_i, E(\boldsymbol{Z}_i)\right) = \log\left(\frac{p_{R_i} \mid S_i, T_i}{1 - p_{R_i} \mid S_i, T_i}\right) = \beta_0 + \beta_1 T_i$$
$$+\sum_{k=1}^{m}\begin{bmatrix}(1-S_i)\{\beta_{2k}E(Z_{ik} \mid S_i = 0) + \beta_{12k}T_iE(Z_{ik} \mid S_i = 0)\} \\ +S_i\{\beta_{2k}E(Z_{ik} \mid S_i = 1) + \beta_{12k}T_iE(Z_{ik} \mid S_i = 1)\}\end{bmatrix}. \quad (4.1)$$

Now, if expected value of expression for predictive gene depends only on sensitivity status of subject, (i.e.- $E(Z_{ik} \mid S_i = s) = \mu_{sk}$) then subscript *i* can be dropped since probability of response no longer depends on *i*, and for a sensitive subject on the treatment arm:

$$\text{logit}(p_{R_{11}}) = \text{logit}(p_R \mid S = 1, T = 1) = \beta_0 + \beta_1 + \sum_{k=1}^{m}\left(\beta_{2j}\mu_{1k} + \beta_{12k}\mu_{1k}\right),$$

and similarly for $p_{R_{10}}$, p_{R01}, and $p_{R_{00}}$.

A multi-gene logistic regression model can be used in an ASD setting. For example using maximum likelihood estimates from the training set substituted for true unknown parameter values, and using genes selected based on their estimated predictive strength in place of the true unknown predictive genes, and z_{ik} the expression value for subject *i* and selected gene *k* in the validation set in place of μ_{sk} is:

$$\text{logit}(\hat{p}_{R_{ti}}) = \text{logit}(\hat{p}_R \mid t_i, z_{ik*}) = \hat{\beta}_0 + \hat{\beta}_1 t_i + \sum_{k*=1}^{m*}\left(\hat{\beta}_{2k^*}z_{ik^*} + t_i\hat{\beta}_{12k^*}z_{ik^*}\right),$$

where $k* = 1, \ldots, m*$ are the *m** genes predicted to be sensitive (using the DR step), and where information to predict s_i must be derived from the z_{ik*} in place of the μ_{sk}, as well as the estimated model coefficients, and their relationship with t_i. The treatment arm odds ratio for subject *i* is then

$$OR_i = \exp\left[\text{logit}(\hat{p}_{R_{1i}}) - \text{logit}(\hat{p}_{R_{0i}})\right] = \exp\left(\hat{\beta}_1 + \sum_{k^*=1}^{m*}\hat{\beta}_{12k^*}z_{ik^*}\right).$$

As in the Freidlin & Simon, prediction of sensitive patients can also be done using weighted voting on the odds ratios for each of the single gene models for the selected genes. The odds ratios for each of selected gene *k** is then

$$OR_{ik^*} = \exp\left[\text{logit}(\hat{p}_{R_{1ik^*}}) - \text{logit}(\hat{p}_{R_{0ik^*}})\right] = \exp\left(\hat{\beta}_{1k^*} + \hat{\beta}_{12k^*} z_{ik^*}\right).$$

Figure 5 gives more details. Nested validation sets are used for selection of the final tuning parameter set, and the final validation set is used for final prediction of sensitive patients. Note that in the Freidlin & Simon's ASD single gene model, parameter β_2 for gene expression main effect is not included. This is equivalent to modeling no gene expression main effect.

3. Extensions and Modifications of ASD Method

The following classification methods were used in the simulation scenarios. They were outlined in detail in Cambon, Baumgartner, Brock, Cooper, Wu, & Rai (2014a [19]) and in Cambon, Baumgartner, Brock, Cooper, Wu, & Rai (2014b [20]).

1. Same weighted voting logistic regression single gene model as in Freidlin & Simon: no parameter for gene main effect (β_2 constrained to be 0). This is called the LR_{TWV} model.
2. Same model as #1 above, but including both main effects and interaction; i.e. - β_1, β_2 and β_{12}. This is called the LR_{TGWV} model.
3. From Cambon et al. (2014a [19]):
    a. Weighted voting kernel density analysis using posterior odds ratio- which is the odds ratio based on treatment arm and class-specific posterior probabilities. In terms of effects included in model, KDA_{TWV} corresponds to LR_{TWV}, and KDA_{TGWV} corresponds to LR_{TGWV}. Note this since these are single gene weighted voting models, independence is assumed between genes. This is also the case for #4 and #5 below.
4. From Cambon et al. (2014b [20]):

a. Weighted voting quadratic discriminant analysis (QDA_{TWV} and QDA_{TGWV}) using posterior treatment odds ratio.

b. Linear discriminant analysis weighted voting (LDA_{TWV} and LDA_{TGWV}) posterior treatment odds ratio.

The LR_{TWV} weighted voting model described in Freidlin & Simon excludes the term for gene expression main effect. In order to have comparable models for KDA_{TWV}, QDA_{TWV}, and LDA_{TWV}, the class-specific densities for expression on the control arm were constrained to be equal in the TWV models.

4. Simulation Study

The distribution of expression of predictive genes depends on sensitivity status; i.e.- $\mathbf{Z} \mid S = s \sim \text{MVN}(\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_S); \forall i = (1,...,n_1,...,n), k = (1,...,m),$ where Z is the *n* by *p* matrix of gene expression, but it is independent of treatment assignment; the results of the assay are not used to assign subjects to a specific treatment arm.

$\Pr(Z_{ik}, T_i) = \Pr(Z_{ik}) \Pr(T_i),\ i = 1,...,n; k = 1,...,p.$ The distribution of expression of predictive genes for sensitive patients is not independent of response:

$\Pr(Z_{ik}, Y_i) \neq \Pr(Z_{ik}) \Pr(Y_i); \forall i$ s.t. $S_i = 1, k = 1,...,m$ but it is independent of response for nonsensitive patients: $\Pr(Z_{ik}, Y_i) = \Pr(Z_{ik}) \Pr(Y_i); \forall i$ s.t. $S_i = 0, k = 1,...,m.$ For the non-predictive genes, expression is independent of response for all patients; i.e.:

$\Pr(Z_{ik}, Y_i) = \Pr(Z_{ik}) \Pr(Y_i); i = 1,...,n;\ k = m+1,...,p.$ The nonpredictive genes are constrained to have the same mean as the predictive genes for non-sensitive subjects, but they are allowed to have different variance: i.e.: $Z_{ik} \sim N(\mu_0, \sigma_{ns}^{\ 2}), i = 1,...,n, k = m+1,...,p.$

Simulation Steps

1. Fix parameters n_1 and n_2.

a. In practice, methods such as those proposed by Dobbins and Simon (2011 [35]) can be used to apportion n to n_1 and n_2.

2. Allocate Type I error to α_1 and α_2.

    a. $\alpha_1 = 0.04$ and $\alpha_2 = 0.01$ were used exclusively in this work.

3. Select list of tuning parameter sets for ASD design.

    a. Since all methods in this simulation were based on odds ratios similar to those used in Freidlin & Simon, the list of tuning parameter sets are similar to those outlined in that work, however a wider range was included. For example the list consisted of 9 sets of tuning parameters (vs. 3 in Freidlin & Simon) with p-values of 0.01 as well as 0.02 included for the dimension reduction step.

4. Simulate subject sensitivity status over all n subjects using Bernoulli distribution with parameter p_S .

5. Conditional on S_i, $\forall i, i = 1,...,n_1,...,n,$ and $k = 1,...,m,$ simulate gene expression Z_{ik} for predictive genes.

6. Simulate gene expression for non-predictive genes, $k=m+1,\ldots,p$.

7. Divide training set $i = 1,...,n_1$ into nested train and validation sets.

    a. In this work, 10-Fold cross validation (CV) was used; approaches such as nested bootstrap could also be used.

8. Use each set of tuning parameters to predict patient sensitivity on the nested validation sets in the training set $i = 1,...,n_1$;

    a. If CV is used, this results in exactly 1 prediction for each subject in the training set for each of the tuning parameter sets.

    b. It was found that use of low p values (such as 0.01) to select a small number of genes may cause the simulation to fail, since at times no genes will be selected.

To prevent this, code was added to select a minimum of 3 genes. This option permitted a wider choice of tuning parameter *p*-values to select genes.

9. Select the set of tuning parameters that results in the lowest *p*-value of treatment-control comparison of predicted sensitive patients, and this set for prediction of sensitive patients in the validation set.
10. Repeat these steps *B* times for each value of p_S used in the simulation.

    In this work there were *B*= 100 simulations for each scenario in Tables 9 through 13. Tables 14 and 15 used 1000 simulations for each scenario in order to elucidate differences between methods.
11. Conserve Type 1 error α as described in Figure 5 using α_1 for the overall test and α_2 for the subset test.

Simulation Scenarios

The following parameters were common to all simulations:

1) The number of predictive genes p^*=10; total number of genes p=1000.
2) Type 1 error for either the overall test or the subset test was controlled at level 0.05 by setting $\alpha_1 = 0.04$ and $\alpha_2 = 0.01$.
3) The distribution of genes used was the same as in Freidlin, Jiang, & Simon (2010 [55]):
    a. For sensitive patients, expression of predictive genes is normally distributed with mean of 1 and variance of 0.25;
    b. For nonsensitive patients, expression of predictive genes is normally distributed with mean of 0 and variance 0.01;
    c. The expression of non-predictive genes is normally distributed with mean of 0 and variance of 0.25.

The following simulation scenarios were used:

1. Values for p_S of 0.05, 0.07, 0.10, 0.15, 0.20, and 0.25.

2. The combination of values for $p_{R_{ST}}$ used in simulation scenarios are shown for each table of results. These include scenarios involving:
    a. Only gene-treatment interaction (from the Freidlin & Simon 2005 work).
    b. Gene main effect together with gene-treatment interaction.
    c. Both main effects and interaction.
3. The different classification methods described in Section 3 were applied over these different scenarios. Note that the simulation scenarios are distinct from model parameters used in the classification methods to predict patient sensitivity.

5. Results

Tables 9, 10, and 11 include results for gene expression treatment interaction only, with Table 9 showing results with no correlation between genes, Table 10 showing results with correlation of 0.6 between predictive gene and 0.6 between nonpredictive genes, and Table 11 including smaller sample sizes for Stage 1 and Stage 2 (100 patients for each stage instead of 200). Table 12 includes a scenario for gene expression main effect as well as expression-treatment interaction. Table 13 includes a scenario for gene expression main effect, treatment main effect, and gene expression-treatment interaction. Under most scenarios considered, the predictive models which included coefficients main effect terms for treatment only (together with treatment-gene expression interaction – the TWV models) had higher empirical power compared to models which included both expression and treatment main effects together with interaction(all models included treatment interaction term).

In order to compare the methods under different scenarios, additional simulations were conducted. Since empirical power of the different methods was similar, 1000 simulation runs per scenario were used to differentiate between the methods. Table 14 compares empirical power of LR_{TWV}, KDA_{TWV}, LDA_{TWV}, and QDA_{TWV} methods under scenarios similar to those used in the Freidlin and Simon work. That is, the ratio of the variances of gene expression for predictive

genes for sensitive subjects vs. nonsensitive subject was kept at 25 to 1. Under this scenario, empirical power of LR_{TWV} was slightly higher than LDA_{TWV}, which in turn was slightly higher than QDA_{TWV} , which was slightly higher than KDA_{TWV}. Table 15 then shows empirical power of LR_{TWV} and LDA_{TWV} when variances of gene expression for predictive genes for sensitive subjects vs. nonsensitive subject is kept equal (ratio of 1 to 1). Under these scenarios, the empirical power of LDA_{TWV} is slightly higher than that of LR_{TWV}.

6. Discussion

Relationship between Probability of Sensitivity and Power of Adaptive Signature Design

As pointed out in Freidlin & Simon (2005 [54]), when p_S is small, the probability that there is a significant overall treatment effect is small, because the sample size of sensitive subjects is small. In this situation, there still may be a significant treatment effect in the predicted set of sensitive patients if the treatment effect in the subset is large. As p_S increases, the subset of sensitive patients also increases, and if the treatment effect in this subset remains constant, the probability of detecting a significant treatment effect in the subset increases. However at the same time the probability that the overall test detects a significant treatment effect also increases. At some point, the added value of the test of treatment effect in the predicted subset decreases, because it is less likely that both the test over all the patients is insignificant while the subset test is significant. This relationship is reflected in Tables 9 through 11. The "Subset only 0.01" column, which is proportion of times that the subset test was significant when the overall test was not, usually increases as p_S increase from 0.05 to 0.15 (results for $p_S = 0.05$ not shown), and then decreases thereafter. However the overall power of the ASD test, which is significant if either the overall test or the predicted subset test is significant, continues to increase as p_S increases.

Main Effects and Interaction Terms in the ASD Model

In clinical applications, models which include interaction terms usually also include corresponding main effect terms. For example, in a treatment subset prediction context, see Scher, Nasso, & Simon (2011 [144]). Therefore it is interesting that the TWV models had superior performance to TGWV models under the scenarios considered. Also KDA_{TWV} and QDA_{TWV} models had power close to performance of LR_{TWV} and KDA_{TGWV} and QDA_{TGWV} similar to LR_{TGWV}. The performance of the corresponding LDA models were slightly lower, as would be expected since variances of predictive genes for sensitive patients were 25 times the variances of predictive genes for nonsensitive patients.

Limitations

1. In this work the response was assumed to be categorical. However in many clinical trials, response to treatment may be continuous. For example progression free response may be determined from a measurement of tumor size or change in size.
2. In some works clinical covariates or gene expression results may be used to assign patients to a specific treatment regimen; for example see Lu, Zhang, & Zeng (2013 [107]). However in this work we limit ourselves to the situation in which the distributions of expression of predictive genes and non-predictive genes are independent of treatment; i.e. patients are randomized independently of results of assay.
3. Real data sets were not used to simulate gene expression. Further work needs to be done examining the performance of the various methods with real data sets.
4. The type 1 error can be split up between the overall test and the subset test in different ways. For example, if it is hypothesized that is only a very small (or even no) overall treatment effect, but a large treatment effect for a subset of patients, most of the Type 1 error can go to the predicted subset. This was the approach used in Scher et al., (2011 [144]).

Future Work

The KNN (K Nearest Neighbors) using in Dudoit et al. (2002 [39]) uses 1 – correlation as a distance measure to determine the nearest neighbors. The decision rule to determine class is $\hat{g}_i = I(p_{i,NN} > 0.5)$ where $\hat{g}_i$ is the 0-1 classification decision for subject *i*, and $p_{i,NN}$ is the proportion of training set nearest neighbors for patient *i* belonging to class g=1. This approach performed well in the Dudoit et al. work (2002 [39]). One approach to apply KNN to treatment subset prediction would be to compare this difference in proportion across the two treatment groups. Subjects which have a large difference in proportions would be predicted to be sensitive (Cambon et al., 2014a [19]). However this difference in proportions does not make direct use of the underlying distance measure (which in the Dudoit et al. work is 1 – correlation) for treatment subset prediction. Work is underway in developing this method for treatment subset prediction. Work is also underway to develop code for multi-gene penalized versions of LDA such as Modified Linear Discriminant Analysis or MLDA (Xu, Brock, & Parrish, 2009 [177]). Use of multi-gene penalized logistic regression models is another appealing avenue for exploration.

Table 9: PR_{11}=0.98, PR_{10}=PR_{01}=PR_{00}=0.25. (Expression-Treatment Interaction Effect Only); $\rho = 0.$

p_S	Method	Subset .01	Subset Only .01	ASD.overall
0.10	LR_{TWV}	0.68	0.46	0.82
	KDA_{TWV}	0.62	0.39	0.75
	QDA_{TWV}	0.61	0.40	0.76
	LDA_{TWV}	0.67	0.44	0.80
	LR_{TGWV}	0.57	0.38	0.74
	KDA_{TGWV}	0.59	0.40	0.76
	QDA_{TGWV}	0.56	0.38	0.74
	LDA_{TGWV}	0.52	0.35	0.71
0.15	LR_{TWV}	0.89	0.35	0.93
	KDA_{TWV}	0.84	0.32	0.90
	QDA_{TWV}	0.84	0.34	0.92
	LDA_{TWV}	0.85	0.33	0.91
	LR_{TGWV}	0.79	0.31	0.89
	KDA_{TGWV}	0.81	0.33	0.91
	QDA_{TGWV}	0.81	0.32	0.90
	LDA_{TGWV}	0.77	0.32	0.90
0.20	LR_{TWV}	0.99	0.24	1.00
	KDA_{TWV}	0.95	0.22	0.98
	QDA_{TWV}	0.97	0.24	1.00
	LDA_{TWV}	0.99	0.24	1.00
	LR_{TGWV}	0.92	0.23	0.99
	KDA_{TGWV}	0.96	0.24	1.00
	QDA_{TGWV}	0.94	0.23	0.99
	LDA_{TGWV}	0.91	0.21	0.97

p_S - probability any patient is sensitive; PR_{ST}-Probability of response given S=s and T=t, where S and T are Sensitivity and Treatment indicators; Subset.01-subset test at level .01; Subset Only .01- Only subset test significant (Overall.04 test not significant); ASD overall - overall empirical power of ASD- either Overall .04 test significant or Subset .01 test significant; Overall .04 test empirical power at Type 1 error 0.04: 0.22, 0.36, 0.58, 0.76 – for p_S =0.07, 0.10,0.15, and 0.20;LR-Logistic Regression; TWV-weighted voting model with only treatment main effect and expression-treatment interaction; TGWV-weighted voting model with expression and treatment main effects and interaction. Mean and variance of expression for predictive genes for sensitive patients =1 and 0.25 respectively; mean and variance for predictive genes for nonsensitive patients =0 and 0.01 respectively. Mean and variance for nonpredictive genes= 0 and 0.25. ρ - correlation within block of predictive genes and within block of nonpredicitve genes. Correlation between the two blocks =0 for all simulations.

Table 10: PR_{11}=0.98, PR_{10}=PR_{01}=PR_{00}=0.25. (Expression-Treatment Interaction Only); $\rho = 0.6$.

p_S	Method	Subset .01	Subset Only .01	ASD.overall
0.10	LR_{TWV}	0.59	0.38	0.74
	KDA_{TWV}	0.59	0.41	0.77
	QDA_{TWV}	0.62	0.43	0.79
	LDA_{TWV}	0.53	0.33	0.69
	LR_{TGWV}	0.49	0.31	0.67
	KDA_{TGWV}	0.52	0.36	0.72
	QDA_{TGWV}	0.48	0.31	0.67
	LDA_{TGWV}	0.47	0.30	0.66
0.15	LR_{TWV}	0.85	0.33	0.91
	KDA_{TWV}	0.84	0.34	0.92
	QDA_{TWV}	0.81	0.32	0.90
	LDA_{TWV}	0.80	0.29	0.87
	LR_{TGWV}	0.75	0.26	0.84
	KDA_{TGWV}	0.68	0.26	0.84
	QDA_{TGWV}	0.73	0.27	0.85
	LDA_{TGWV}	0.64	0.20	0.78
0.20	LR_{TWV}	0.95	0.21	0.97
	KDA_{TWV}	0.91	0.21	0.97
	QDA_{TWV}	0.92	0.22	0.98
	LDA_{TWV}	0.95	0.21	0.97
	LR_{TGWV}	0.90	0.21	0.97
	KDA_{TGWV}	0.87	0.20	0.96
	QDA_{TGWV}	0.88	0.21	0.97
	LDA_{TGWV}	0.86	0.20	0.96

See Table 9 for notation/abbreviations. Overall empirical power at Type 1 error 0.04: 0.22, 0.36, 0.58, 0.76 – for p_S =0.07, 0.10, 0.15, and 0.20.

Table 11: Small Sample Size Simulation Scenario PR_{11}=0.98, PR_{10}=0.25, PR_{01}=PR_{00}=0.25; n_1=n_2=100, $\rho = 0$.

p_S	Method	Subset .01	Subset Only .01	ASD.overall
0.10	LR_{TWV}	0.08	0.08	0.24
	KDA_{TWV}	0.05	0.05	0.21
	LDA_{TWV}	0.08	0.07	0.23
	QDA_{TWV}	0.08	0.07	0.23
0.15	LR_{TWV}	0.22	0.12	0.43
	KDA_{TWV}	0.18	0.09	0.40
	LDA_{TWV}	0.19	0.12	0.43
	QDA_{TWV}	0.19	0.13	0.44
0.20	LR_{TWV}	0.50	0.14	0.76
	KDA_{TWV}	0.45	0.09	0.71
	LDA_{TWV}	0.44	0.09	0.71
	QDA_{TWV}	0.50	0.11	0.73
0.25	LR_{TWV}	0.72	0.13	0.91
	KDA_{TWV}	0.65	0.11	0.89
	LDA_{TWV}	0.71	0.13	0.91
	QDA_{TWV}	0.67	0.13	0.91

See Table 9 for notation/abbreviations. Overall power at 0.04: 0.16,0.31,0.62, and 0.78 for p_S=0.10,0.15, 0.20,.25.

Table 12: Gene Main Effect, and Gene-Treatment Interaction:PR_{11}=0.98, PR_{10}=0.35, PR_{01}=PR_{00}=0.25, $\rho = 0$.

p_S	Method	Subset .01	Subset Only .01	ASD.overall
0.07	LR_{TWV}	0.10	0.07	0.26
	KDA_{TWV}	0.12	0.07	0.26
	QDA_{TWV}	0.11	0.08	0.27
	LDA_{TWV}	0.08	0.05	0.24
	LR_{TGWV}	0.07	0.06	0.25
	KDA_{TGWV}	0.09	0.07	0.26
	QDA_{TGWV}	0.08	0.05	0.24
	LDA_{TGWV}	0.07	0.06	0.25
0.10	LR_{TWV}	0.43	0.35	0.57
	KDA_{TWV}	0.38	0.32	0.54
	QDA_{TWV}	0.44	0.36	0.58
	LDA_{TWV}	0.41	0.34	0.56
	LR_{TGWV}	0.36	0.30	0.52
	KDA_{TGWV}	0.33	0.27	0.49
	QDA_{TGWV}	0.40	0.32	0.54
	LDA_{TGWV}	0.29	0.25	0.47
0.15	LR_{TWV}	0.77	0.39	0.85
	KDA_{TWV}	0.68	0.34	0.80
	QDA_{TWV}	0.72	0.38	0.84
	LDA_{TWV}	0.76	0.39	0.85
	LR_{TGWV}	0.66	0.36	0.82
	KDA_{TGWV}	0.65	0.36	0.82
	QDA_{TGWV}	0.61	0.30	0.76
	LDA_{TGWV}	0.62	0.30	0.76
0.20	LR_{TWV}	0.90	0.33	0.94
	KDA_{TWV}	0.86	0.28	0.89
	QDA_{TWV}	0.89	0.32	0.93
	LDA_{TWV}	0.85	0.29	0.90
	LR_{TGWV}	0.79	0.26	0.87
	KDA_{TGWV}	0.79	0.27	0.88
	QDA_{TGWV}	0.81	0.29	0.90
	LDA_{TGWV}	0.76	0.26	0.87

See Table 9 for notation/abbreviations. Overall power at 0.04: 00.22,0.36,0.58, and 0.76 for p_S =0.07, 0.10,0.15, and 0.20.

Table 13: Sensitivity Main Effect, and Treatment Main Effect, and Sensitivity-Treatment Interaction: PR_{11}=0.98, PR_{10}=0.35, PR_{01}=0.35, PR_{00}=0.25, $\rho = 0$.

p_S	Method	Subset .01	Subset Only .01	ASD.overall
0.07	LR_{TWV}	0.07	0.00	0.78
	KDA_{TWV}	0.10	0.01	0.79
	QDA_{TWV}	0.06	0.01	0.79
	LDA_{TWV}	0.05	0.00	0.78
	LR_{TGWV}	0.10	0.00	0.78
	KDA_{TGWV}	0.11	0.00	0.78
0.10	LR_{TWV}	0.35	0.03	0.94
	KDA_{TWV}	0.30	0.02	0.93
	QDA_{TWV}	0.33	0.01	0.92
	LDA_{TWV}	0.26	0.03	0.94
	LR_{TGWV}	0.27	0.01	0.92
	KDA_{TGWV}	0.27	0.02	0.93
0.15	LR_{TWV}	0.67	0.01	1.00
	KDA_{TWV}	0.55	0.01	1.00
	QDA_{TWV}	0.58	0.01	1.00
	LDA_{TWV}	0.60	0.01	1.00
	LR_{TGWV}	0.46	0.01	1.00
	KDA_{TGWV}	0.51	0.01	1.00
0.20	LR_{TWV}	0.76	0.02	0.99
	KDA_{TWV}	0.76	0.02	0.99
	QDA_{TWV}	0.76	0.03	1.00
	LDA_{TWV}	0.73	0.01	0.98
	LR_{TGWV}	0.68	0.02	0.99
	KDA_{TGWV}	0.72	0.02	0.99

See Table 9 for notation/abbreviations. Overall power at 0.04: 0.78,0.91,0.99, and 0.97 for p_S =0.07, 0.10,0.15, and 0.20.

Table 14: Comparison of TWV Methods under simulation scenarios similar to Freidlin and Simon work (unequal variances for gene expression for predictive genes between sensitivity classes). PR_{11}=0.98, PR_{10}=0.25, PR_{01}=0.25, PR_{00}=0.25, $\rho = 0$.

Method	Ps	Overall.04	Subset.01	Subset Only .01	ASD Overall
LR_{TWV}	0.10	0.35	0.63	0.41	0.75
KDA_{TWV}	0.10	0.35	0.57	0.36	0.70
LDA_{TWV}	0.10	0.35	0.61	0.40	0.74
QDA_{TWV}	0.10	0.35	0.59	0.38	0.72
LR_{TWV}	0.15	0.50	0.78	0.36	0.86
KDA_{TWV}	0.15	0.50	0.73	0.32	0.83
LDA_{TWV}	0.15	0.50	0.76	0.35	0.86
QDA_{TWV}	0.15	0.50	0.74	0.34	0.84
LR_{TWV}	0.20	0.64	0.85	0.28	0.91
KDA_{TWV}	0.20	0.64	0.82	0.25	0.89
LDA_{TWV}	0.20	0.64	0.84	0.27	0.91
QDA_{TWV}	0.20	0.64	0.83	0.26	0.90

See Table 9 for notation/abbreviations.

Table 15: Comparison of LR_{TWV} and LDA_{TWV} methods when variances of gene expression for predictive genes are constrained to be equal PR_{10}=0.25, PR_{01}=0.25, PR_{00}=0.25, $\rho = 0$.

Method	Stdev	PR_{11}	Overall 0.04	Subset.01	Subset .01 Only	ASD Overall
LDA_{TWV}	0.40	0.90	0.53	0.24	0.08	0.61
LDA_{TWV}	0.35	0.90	0.53	0.28	0.10	0.63
LDA_{TWV}	0.30	0.90	0.54	0.34	0.12	0.66
LRT_{TWV}	0.40	0.90	0.53	0.21	0.06	0.59
LRT_{TWV}	0.35	0.90	0.53	0.26	0.08	0.61
LRT_{TWV}	0.30	0.90	0.54	0.31	0.10	0.63
LDA_{TWV}	0.40	0.98	0.62	0.48	0.16	0.77
LDA_{TWV}	0.35	0.98	0.63	0.56	0.17	0.80
LDA_{TWV}	0.30	0.98	0.63	0.64	0.20	0.83
LRT_{TWV}	0.40	0.98	0.62	0.45	0.14	0.76
LRT_{TWV}	0.35	0.98	0.63	0.54	0.16	0.79
LRT_{TWV}	0.30	0.98	0.63	0.62	0.18	0.82
LDA_{TWV}	0.40	0.98	0.49	0.27	0.11	0.597
LDA_{TWV}	0.35	0.98	0.50	0.35	0.14	0.633
LDA_{TWV}	0.30	0.98	0.50	0.43	0.18	0.677
LRT_{TWV}	0.40	0.98	0.49	0.28	0.10	0.595
LRT_{TWV}	0.35	0.98	0.50	0.33	0.13	0.626
LRT_{TWV}	0.30	0.98	0.50	0.41	0.16	0.664

Stdev- standard deviation for gene expression for predictive genes and nonpredictive genes; difference between means of gene expression for nonsensitive vs sensitive subjects = 1. See Table 9 for other notation/abbreviations.

REFERENCES

[1] Altman DG, Vergouwe Y, Royston P, and Moons KGM. Prognosis and prognostic research: validating a prognostic model. *BMJ* 338: 2009.

[2] Anderson JA. Separate Sample Logistic Discrimination. *Biometrika* 59: 19-35, 1972.

[3] Argenziano G, Fabbrocini G, Carli P, De Giorgi V, Sammarco E, and Delfino M. Epiluminescence microscopy for the diagnosis of doubtful melanocytci skin lesions: comparison of the ACD rule of dermascopy and a new 7-point checklist based on pattern analysis. *Arch Dermatol* 134: 1363-1370, 1998.

[4] Balch CM, Soong S-J, Atkins MB, Buzaid AC, Cascinelli N, Coit DG, Fleming ID, Gershenwald JE, Houghton A, Kirkwood JM, McMasters KM, Mihm MF, Morton DL, Reintgen DS, Ross MI, Sober A, Thompson JA, and Thompson JF. An evidence-based staging system for cutaneous melanoma. *CA: A Cancer Journal for Clinicians* 54: 131-149, 2004.

[5] Balch CM, Gershenwald JE, Soong S-j, Thompson JF, Atkins MB, Byrd DR, Buzaid AC, Cochran AJ, Coit DG, Ding S, Eggermont AM, Flaherty KT, Gimotty PA, Kirkwood JM, McMasters KM, Mihm MC, Morton DL, Ross MI, Sober AJ, and Sondak VK. Final version of 2009 AJCC melanoma staging and classification. *Journal of Clinical Oncology* 27: 6199-6206, 2009.

[6] Balch CM, Gershenwald JE, Soong S-j, Thompson JF, Ding S, Byrd DR, Cascinelli N, Cochran AJ, Coit DG, Eggermont AM, Johnson T, Kirkwood JM, Leong SP, McMasters KM, Mihm MC, Morton DL, Ross MI, and Sondak VK. Multivariate analysis of prognostic factors among 2,313 patients with stage III melanoma: comparison of nodal micrometastases versus macrometastases. *Journal of Clinical Oncology* 28: 2452-2459, 2010.

[7] Belhumeur PN, Hespanha JP, and Kriegman D. Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 19: 711-720, 1997.

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

[9] Bickel PJ, and Levina E. Some theory for Fisher's linear discriminant function, 'Naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli* 10: 989-1010, 2004.

[10] Bishop C. *Pattern Recognition and Machine Learning*. Springer, 2006.

[11] Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben-Dor A, Sampas N, Dougherty E, Wang E, FMarincola F, Gooden C, Lueders J, Glatfelter A, Pollock P, Carpten J, Gillanders E, Leja D, Dietrich K, Beaudry C, Berens M, Alberts D, Sondak V, Hayward N, and Trent J. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* 406: 536-540, 2000.

[12] Boser BE, Guyon IM, and Vapnik VN. A training algorithm for optimal margin classifier. In *Proc 5th ACM Workshop on Computational Learning Theory* 144-152, 1992.

[13] Boulesteix A-L, and Strimmer K. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics* 8: 32-44, 2007.

[14] Box GEP, and Cox DR. An Analysis of Transformations. *Journal of the Royal Statistical Society Series B (Methodological)* 26: 211-252, 1964.

[15] Breiman L, Friedman J, Olshen RA, and Stone CJ. *Classification and Regression Trees*. Chapman and Hall, 1984.
[16] Breiman L. Bagging predictors. *Machine Learning* 24: 123-140, 1996.
[17] Breiman L. Random forests-random features. Berkeley: Department of Statistics,University of California, CA, 1999.
[18] Breiman L. Random forests. *Machine Learning* 45: 5-32, 2001.
[19] Cambon A, Baumgartner KB, Brock GN, Cooper NGF, Wu D, and Rai SN. Classification of Clinical Outcomes Using High-throughput Informatics: Part 1- Nonparametric Method Reviews. *Model Assisted Statistics and Applications* (In Press): 2015.
[20] Cambon A, Baumgartner KB, Brock GN, Cooper NGF, Wu D, and Rai SN. Classification of Clinical Outcomes Using High-throughput Informatics: Part 2- Parametric Method Reviews. *Model Assisted Statistics and Applications* (Accepted): 2015.
[21] Carpenter J, and Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicine* 19: 1141-1164, 2000.
[22] Chan K-Y, and Loh W-Y. LOTUS: an algorithm for building accurate and comprehensible logistic regression trees. *Journal of Computational and Graphical Statistics* 13: 826-852, 2004.
[23] Chang HY, Nuyten DSA, Sneddon JB, Hastie T, Tibshirani R, Sørlie T, Dai H, He YD, van't Veer LJ, Bartelink H, van de Rijn M, Brown PO, and van de Vijver MJ. Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proceedings of the National Academy of Sciences of the United States of America* 102: 3738-3743, 2005.
[24] Chapman PB, Hauschild A, Robert C, Haanen JB, Ascierto P, Larkin J, Dummer R, Garbe C, Testori A, Maio M, Hogg D, Lorigan P, Lebbe C, Jouary T, Schadendorf D, Ribas A, O'Day SJ, Sosman JA, Kirkwood JM, Eggermont AMM, Dreno B, Nolop K, Li J, Nelson B, Hou J, Lee RJ, Flaherty KT, McArthur, and A. G. Improved survival with Vemurafenib in melanoma with BRAF V600E mutation. *New England Journal of Medicine* 364: 2507-2516, 2011.
[25] Chen S-H, Sun J, Dimitrov L, Turner AR, Adams TS, Meyers DA, Chang B-L, Zheng SL, Grönberg H, Xu J, and Hsu F-C. A support vector machine approach for detecting gene-gene interaction. *Genetic Epidemiology* 32: 152-167, 2008.
[26] Chow S-C, Shao J, and Wang H. A note on sample size calculation for mean comparisons based on noncentral t-statistics. *Journal of Biopharmaceutical Statistics* 12: 441-456, 2002.
[27] Chun H, and Keleş S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72: 3-25, 2010.
[28] Claridge E, Hall PN, Keefe M, and Allen JP. Shape analysis for classification of malignant melanoma. *Journal of Biomedical Engineering* 14: 229-234, 1992.
[29] Conover WJ, and Iman RL. Rank Transformations as a Bridge Between Parametric and Nonparametric Statistics. *The American Statistician* 35: 124-129, 1981.
[30] Cordell HJ. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet* 10: 392-404, 2009.
[31] Cox DR. Some procedures connnected with the logistic quality response curve. *Research papers in statistics* 55-71, 1966.
[32] Cristianini N, and Shawe-Taylor J. *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.
[33] Cruz JA, and Wishart DS. Applications of machine learning in cancer prediction and prognosis. *Cancer Informatics* 2: 2006.

[34] Dempster AP, Laird NM, and Rubin DB. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society Series B (Methodological)* 39: 1-38, 1977.
[35] Dobbin K, and Simon R. Optimally splitting cases for training and testing high dimensional classifiers. *BMC Medical Genomics* 4: 31, 2011.
[36] Dobbin KK, and Simon R. Sample size planning for developing classifiers using high dimensional DNA microarray data. *Biostatistics* 8: 101-117, 2007.
[37] Dobson AJ, and Gebski VJ. Sample Sizes for Comparing Two Independent Proportions Using the Continuity-Corrected Arc Sine Transformation. *Journal of the Royal Statistical Society Series D (The Statistician)* 35: 51-53, 1986.
[38] Duda R, Hart P, and Stork DG. *Pattern Classification*. John Wiley & Sons, Inc, 2001.
[39] Dudoit S, Fridlyand J, and Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. *JASA* 2002.
[40] Duncan LM. The classification of cutaneous melanoma. *Hematology/Oncology Clinics of North America* 23: 501-513, 2009.
[41] Duong T. ks: kernel density estimation and dernel discriminant analysis for multivariate data in R. *Journal of Statistical Software* 21: 1-16, 2007.
[42] Efron B, and Tibshirani RJ. *An Introduction to the Bootstrap*. Chapman & Hall/CRC 1994.
[43] Efron B, and Tibshirani R. On testing the significance of sets of genes. Standford: Standford University, 2006.
[44] Elith J, Leathwick JR, and Hastie T. A working guide to boosted regression trees. *Journal of Animal Ecology* 77: 802-813, 2008.
[45] Ercal F, Chawla A, Stoecker WV, Lee H-C, and Moss RH. Neural network diagnosis of malignant melanoma from color images. *Biomedical Engineering, IEEE Transactions* 41: 837-845, 1994.
[46] Fan J, and Fan Y. High-dimensional classification using features annealed independence rules. *Annals of Statistics* 6: 2605-2637, 2008.
[47] Fan J, and Lv J. A selective overview of variable selection in high dimensional feature space. *Stat Sin* 10: 101-148, 2010.
[48] Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7 179-188, 1936.
[49] Fix E, and Hodges JLJ. Discriminatory analysis. Nonparametric discrimination: consistency properties. *International Statistical Review / Revue Internationale de Statistique* 57: 238-247, 1951.
[50] Florek K, Lukaszewicz J, Perkal J, and Zubrzycki S. Sur la Liaison et la Division des Points d'un Ensemble Fini. *Colloquium Mathematicae* 2: 282-285, 1951.
[51] Foster JC, Taylor JMG, and Ruberg SJ. Subgroup identification from randomized clinical trial data. *Statistics in Medicine* 30: 2867-2880, 2011.
[52] Fred ALN, and Jain AK. Data clustering using evidence accumulation. In: *Pattern Recognition, 2002 Proceedings 16th International Conference on*2002, p. 276-280 vol.274.
[53] Freidlin B, Zheng G, Li Z, and Gastwirth J. Trend tests for case-control studies of genetic markers: power, sample size and robustness. *Human Heredity* 53: 146-152, 2002.
[54] Freidlin B, and Simon R. Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clinical Cancer Research* 11: 7872-7878, 2005.
[55] Freidlin B, Jiang W, and Simon R. The cross-validated adaptive signature design. *Clinical Cancer Research* 16: 691-698, 2010.

[56] Freund Y, and Schapire R. A desicion-theoretic generalization of on-line learning and an application to boosting

In: *Computational Learning Theory*, edited by Vitányi PSpringer Berlin / Heidelberg, 1995, p. 23-37.
[57] Freund Y, and Schachter J. Experiments with a new boosting algorithm. *Machine Learning* 1996.
[58] Friedman JH. Regularized Discriminant Analysis. *Journal of the American Statistical Association* 84: 165-175, 1989.
[59] Friedman JH. Multivariate Adaptive Regression Splines. *The Annals of Statistics* 19: 1-67, 1991.
[60] Friedman JH. Stochastic gradient boosting. *Computational Statistics & Data Analysis* 38: 367-378, 2002.
[61] Friedman RJ, Rigel DS, and Kopf AW. Early detection of malignant melanoma: the role of physician examination and self-examination of the skin. *CA: A Cancer Journal for Clinicians* 35: 130-151, 1985.
[62] Gan G, Ma C, and Wu J. *Data Clustering Theory, Algorithms, and Applications*. ASA-SIAM, 2007.
[63] Geladi P, and Kowalski BR. Partial least-squares regression: a tutorial. *Analytica Chimica Acta* 185: 1-17, 1986.
[64] Gini C. *Variabilità e mutabilità* Bologna: C. Cuppini, 1912.
[65] Ginsberg ML. Counterfactuals. *Artificial Intelligence* 30: 35-79, 1986.
[66] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, and Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286: 531-537, 1999.
[67] Gower JC. A Comparison of Some Methods of Cluster Analysis. *Biometrics* 23: 623-637, 1967.
[68] Guo Y, Hastie T, and Tibshirani R. Regularized linear discriminant analysis and its application in microarrays. *Biostatistics* 8: 86-100, 2007.
[69] Guyon. An introduction to variable and feature selection. *Journal of Machine Learning Research 3 (2003)* 3: 1157-1182, 2003.
[70] Hall M. Correlation-based feature selection for machine learning. In: *Department of Computer Science*. New Zealand: Waikato University, 1999.
[71] Hall P, and Wand MP. On nonparametric discrimination using density differences. *Biometrika* 75: 541-547, 1988.
[72] Harrison DA, and Brady AR. Sample size and power calculations using the noncentral t-distribution. *The Stata Journal* 4: 142-153, 2004.
[73] Hastie T, and Tibshirani R. Generalized Additive Models: Some Applications. *Journal of the American Statistical Association* 82: 371-386, 1987.
[74] Hastie T, Buja A, and Tibshirani R. Penalized Discriminant Analysis. *The Annals of Statistics* 23: 73-102, 1995.
[75] Hastie T, Tibshirani R, and Friedman J. *The Elements of Statistical Learning*. Springer, 2009.
[76] Heinze G, and Schemper M. A solution to the problem of separation in logistic regression. *Statistics in Medicine* 21: 2409-2419, 2002.
[77] Henning JS, Dusza SW, Wang SQ, Marghoob AA, Rabinovitz HS, Polsky D, and Kopf AW. The CASH (color, architecture, symmetry, and homogeneity) algorithm for dermoscopy. *Journal of the American Academy of Dermatology* 56: 45-52, 2007.

[78] Hills M. Allocation Rules and their Error Rates. *Journal of the Royal Statistical Society Series B (Methodological)* 28: 1-31, 1966.
[79] Hodi FS, O'Day SJ, McDermott DF, Weber RW, Sosman JA, Haanen JB, Gonzalez R, Robert C, Schadendorf D, Hassel JC, Akerley W, van den Eertwegh AJM, Lutzky J, Lorigan P, Vaubel JM, Linette GP, Hogg D, Ottensmeier CH, Lebbé C, Peschel C, Quirt I, Clark JI, Wolchok JD, Weber JS, Tian J, Yellin MJ, Nichol GM, Hoos A, and Urba WJ. Improved survival with Ipilimumab in patients with metastatic melanoma. *New England Journal of Medicine* 363: 711-723, 2010.
[80] Hoerl AE, and Kennard RW. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* 12: 55-67, 1970.
[81] Hofling H, and Tibshirani R. A study of prevalidation. *The Annals of Applied Statistics* 2: 643–664, 2008.
[82] Holland J. *Adaptation in Natural and Artificial Systems*. Ann Arbor: University of Michigan Press, 1975.
[83] Hotelling H. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* 24: 417-441, 1933.
[84] Hothorn T, and Lausen B. Bundling classifiers by bagging trees. *Computational Statistics & Data Analysis* 49: 1068-1078, 2005.
[85] Huang DW, Sherman BT, and Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protocols* 4: 44-57, 2008.
[86] Huang J, Lin A, Narasimhan B, Quertermous T, Hsiung CA, Ho L-T, Grove JS, Olivier M, Ranade K, Risch NJ, and Olshen RA. Tree-structured supervised learning and the genetics of hypertension. *Proceedings of the National Academy of Sciences of the United States of America* 101: 10529-10534, 2004.
[87] Irvin Jr WJ, and Carey LA. What is triple-negative breast cancer? *European Journal of Cancer* 44: 2799-2805, 2008.
[88] Isakoff SJ. Triple-Negative Breast Cancer: Role of Specific Chemotherapy Agents. *The Cancer Journal* 16: 53-61 2010.
[89] Jacob L, Obozinski G, and Vert J-P. Group lasso with overlap and graph lasso. In: *Proceedings of the 26th Annual International Conference on Machine Learning*. Montreal, Quebec, Canada: ACM, 2009, p. 433-440.
[90] Jain AK, and Chandrasekaran B. Dimensionality and Sample Size Considerations in Pattern Recognition Practice. In: *Handbook of Statistics*, edited by Krishnaiah PR, and Kanal LNNorth Holland, 1982, p. 835-855.
[91] Jeffery I, Higgins D, and Culhane A. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics* 7: 359, 2006.
[92] Johnson R, and Wichern D. *Applied Multivariate Statistical Analysis*. Pearson 2007.
[93] Kim J, and Scott C. L2 kernel classification. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 32: 1822-1831, 2010.
[94] Kittler J. Feature set search algorithms. In: *Pattern Recognition and Signal Processing*, edited by Chen CSijthoff and Noordhoff, Alphen aan den Rijn,, 1978.
[95] Koenker R, and Bassett G, Jr. Regression Quantiles. *Econometrica* 46: 33-50, 1978.
[96] Kosko B. *Fuzzy thinking: The new science of fuzzy logic*. New York: Hyperion, 1993.
[97] Lachenbruch PA. On expected probabilities of misclassification in discriminant analysis, necessary sample size, and a relation with the multiple correlation coefficient. *Biometrics* 24: 823-834, 1968.
[98] Leblanc M, and Tibshirani R. Combining estimates in regression and classification,. *Journal of the American Statistical Association* 91: 1641–1650, 1996.

[99] Ledoit O, and Wolf M. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis* 88: 365-411, 2004.
[100] Lee TK, and Claridge E. Predictive power of irregular border shapes for malignant melanomas. *Skin Research & Technology* 11: 1-8, 2005.
[101] Lipkovich I, Dmitrienko A, Denne J, and Enas G. Subgroup identification based on differential effect search—A recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in Medicine* 30: 2601-2621, 2011.
[102] Liu H, Li J, and Wong L. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome Informatics* 13: 51-60, 2002.
[103] Lloyd S. Least squares quantization in PCM. *Information Theory, IEEE Transactions on* 28: 129-137, 1982.
[104] Loh W-Y. Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica* 12: 361-386, 2002.
[105] Loh W-Y. Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1: 14-23, 2011.
[106] Lou X-Y, Chen G-B, Yan L, Ma JZ, Zhu J, Elston RC, and Li MD. A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. *The American Journal of Human Genetics* 80: 1125-1137, 2007.
[107] Lu W, Zhang HH, and Zeng D. Variable selection for optimal treatment decision. *Statistical Methods in Medical Research* 22: 493-504, 2013.
[108] Lunetta K, Hayward LB, Segal J, and Van Eerdewegh P. Screening large-scale association study data: exploiting interactions using random forests. *BMC Genetics* 5: 32, 2004.
[109] MacKie RM. *Illustrated Guide to the Recognition of Early Malignant Melanoma*. Glasgow University Department of Dermatology, 1985.
[110] Maglietta R, Distaso A, Piepoli A, Palumbo O, Carella M, D'Addabbo A, Mukherjee S, and Ancona N. On the reproducibility of results of pathway analysis in genome-wide expression studies of colorectal cancers. *Journal of Biomedical Informatics* 43: 397-406, 2010.
[111] Mahalanobis P. On the generalized distance in statistics. *Proceedings of the National Institute of Science of India* 2: 1936.
[112] Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, and Wingender E. TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes. *Nucleic Acids Research* 34: D108-D110, 2006.
[113] McCullagh P, and Nelder JA. *Generalized Linear Models*. London: Chapman & Hall., 1989.
[114] Monheit G, Cognetta AB, Ferris L, Rabinovitz H, Gross K, Martini M, Grichnik JM, Mihm M, Prieto VG, Googe P, King R, Toledano A, Kabelev N, Wojton M, and Gutkowicz-Krusin D. The performance of melafind: a prospective multicenter study. *Arch Dermatol* 147: 188-194, 2011.
[115] Moore JH, Asselbergs FW, and Williams SM. Bioinformatics challenges for genome-wide association studies. *Bioinformatics* 26: 445-455, 2010.
[116] Neal R, and Zhang J. High dimensional classification with bayesian neural networks and dirichlet diffusion trees,. In: *Feature Extraction, Foundations and Applications*, edited by Guyon I, Gunn S, Nikravesh M, and Zadeh L. New York: Springer, 2006, p. 265-296.
[117] Neyman J, and Pearson ES. On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society of London Series A, Containing Papers of a Mathematical or Physical Character* 231: 289-337, 1933.
[118] Nguyen DV, and Rocke DM. Multi-class cancer classification via partial least squares with gene expression profiles. *Bioinformatics* 18: 1216-1226, 2002.

[119] Page L, Brin S, Motwani R, and Winograd T. The PageRank citation ranking: bringing order to the Web 1999.
[120] Park MY, and Hastie T. Penalized logistic regression for detecting gene interactions. *Biostatistics* 9: 30-50, 2008.
[121] Parrish RS, Spencer HJ, and Xu P. Distribution modeling and simulation of gene expression data. *Computational Statistics and Data Analysis* 53: 53 (55), 1650–1660, 2009.
[122] Pattengale N, Alipour M, Bininda-Emonds O, Moret B, and Stamatakis A. How many bootstrap replicates are necessary? *Journal of Computational Biology* 17: 184-200, 2010.
[123] Pearson K. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London A* 185: 71-110, 1894.
[124] Pearson K. On lines and planes of closest fit to systems of points in space. *Philisophical Magazine* 2: 559-572, 1901.
[125] Peters A, and Hothorn T. ipred: Improved Predictors. R package version 0.9-1. 2012.
[126] Pittman J, Huang E, Dressman H, Horng C-F, Cheng SH, Tsou M-H, Chen C-M, Bild A, Iversen ES, Huang AT, Nevins JR, and West M. Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. *Proceedings of the National Academy of Sciences of the United States of America* 101: 8431-8436, 2004.
[127] Press SJ, and Wilson S. Choosing Between Logistic Regression and Discriminant Analysis. *Journal of the American Statistical Association* 73: 699-705, 1978.
[128] Quenouille MH. Approximate Tests of Correlation in Time-Series. *Journal of the Royal Statistical Society Series B (Methodological)* 11: 68-84, 1949.
[129] Rai S, Pan J, Cambon A, Gargett N, and Chaires JB. Group classification based on high-dimensional data: application to differential scanning calorimetry plasma thermogram analysis of cervical cancer and control samples. *Open Access Medical Statistics* 3: 1-9, 2013.
[130] Rao CR. The problem of classification and distance between two populations. *Nature* 159: 30-31, 1947.
[131] Rencher AC. *Methods of Multivariate Analysis*. Wiley, 1995.
[132] Richardson MW. Multidimensional psychophysics. *Psychological Bulletin* 35: 659-660, 1938.
[133] Rigel DS, Russak J, and Friedman R. The evolution of melanoma diagnosis: 25 years beyond the ABCDs. *CA: A Cancer Journal for Clinicians* 60: 301-316, 2010.
[134] Ripley BD. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.
[135] Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, and Moore JH. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *The American Journal of Human Genetics* 69: 138-147, 2001.
[136] Ritchie MD, Hahn LW, and Moore JH. Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. *Genetic Epidemiology* 24: 150-157, 2003.
[137] Robnik-Šikonja M, and Kononenko I. Theoretical and empirical analysis of ReliefF and RReliefF. *Machine Learning* 53: 23-69, 2003.
[138] Ruczinski I, Kooperberg C, and L. LeBlanc M. Exploring interactions in high-dimensional genomic data: an overview of Logic Regression, with applications. *Journal of Multivariate Analysis* 90: 178-195, 2004.
[139] Saenger YM, and Wolchok dD. The heterogeneity of the kinetics of response to ipilimumab in metastatic melanoma: patient cases. *Cancer Immunity* 8: 2008.
[140] Saeys Y, Inza I, and Larrañaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics* 23: 2507-2517, 2007.

[141] Sasieni PD. From genotypes to genes: doubling the sample size. *Biometrics* 53: 1253-1261, 1997.
[142] Satterthwaite F. An approximate distribution of estimates of variance components. *Biometrics Bulletin* 2: 110-114, 1946.
[143] Satterthwaite FE. Synthesis of variance. *Psychometrika* 6: 309-316, 1941.
[144] Scher HI, Nasso SF, Rubin EH, and Simon R. Adaptive Clinical Trial Designs for Simultaneous Testing of Matched Diagnostics and Therapeutics. *Clinical Cancer Research* 17: 6634-6640, 2011.
[145] Schmoor C, Ulm K, and Schumacher M. Comparison of the cox model and the regression tree procedure in analysing a randomized clinical trial. *Statistics in Medicine* 12: 2351-2366, 1993.
[146] Schott JR. *Matrix Analysis for Statistics*. Wiley, 2005.
[147] Segura MF, Belitskaya-Lévy I, Rose AE, Zakrzewski J, Gaziel A, Hanniford D, Darvishian F, Berman RS, Shapiro RL, Pavlick AC, Osman I, and Hernando E. Melanoma microRNA signature predicts post-recurrence survival. *Clinical Cancer Research* 16: 1577-1586, 2010.
[148] Self SG, Mauritsen RH, and Ohara J. Power Calculations for Likelihood Ratio Tests in Generalized Linear Models. *Biometrics* 48: 31-39, 1992.
[149] Shieh G. A comparison of two approaches for power and sample size calculations in logistic regression models. *Communications in Statistics - Simulation and Computation* 29: 763-791, 2000.
[150] Smyth GK. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology* 3: 2004.
[151] Sorenson T. A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content and Its Application to Analyses of the Vegetation on Danish Commons. *Biologiske Skrifter* 5: 1-34, 1948.
[152] Strobl C, Malley J, and Tutz G. An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods* 14: 323-348, 2009.
[153] Su X, Tsai C-L, Wang H, Nickerson D, and Li B. Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research* 10: 141-158, 2009.
[154] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, and Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* 102: 15545-15550, 2005.
[155] Tarca A, Draghici S, Bhatti G, and Romero R. Down-weighting overlapping genes improves gene set analysis. *BMC Bioinformatics* 13: 136, 2012.
[156] Thomas JG, Olson JM, Tapscott SJ, and Zhao LP. An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Research* 11: 1227-1236, 2001.
[157] Tibshirani R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B (Methodological)* 58: 267-288, 1996.
[158] Tibshirani R, and Knight K. Model search by bootstrap "bumping". *Journal of Computational and Graphical Statistics* 8: 671-686, 1999.
[159] Tibshirani R, and Efron B. Pre-validation and inference in microarrays. *Statistical Applications in Genetics and Molecular Biology* 1: 2002.

[160] Tibshirani R, Hastie T, Narasimhan B, and Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences* 99: 6567-6572, 2002.
[161] Tibshirani R, Hastie T, Narasimhan B, and Chu G. Class prediction by nearest shrunken centroids, with applications to dna microarrays. *Statistical Science* 18: 104-117, 2003.
[162] Tuma RS. Melanoma: two experts on their first-line treatment choices for patients with advanced disease: start with the tortoise or the hare? *Oncology Times* 34: 7-8, 2012.
[163] Vanneman M, and Dranoff G. Combining immunotherapy and targeted therapies in cancer treatment. *Nat Rev Cancer* 12: 237-251, 2012.
[164] Vapnik VN. *Statistical Learning Theory*. Wiley, 1998.
[165] Venables WN, and Ripley BD. *Modern Applied Statistics with S*. Springer, 2002.
[166] Viros A, Fridlyand J, Bauer J, Lasithiotakis K, Garbe C, Pinkel D, and Bastian BC. Improving melanoma classification by integrating genetic and morphologic features. *PLoS Med* 5: e120, 2008.
[167] Wald A. On a Statistical Problem Arising in the Classification of an Individual into One of Two Groups. *The Annals of Mathematical Statistics* 15: 145-162, 1944.
[168] Wand MP, and Jones MC. Comparison of smoothing parameterizations in bivariate kernel density estimation. *Journal of the American Statistical Association* 88: 520-528, 1993.
[169] Wang Y, Tetko IV, Hall MA, Frank E, Facius A, Mayer KFX, and Mewes HW. Gene selection from microarray data for cancer classification—a machine learning approach. *Computational Biology and Chemistry* 29: 37-46, 2005.
[170] Ward JH, Jr. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association* 58: 236-244, 1963.
[171] Welch BL. Note on Discriminant Functions. *Biometrika* 31: 218-220, 1939.
[172] Welch BL. The generalization of "Student's" problem when several different population variances are involved. *Biometrika* 34: 28-35, 1947.
[173] Winham S, Colby C, Freimuth R, Wang X, de Andrade M, Huebner M, and Biernacka J. SNP interaction detection with random forests in high-dimensional genetic data. *BMC Bioinformatics* 13: 164, 2012.
[174] Winter C, Kristiansen G, Kersting S, Roy J, Aust D, Knösel T, Rümmele P, Jahnke B, Hentrich V, Rückert F, Niedergethmann M, Weichert W, Bahra M, Schlitt HJ, Settmacher U, Friess H, Büchler M, Saeger H-D, Schroeder M, Pilarsky C, and Grützmann R. Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes. *PLoS Comput Biol* 8: e1002511, 2012.
[175] Wolpert DH. Stacked generalization. *Neural Networks* 5: 241-259, 1992.
[176] Wolpert DH, and Macready WG. No free lunch theorems for optimization. *Evolutionary Computation, IEEE Transactions on* 1: 67-82, 1997.
[177] Xu P, Brock GN, and Parrish RS. Modified linear discriminant analysis approaches for classification of high-dimensional microarray data. *Computational Statistics & Data Analysis* 53: 1674-1687, 2009.
[178] Yuan M, and Lin Y. Model Selection and Estimation in Regression with Grouped Variables. *Journal of the Royal Statistical Society Series B (Statistical Methodology)* 68: 49-67, 2006.
[179] Zhang B, Tsiatis AA, Laber EB, and Davidian M. Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika* 2013.
[180] Zhang Y, and Liu JS. Bayesian inference of epistatic interactions in case-control studies. *Nat Genet* 39: 1167-1173, 2007.

[181] Zhao L, Tian L, Cai T, Claggett B, and Wei LJ. Effectively Selecting a Target Population for a Future Comparative Study. *Journal of the American Statistical Association* 108: 527-539, 2013.
[182] Zhou Z-H. *Ensemble Methods:Foundations and Algorithms*. CRC Press, 2012.
[183] Zou H, and Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67: 301-320, 2005.
[184] Zou H. The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association* 101: 1418-1429, 2006.
[185] Zou H, Hastie T, and Tibshirani R. Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics* 15: 265-286, 2006.

CURRICULUM VITA

1. CONTACT INFORMATION

Alexander Cambon
Department of Bioinformatics and Biostatistics
Biostatistics Shared Facility
University of Louisville
Louisville, KY 40202
Phone 502-852-4111 Fax: 502-852-7979 Cell 502-541-8869
Alex.Cambon@Louisville.Edu
cambon5810@gmail.com

2. ACADEMIC DEGREES

2007-2014 Ph. D., Biostatistics, University of Louisville, Louisville, KY

1992-1998 Masters of Engineering, Industrial Engineering, Pennsylvania State University, Penn State Great Valley School of Graduate Professional Studies, Malvern, Pennsylvania

1979-1982 B.S. Mechanical Engineering, Syracuse University, Syracuse, NY.

3. ACADEMIC AND PROFESSIONAL EXPERIENCE

University of Louisville, Biostatistician, Louisville, KY 2003 - present

a. Ongoing grant support

Bhatnagar (PI) 2013 - present
Grant Number: OGMB130642, AHA Tobacco Regulation and Addiction Center.

This project is to build and develop a Tobacco Regulation and Addiction Center at the American Heart Association (AHA). This virtual Center will be a network of consortia between AHA and several leading academic institutions. It will house programs of multidisciplinary research that will inform the manufacture, distribution and marketing of tobacco products related to the regulatory authority of the FDA. Center investigators will study the cardiovascular toxicity of tobacco products and the relationship between tobacco product use and subclinical progression of cardiovascular disease to identify sensitive and robust biomarkers of cardiovascular injury related to tobacco product exposure. These studies will help to identify the relative sensitivity of specific domains of cardiovascular injury that are sensitive to tobacco exposure and facilitate the identification of harmful and potentially harmful constituents of current, new and emerging tobacco products that cause cardiovascular injury and how the impact of these constituents on cardiovascular disease outcomes could be measured.

Percentage: 40

Role: Biostatistician

Litvan (PI) 2006 – present
Grant Number: OICB070384, Genetic and Environmental Risk Factors for PSP National Institutes of Health.

The long-range goal of this research is to develop efficacious interventions to delay, slow, or stop the progression of progressive supranuclear palsy (PSP) and related tauopathies.

I co-wrote statistical section of NIH grant proposal for a longitudinal study for PSP, and for CURE PSP longitudinal study. I also reviewed grant proposals and manuscripts. I do data analysis, sample size planning, report generation, and manuscript review and editing.

Percentage: 15 (150% total given in June and July of 2013 to cover 2013-2014 period)
Role: Biostatistician

b. Past Grant Support

DeFilippis (PI) 2012-2013

Grant Number OICB120508, AHA-Great Rivers, "Early Diagnosis of Atherothrombotic Acute Coronary Syndrome"
Percentage: 25
Role: Biostatistician

I did statistical analysis, wrote up reports, and participated in manuscript editing and review.

Ramos (PI) 2007-2011

NIH/NIEHS
2007-2011
Grant Number: 5P30ES014443-02 (Ramos)
Project Title: Center for Environmental Genomics and Integrative Biology (Bioinformatics, Biostatistics and Computational Biology Core)
To create a "Knowledge Exchange Network" that results in promotion of research and educational programs in environmental genomics and integrative biology; provide University of Louisville investigators an outstanding intellectual environment that supports innovation in environmental health sciences; promote collaborative research activities among Center investigators and recruit new talent into the field of environmental health sciences; continue developing genomics and integrative biology approaches to support investigations focusing on environmental cardiology, environmental carcinogenesis and developmental origins of health and disease; and to expand bench-to-bedside, bench-to-community, and reciprocal activities at University of Louisville that benefit underserved communities in the Louisville area. Role: Biostatistician/perform microarray analysis

Role: Biostatistician (including bioinformatics), Data Management

c. Contract Support

Diana Han (PI)
GE Industrial Athlete Program

The purpose of the study is to determine if a planned intervention consisting of ergonomic improvements and a focused training and exercise regimen will increase Functional Movement Screen (FMS) scores and reduce worker-related injuries.

Percentage: 20
Role: Biostatistician

d. Supervisory

I jointly supervise 3 Master-Level Biostatisticians.

e. Projects

- Analysis of Phase I, II, and III Clinical trials.
- Reviewing statistician for over 20 clinical trial protocols. Includes assisting investigators with defining research objectives, generating professional-quality reports including statistical reports, checking and performing sample size calculations, giving guidance on randomization procedures.
- Dissertation topic: Adaptive Signature Design to predict a subset of subjects responding differently to treatment using genomic data.
- Wrote statistical reports and performed survival analysis for study involving endometrial cancer.
- Analysis of Matched Case-Control Study.
- Pathway, cluster analysis, differential gene expression, classification and machine learning for bioinformatics/high throughput data projects.
- Microarray Analysis for Agilent, Affymetrix, two-color Platforms.
- Creation of Phase III Clinical Trial Protocol template for department staff and faculty.
- Multi-rater reliability analysis.
- Statistical Consulting Center projects in microarray analysis: involved helping investigators with research objectives and study design issues.
- Probe level linear mixed model analysis for microarray data involving overproduction of sphingosine kinase.
- Microarray analysis for experiment involving apoptosis induced retinal ganglion cells.
- Power analysis to compare diagnostic tests for flu.
- HW (Hardy Weinberg) equilibrium analysis for case control study involving diseased (Progressive Supranuclear Palsy) and normal offspring.
- Consultant for meta analysis study for small cell lung cancer.
- Survival analysis, recurrent events analysis, and longitudinal analysis for congestive heart failure study, using SAS, S-PLUS, and R.
- SAS analysis for BRFSS (Behavioral Risk Factor Surveillance System) study for APHA presentation “Comparable compliance for comparable preventive health services in American men and women” by Martin Weinrich.
- National Health Interview Survey analysis on prostate cancer using SUDAAN.

GE Appliances, Reliability Statistician, Louisville, KY 1999-2003

- Implemented survival analysis methodology for Supplier Quality, including accelerated testing (e.g. Arrhenius, Inverse Power models), degradation analysis, power analysis, distribution fitting, mixture distribution analysis, comparison testing using relative likelihood, and calculation and application of confidence intervals.

- Led projects qualifying more than 15 components on new product introductions, performing process capability analysis and variance components analysis.
- Led projects or played key role in projects saving more than $3 MM annually in service calls.
- Taught graduate course lecture in Reliability Testing (Survival Analysis). This course receives graduate credit at University of Louisville. Provided statistical consulting for managers, quality leaders, and engineers.
- Developed and conducted training for reliability testing and analysis. Co-developed and taught curriculum for logistic regression short course.
- Implemented Internet-based Supplier Survival Analysis System. Suppliers can enter time-to-event data for components. Software allows GE contact to set sample size and alarm limits based on Type I and Type II errors. Received $1000 GE patent award and US patent.
- Designed and tested smoothing algorithm for new GE dryer. Received $1000 GE patent award. Algorithm implemented in new product launch.
- Co-developed and implemented standing instruction template to standardize reliability testing and qualification of components and systems.
- Statistical Methods include design of experiments for manufacturing processes, logistic regression, nonlinear models, power analysis, comparison analysis using relative likelihood, general linear models, mixture models, categorical data analysis, smoothing and forecasting methods.

GE Plastics, Internal Statistical Consultant, Pittsfield, MA 1996-1999

- Implemented Monte Carlo Simulation for transactional and manufacturing processes. Conducted training in Monte Carlo simulation. Corroborated with GE Corporate Research Ph.D. Statisticians in selection of corporate simulation package.
- Conducted extensive training for six sigma projects. Training included Design of Experiments, Regression, ANOVA.
- Co-developed statistical process control module and training for GE Plastics.
- Statistical Methods included Mixed Models, Logistic Regression, Response Surface Analysis, Design of Experiments, General Linear Models, ANOVA, Monte Carlo simulations.

Graco Children's Products, Manufacturing Statistician, Elverson, PA
1987-1996

- Implemented statistical process control in metal stamping, plastics injection molding, paint line, tube bending, and assembly line.
- Designed and implemented design of experiments on injection molding equipment – achieved >$30,000 cost avoidance.
- Designed statistical process control chart and decision rules to reduce inventory fluctuations and smooth out delivery to major customer.
- Taught basic statistics, measurement analysis, statistical process control, design of experiments.
- Statistical methods include ARIMA time series, random effects models, statistical process control, design of experiments, regression.

Mennonite Central Committee, Water Resource and Health Development Engineer, Burkina Faso West Africa 1982-1985

- Trained villagers to build small earthen dams to hold water for use during the dry season.

- Taught basic health related to water usage.

4. PUBLICATIONS

Cambon AC, Khalyfa A, Cooper N, and Thompson C. Analysis of probe level patterns in Affymetrix microarray data. *BMC Bioinformatics*, 8(1):146, 2007.

Kalbfleisch T, Cambon AC, and Wattenberg BW. A Bioinformatics Approach to Identifying Tail-anchored Proteins in the Human Genome. *Traffic*, 8(12):1687-1694, 2007.

Litvan I, Chism A, Litvan J, Cambon AC, and Hutton M. H1/H1 genotype influences symptom severity in corticobasal syndrome. *Movement Disorders*, 25:760-763, 2010.

Luo D, Cambon AC, and Wu D. Evaluating the long-term effect of FOBT in colorectal cancer screening. *Cancer Epidemiology*, 36:e54-e60, 2011.

Eng M, Zhang J, Cambon AC, Marvin MR, and Gleason J. Employment outcomes following successful renal transplantation. Clinical Transplantation, 26:242-246, 2011.

Kong M, Cambon AC, and Smith M. Extended logistic regression model for studies with interrupted events, seasonal trend and serial correlation. *Communications in Statistics-*Theory and Methods, 41:3528-3543, 2012.

Potts LF, Cambon AC, Ross OA, Rademakers R, Dickson DW, Uitti RJ, Wszolek ZK, Rai SN, Farrer MJ, and Hein DW. Polymorphic genes of detoxification and mitochondrial enzymes and risk for progressive supranuclear palsy: A case control study. *BMC Medical Genetics*, 13:16, 2012.

Smith MJ, Kong M, Cambon AC, and Woods CR. Effectiveness of Antimicrobial Guidelines for Community-Acquired Pneumonia in Children. *Pediatrics*, 129:e1326-e1333, 2012.

Newton TL, Fernandez-Botran R, Miller JJ, Cambon AC, Burns VE, and Allison KE. Posttraumatic Stress Symptom Severity and Inflammatory Processes in Midlife Women. *Psychological Trauma: Theory, Research, Practice, and Policy*, 5(5): 439-447, 2012.

Rai SN, Pan J, Cambon AC, Gargett N, and Chaires JB. Group Classification based on High-Dimensional Data: Application to Differential Scanning Calorimetry Plasma Thermogram Analysis of Cervical Cancer and Control Samples. *Open Access Medical Statistics*, 2013(3): 1-9, 2013.

DeFilippis AP, Rai SN, Cambon AC, Miles RJ, Jaffe AS, Moser AB, Jones RO, Bolli R, and Schulman SP. Fatty acids and TxA2 generation, in the absence of platelet-COX-1 activity. *Nutrition, Metabolism & Cardiovascular Diseases*, 24(4): 428-433, 2014.

Barton C, Kouokam JC, Lasnik AB, Foreman O, Cambon AC, Brock GN, Montefiori DC, Vojdani F, McCormick AA, O'Keefe BR, and Palmer KE. Activity of and Effect of Subcutaneous Treatment with the Broad-Spectrum Antiviral Lectin Griffithsin in Two Laboratory Rodent Models. *Antimicrobial Agents and Chemotherapy*, 58(1):120-127, 2014.

Rai SN, Ray HE, Pan J, Barnes C, Cambon AC, Wu X, Bonassi S, and Srivastava DK. Phase II Clinical Trials: Issues and Practices. *Biometrics and Biostatistics International Journal*, 1(2): 1-3, 2014.

Cambon AC, Baumgartner KB, Brock GN, Cooper NGF, Wu D, and Rai SN. Classification of Clinical Outcomes Using High-Throughput Informatics: Part 1- Nonparametric Method Reviews. *Model Assisted Statistics and Applications*, In Press, 2014.

Cambon AC, Baumgartner KB, Brock GN, Cooper NGF, Wu D, and Rai SN. Classification of Clinical Outcomes Using High-Throughput Informatics: Part 2- Parametric Method Reviews. *Model Assisted Statistics and Applications*, Accepted, 2014.

Submitted

Steiner RWP, Brock GN, Cambon AC, Anderson SA, Lewis JN, and Morse JH. Evaluating the Impact of a Work Site Tobacco Smoking Ban on Healthcare Utilization among Active Employees. *Nicotine and Tobacco Research*, April 2014.

Amraotkar AR, Cambon AC, Rai SN, Keith MCL, Ghafghazi S, Bolli R, and DeFilippis AP. Risk of E. coli Contamination in Non-Municipal Waters Consumed by Mennonites versus Other Rural Populations. *The American Journal of Public Health*, June 2014.

In Preparation

Amraotkar AR, Boman M, Nair R, Cambon AC, Rai SN, Bolli R, and DeFilippis AP. Sensory Integration in Individuals with Autism Spectrum Disorders Entering White versus Black Sensory Room.

Cambon AC, Baumgartner KB, Brock GN, Cooper NGF, Wu D, and Rai SN. Properties of Adaptive Clinical Trial Signature Design in the Presence of Gene and Gene –Treatment Interaction.

Cambon AC, Baumgartner KB, Brock GN, Cooper NGF, Wu D, and Rai SN. Estimating Design Parameters in the Presence of Gene and Gene -Treatment Interaction Using High-Throughput Informatics in Clinical Trials.

5. VOLUNTEER SERVICE

Water and Health Development Engineer, Mennonite Central Committee, 1982-1985 – see #3 above for more detail.

GE Elfun Society – 1997-2003 – participated in numerous projects including Kentucky School for the Blind renovations, school tutoring, etc.

Walk for Diabetes Fundraiser, Louisville, KY, 2000.

Part of group to Eastern Kentucky (Appalachia) to provide needed home repairs for elderly residents, 2001.

Tutored grade K-12 inner city in Math, English, Spelling (Here's Life Inner City), 2001-2002.

Taught English to Chinese Middle School Students in Hubei Province, China, Summer 2002.

Lead volunteer group to gut houses in 9th ward and other hard hit areas in New Orleans, 2006.

English tutoring for Burundi Refugees through Kentucky Refugee Services, 2007.

Susan R Komen Race for the Cure, 10K Run, 2012.

6. TEACHING

PHPH 610, Data and Statistics Management for Public Health using SPSS, Co-Instructor, Spring 2013.

PHST 724, Advanced Clinical Trials, Co-Instructor and Teaching Assistant, Spring 2013.

Graduate Level Course Lecture, Reliability Testing and Analysis, GE Appliances, Course approved for graduate credit in Engineering by University of Louisville, (2000-2002).

One-week Reliability Analysis Course, Shanghai, China, and Seoul, Korea to GE Design Engineers, Quality Engineers, and Sourcing Managers (August, 2001).

One-week Reliability Course to GE Supplier Engineers in Guangzhou, China (February, 2002).

Co-developed and taught Short Course in Logistic Regression (GE Appliances, 2001).

Developed and taught Accelerated Testing Short Course at GE Appliances (2000-2003).

Taught English to Chinese Middle School Students in Hubei Province, China (Summer 2002).

Monte Carlo Simulation short course instructor at GE Plastics (1998).

40 hour course (Statistical Process Control), Graco Children's Products (1991-1996).

Short Course in Design of Experiments – Graco Children's Products (1996).

Basic health course for West African Village (1982-1985) – focus on water usage and health.

7. LANGUAGES

English (fluent), French (spoke daily for 2 ½ years), Chinese Mandarin (Advanced), Italian (Intermediate), Birifor (West African Language – basic), Spanish (basic).

8. STATISTICAL SOFTWARE PACKAGES/PROGRAMMING

R	Partek
SAS	East
Bioconductor	NCSS/PASS
dChip	Xemacs
S-PLUS	Ingenuity
SPSS	Excel
nQuery Advisor	StatXact
East	Weibull++
ALTA	RG (Growth Models)
Process Model (Simulation)	ProModel – (Simulation)
Link Plus (Probability Matching)	Minitab
GSEA/GSA (Pathway analysis software)	Access
Genetic/SNP Analysis Software packages	DAVID

8b. TRAINING

Bayesian Methods in Pharmaceutical Development, ASA Webinar, Biopharmaceutical Section, Instructors Karen Price, Mani Lakshminarayanan (November 2014)

Genomic Clinical Trials and Predictive Medicine, Instructor-Richard Simon, One day short course at Joint Statistical Meeting, (August 2014).

Partek software for NextGen sequencing- Alignment, QA/QC, Downstream Analysis, etc. (February 2013).

National Institute of Environmental Health Sciences (NIEHS) SNPs Two-day Workshop, 1/11/2008 (Training in Genetic/SNP Analysis)

National Science Foundation (NSF) - Funded Short Course on Statistical Genetics and Statistical Genomics (1 Week, UAB, July, 2008)

Joint Statistical Meeting (JSM) One-Day Short Course: Statistical Methods for Genetic Analysis, Kenneth Lange (August 2007).

JSM One-Day Short Course: Probability Linkage (August 2007).

8c. COMMITTEES

Software Committee- developed Access database to keep track of department software license statistics by person, by contract, company, etc.

9. CERTIFICATIONS

American Society for Quality (ASQ) Certified Reliability Engineer since 1992.
ASQ Certified Quality Engineer since 1991.
ASQ Certified Quality Auditor since 1992.
Certified Six Sigma Instructor (General Electric, 1998-2003).
Certified Six Sigma Leader (General Electric, 2000-2003).
Certified Statistical Process Control Instructor (Quality Institute ~1994).
GE Reliability Expert (1999-2003).
GE Certified Reliability Practitioner (2000).
Certified over 25 Reliability Practitioners (2000-2003).

10. PROFESSIONAL SOCIETIES

American Statistical Association (Kentucky Chapter) since 1996.

11. EDITORIAL BOARD, COUNCILS AND COMMITTEES – SERVICE.

Reviewer for Journal of Biometrics & Biostatistics, 2014 - present.
American Statistical Association (ASA), Vice Chair District 2, Region 1, Council of Chapters Governing Board, 2012-2014.
American Statistical Association, Kentucky Chapter Judge, duPont Manual High School Regional Science Fair, Louisville, KY, 2012.
American Statistical Association, Kentucky Chapter Representative, 2004- 2012.
American Statistical Association Judge, International Science and Engineering Fair, Louisville, KY, 2002.
Chaired GE Appliances Reliability Practitioner (Survival Analysis) Best Practices monthly conference meetings 2000-2003.
ASME (American Society for Mechanical Engineers) Student Chairman, Syracuse University, 1982.

12. AWARDS AND NAMED LECTURESHIPS

Magna Cum Laude, Syracuse University, 1982.
Outstanding Academic Achievement Award, Pennsylvania State University, 1998.
GE Call Center Simulation Award, 1998.
GE Instructor for Graduate Level Engineering Course Lecture in Reliability, 1999-2003.
GE Certified Six Sigma Trainer, 1997-2003.
GE Appliances Survival Analysis Expert, 2000-2003.
GE Certified Six Sigma Black Belt, 2000-2003.
GE Patent Award ($1,000) – Smoothing Algorithm for Dryer Sensor, 2001.
GE Patent Award ($1,000) - Design of Internet Based Supplier Reliability System, 2001.
ASA Judge, International Science and Engineering Fair, Louisville, Kentucky, 2002.
US Patent 6,675,129 - Internet Based Supplier Process Reliability System, 2004.
US Patent 7,013,578 - System and method for controlling a dryer appliance, 2006.

13. POSTERS/PRESENTATIONS/ABSTRACTS

Classification of Clinical Outcomes Using High-Throughput Data, Alexander Cambon, Kathy B Baumgartner, Guy N Brock, Nigel GF Cooper, Dongfeng Wu, Shesh N Rai; Joint Statistical Meeting, Boston (August 2014).

An Evaluation of a Simon 2-Stage Phase II Clinical Trial Design Incorporating Continuous Toxicity Monitoring, Herman Ray, D Kumar Srivastava, Alexander Cambon, Shesh N Rai (2014).

Model Based Classifications of High-Throughput Data- Review, Design and Application to a Cancer Clinical Study, Alexander Cambon, Shesh N Rai; Joint Statistical Meeting, Montreal (July-August 2013).

A review of classification methods which could be used to identify a subset of patients in a clinical study", Alexander Cambon, Shesh N Rai; Joint Statistical Meeting, San Diego (July-August 2012).

Invited Guest Lecturer for Graduate Course Seminar Series PHST 602. Presentation entitled "Gene Set Analysis" (January 2009).

Presentation of statistical analysis of bioinformatics project: "Response of Oral Cavity Cells to Cigarette Smoke Components", Louisville, GEGIB-BBCB (February 2008).

Poster Presentation: "Using Link Plus for Probability Matching in Kentucky's Newborn Screening and Birth Defects Data", Alexander Cambon, Sandy Fawbush, Charles Mundt, Joyce Robl; Maternal Child Health Conference, Atlanta (December 2007).

Invited Guest Lecturer for Biostatistics Seminar (coordinated by Dr. Rempala) Presentation entitled "Analysis of a Probe Level Linear Mixed Model for Oligonucleotide Arrays" (March 2007).

Contributed Presentation: "Analysis of a Probe Level Linear Mixed Model for Oligonucleotide Arrays", Alexander Cambon, Dr. Caryn Thompson, Dr. Brian Wattenberg, Joint Statistical Meeting (August 2006).

Poster Presentation: “Probe Level Patterns in Affymetrix Microarrays”, Alexander Cambon, Abdelnaby Khalyfa, Caryn Thompson, Nigel Cooper, Kentucky KBRIN conference, Land-Between-the-Lakes (April 2006).

Poster Presentation: “Pathway Analysis and Gene Signal Identification of Microarray Data for Apoptosis Induced Retinal Ganglion Cells”, Alexander Cambon, Abdelnaby Khalyfa, Caryn Thompson, Nigel Cooper; CHI Pathway Analysis Conference, San Francisco (February 2006).

Invited Guest Lecturer for Graduate Course in Survival Analysis. Presentation entitled "Recurrent Events Analysis" (November 2004).

An Application of Recurrent Events Analysis, Joint Statistical Meeting Proceedings, Toronto, Canada (Summer 2004).

Invited Guest Lecturer for PHDA 602, University of Louisville; Survival Analysis Methods in Industry (2003).

ASA/QPRC Northeast Meeting; “Using Monte Carlo simulation to Estimate an Optimum Ratio of “Good” to “Bad” Parts in an Attribute Gage Study” (1998).

“An Application of Design Experiments in Design of Baby Swings”, American Society for Quality, Reading, PA Chapter, ~1993.