WEAKEST-LINK METHODS AND APPLICATIONS FOR DETECTING JOINT EFFECTS IN BIOLOGY

by

The Minh Luong

B.Sc., McGill University, 1998

M.Sc., McGill University, 2000

Submitted to the Graduate Faculty of

the Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2011

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

The Minh Luong

It was defended on

October 1, 2010

and approved by

Roger S. Day, Sc. D, Associate Professor
Department of Biomedical Informatics, School of Medicine
Department of Biostatistics, Graduate School of Public Health
University of Pittsburgh

Stewart J. Anderson, Ph. D., Professor
Department of Biostatistics, Graduate School of Public Health
University of Pittsburgh

Eleanor Feingold, Ph. D., Professor
Department of Human Genetics, Graduate School of Public Health
Department of Biostatistics, Graduate School of Public Health
University of Pittsburgh

George C. Tseng, Sc. D., Assistant Professor
Department of Biostatistics, Graduate School of Public Health
University of Pittsburgh

Dissertation Director:
Roger S. Day, Sc. D, Associate Professor
Department of Biomedical Informatics, School of Medicine
Department of Biostatistics, Graduate School of Public Health
University of Pittsburgh

WEAKEST-LINK METHODS AND APPLICATIONS FOR DETECTING JOINT EFFECTS IN BIOLOGY

The Minh Luong, PhD

University of Pittsburgh, 2010

The joint effect of several variables is a prevailing statistical concept in biology. The public health importance of developing methods to better assess joint effects is evident when studying gene combinations that function together to produce a disease phenotype, or biomarker pairs that jointly affect prognosis or treatment response.

The "weakest-link" paradigm, introduced earlier by Richards and Day, constructs derived covariates accounting for the joint effect of multiple variables. The weakest-link method posits a one-dimensional locus in covariate space, called the curve of optimal use (COU). For a data set with two predictors and an associated outcome, the COU separates the two-dimensional covariate space into two subsets. The subset of an observation determines its *weakest-link* covariate, which alone locally affects the corresponding outcome. With a modest generalization, one can extend weakest-link methods to assess interactions between more than two variables.

Current methods for detecting interesting variable combinations have shortcomings. Some methods, such as logic regression, require dichotomization, and lose information. Other methods such as support vector machines, are too computationally intensive, especially with large data sets.

With these issues in mind, the primary objectives in expanding the practical applications of weakest-link methodology are: (1) to develop a semi-parametric method to screen hundreds or thousands of variables for combinations associated with an outcome, (2) to adapt the method for a more complicated data structure found in a multi-parameter cell-

based cytometry study, where data sets typically consist of thousands of cell observations per outcome.

In a high-throughput microarray data set of breast cancer patients, conventional additive linear models and weakest-link models identified multiple combinations of biomarkers associated with lymph node positivity. Simulations of high-throughput data sets found that weakest-link models had better success than additive models in detecting covariate pairs used to generate outcomes; weakest-link models were preferable even in some situations when the additive model was the true outcome-generating model.

The weakest-link approach showed promising results in modeling recurrence-free survival in a cytometry data set of lung cancer samples. Weakest-link models, compared to logic regression and linear regression, provided the best results according to cross-validation assessments.

TABLE OF CONTENTS

PREFACE x

1.0 INTRODUCTION 1

2.0 LITERATURE REVIEW 5

2.1 Statistical methods used in detecting and assessing joint effects 5

2.1.1 Model fitting, assessment, and selection 8

2.2 Weakest-link methods 9

2.3 Cytometry 10

2.3.1 Multiplex biomarker data for predicting survival 12

2.4 Limitations of current methods 12

3.0 OBJECTIVES 14

4.0 WEAKEST-LINK METHODOLOGY 16

4.1 Weakest-link notation 16

4.2 Stitched weakest-link models 20

4.2.0.1 Quantile stitching 22

4.2.0.2 Probit stitching 22

4.2.1 Maximum likelihood estimation for weakest-link models 25

4.2.2 Directionality in stitched weakest-link methods 26

4.3 Model assessment and selection through cross-validation 29

4.4 Variable combination screening for high-throughput data sets 30

4.4.1 Filtered subsets of covariates 30

4.4.2 Greedy search algorithm for locally optimal combinations 31

4.4.2.1 Finding a locally optimal combination 31

4.4.2.2 Finding multiple combinations 32

4.4.3 Simulated annealing procedures . . . 33
4.5 General notation for hierarchical data . . . 34
4.6 Logic regression methods for detecting joint combinations of binary variables 37
5.0 STUDY PLAN . . . 39
6.0 COVARIATE PAIR DETECTION FOR LARGE SCALE DATA SETS 43
6.1 Screening covariate pairs from a data set of breast cancer patients . . . 43
6.1.1 Combinations of biomarkers associated with lymph node status . . . 43
6.1.1.1 Exhaustive filtered search for pairwise combinations . . . 45
6.1.1.2 Greedy search for locally optimal pairwise combinations . . . 47
6.1.1.3 Linear combinations of pairs . . . 53
6.1.2 Discussion of analysis of breast cancer data set . . . 56
6.2 Covariate pair detection comparisons through simulation . . . 57
6.2.1 Simulation procedures . . . 58
6.2.2 Simulation results . . . 61
6.2.3 Sample plots of covariates by outcome type . . . 74
6.2.4 Computation times . . . 80
6.2.5 Discussion of simulation studies . . . 85
6.2.5.1 Comparisons between models . . . 85
6.2.5.2 Comparison of different search algorithms . . . 88
7.0 ASSESSING JOINT EFFECTS IN A CYTOMETRY DATA SET . . . 90
7.1 Pennsylvania lung cancer data set . . . 90
7.1.1 Results using quantile stitching weakest-link . . . 91
7.1.2 Weakest-link for identifying high-risk groups of patients . . . 94
7.1.3 Effects of level of weakest-link assessment in cell-based data . . . 99
7.2 Discussion of analysis of cytometry data . . . 101
8.0 R PACKAGE WEAKESTLINK . . . 104
BIBLIOGRAPHY . . . 106

LIST OF TABLES

1 Weakest-link without main effects example 31

2 Steps to obtain two different quantile stitched covariates 35

3 Genes with highest marginal association with LN+ 46

4 Covariate pairs associated with lymph node status, after exhaustive filtered search within the subset of genes with marginal $p < 0.001$ 47

5 Covariate pairs associated with lymph node status, after greedy search for multiple local minima using probit stitching weakest-link model 49

6 Covariate pairs associated with lymph node status, after greedy search for multiple local minima using main and interaction effects model 52

7 Covariate pairs with lowest misclassification rates for predicting lymph node status by cross-validation 55

8 Lymph node status constructed by linear combinations of covariate pairs . . . 56

9 Observed proportions of replicates detecting correct pair out of 10 simulated normally-distributed covariates of size 50, 3 binary outcomes 65

10 Observed mean number of significant pairs detected per replication, from 10 simulated normally-distributed covariates of size 50, 3 binary outcomes 66

11 Observed proportions of replicates detecting correct pair out of 10 simulated normally-distributed covariates of size 100, 3 binary outcomes 67

12 Observed proportions of replicates detecting correct pair out of 10 simulated normally-distributed covariates of size 100, one set of binary outcomes simultaneously generated from 3 covariate pairs 68

13 Observed proportions of replicates detecting correct pair out of 100 simulated normally-distributed covariates of size 100, 3 binary outcomes 69

14 Observed mean number of significant pairs detected per replication, from 100 simulated normally-distributed covariates of size 100, 3 binary outcomes . . . 71
15 Observed proportions of replicates detecting correct pair out of 100 simulated right-skewed covariates of size 100, 3 binary outcomes 73
16 T-test vs logistic regression: observed proportions of replicates detecting correct pair out of 100 simulated covariates of size 100, 3 binary outcomes. . . . 83
17 Computation time of methods in screening for pairs 84
18 Recurrence-free survival: results from Panel 1 92
19 Recurrence-free survival: results from Panel 2 93
20 Recurrence-free survival in subgroups by unsupervised clustering 97
21 Kaplan-Meier median survival in subgroups detected by maximally selected weakest-link derived covariates 100

LIST OF FIGURES

1 Curve of optimal use (COU) plot for 2 covariates 17
2 Curve of optimal use (COU) plot for quantile stitching 23
3 Curve of optimal use (COU) plot for probit stitching 24
4 Curve of optimal use (COU) plots for weakest-link directionalities 28
5 Dot plots of biomarkers Oncogene JUN-D and hBrm 50
6 Probit stitching weakest-link between Oncogene Jun-D and hBrm protein . . 51
7 Scatterplots of pairs associated with LN+ detected by both models 54
8 Sample scatterplot of weakest-link generated outcomes 75
9 Sample scatterplot of interaction model generated outcomes 77
10 Comparison of contours from weakest-link and interaction models 78
11 Sample scatterplot of interaction model generated outcomes, uniform covariates 79
12 Sample scatterplot of outcomes generated by neg. correlated covariates 81
13 Scatterplot between her-2/neu and DNA 95
14 Plots of derived covariates from quantile stitching weakest-link 98
15 Weakest-link covariates when taken at the patient level and cell level 102

PREFACE

This dissertation was supported by Cancer Center Support Grant # P30 CA4790413.

There are many people who, both directly and indirectly, made this dissertation possible. I would first like to thank my advisor, Roger Day. Over several years and countless regular meetings, his input, knowledge, suggestions and critiques have been invaluable. I would also like to thank the members of my committee, Drs. Anderson, Feingold and Tseng, for their suggestions and willingness to work with me on this project.

A deep thanks also goes to the faculty and staff of University of Pittsburgh Cancer Institute Biostatistics Facility, most notably its directors James Schlesselman and Daniel Normolle. While the financial support with my graduate student research-ship was essential for completing my doctoral study, the experience I got in terms of both statistical knowledge and collaboration will be essential to my professional development.

I would like to note the valuable help of the Pittsburgh Supercomputing Center in developing simulations by parallel computing that provided an essential part of this dissertation. Also, I would like to thank Stanley Shackney for providing and explaining the invaluable cytometry data analyzed in this study. I would also like to thank the many friends that I have made in Pittsburgh, both inside and outside of the Graduate School of Public Health, partly for help academically but mostly for enriching my experience over five years in a new city.

Finally, and most of all, I would like to thank my parents The Man Luong and Corina Mongcal, both of whom have passed through this PhD path before I did. Even though they are here only in spirit to witness my graduation, it would not have been possible without their love, support, example, sacrifice and inspiration. This dissertation is forever dedicated to them.

1.0 INTRODUCTION

Interaction, the non-additive combined effects of several variables, is one the most prevailing concepts in statistical modeling in medical studies. The conventional way to model combinations of covariates in statistical procedures is to place them in a linear regression model, with terms denoting the main effects and interaction effects. Interaction terms are typically expressed as the product of several variables, and are usually only investigated if their individual main effects are significant. However, this may not be an accurate method of modeling biological interaction, the true relationship of multiple characteristics and their effect on outcome variables.

Interaction plays a key role across a wide variety of biological studies. One area where conventional statistical interaction may not adequately account for biological interaction is in genetics. The concept of epistasis relates to the effect of combinations of genes. In a purely epistatic model, a phenotype requires the simultaneous presence of specific alleles from several genes. These alleles may be associated with biomarkers such as proteins. If the marginal associations between outcomes and these biomarker variables are weak they may be undetectable. Many model selection procedures do not investigate the interaction term unless marginal effects are present. As a result, we may fail to detect many interesting associations.

A previous dissertation introduced methods based on the "weakest-link" paradigm as a way of modeling interactions. A simple illustration of the weakest-link is recovery from an advanced breast cancer. Two important components in treating Stage II breast cancers are chemotherapy, to destroy cancer cells and prevent their spread throughout the body, and radiation, to destroy cells missed during surgery. If sufficient chemotherapy is being given but radiation is insufficient, recovery will not be aided by simply increasing the dose

of chemotherapy as it has a different target. In this case, *radiation* is the *weakest-link* and increasing this part of therapy should be the target. Alternatively, in another patient increasing the dose of *chemotherapy* would be more helpful in recovery.

In general in severe diseases, the *weakest-link* covariate for multiparametric data from an individual may point to a mechanism that should be targeted by treatment. Expression levels of proteins provide a more quantitative example. Many biological functions are performed by complexes of pathways. The function of these systems depends on the relative proportions of their constituents.

For a bivariate set of data with an associated outcome, the weakest-link method posits a one-dimensional locus in covariate space, called the curve of optimal use (COU). Along the COU, reducing either of the covariates alters the expected value of the outcome. Furthermore, the COU separates the two-dimensional covariate space into two subsets. The location of an observation relative to this COU identifies the subset and its corresponding *weakest-link* covariate for that data point. The outcome for this observation can only change by altering its weakest-link covariate. Unlike expressions from additive models, the effects of this covariate cannot be compensated by changing another covariate.

In current studies in fields such as cancer, the joint effect of variables for patient characteristics, or different biomarkers from their tissue or blood samples, is often of interest. However, data sets can easily consist of hundreds or thousands of possible predictor covariates for each outcome. A practical method should then screen this huge amount of data for interesting combinations of variables in an efficient manner. Current methods, such as logic regression, are possible tools for detecting interesting covariate combinations. However, these methods often require the features to be binary or categorized, and thus may discard valuable information. The categorization may thus reduce the usefulness of the interaction model for practical applications.

We will further develop the practical applications of the weakest-link paradigm with several practical issues of current methods in mind:

1. The computational efficiency in screening for combinations of two or more variables, out of hundreds or thousands of variables, associated with an outcome.

2. The optimal characterization of the association between combinations of these screened variables and a set of outcomes.

3. More complicated data reduction problems found in cell-based studies such as cytometry, whose data consist of multiple observations per outcome.

The weakest-link procedure would thus be a method that is less computationally intensive and does not require extensive user input or statistical expertise. Additionally, it also does not require dichotomization and preserves the continuous nature of continuous data if chosen to do so.

In terms of cell-based studies such as cytometry, the covariates derived through the weakest-link method can also summarize information across many covariates within a cell, and across many cells from a patient. A practical weakest-link model can account for the interaction within individual cells; a previous study suggested the joint relationship from several biomarkers within individual cells of a patient can provide more helpful information than the overall joint effect of biomarker levels in the patient.

The organization of this dissertation is as follows. Chapter 2 summarizes the current literature on methods for assessing statistical interaction. These methods include linear regression, machine learning, CART, mixture models, and others. It also gives a short introduction to cytometry procedures and the biological markers typically analyzed in these studies. Chapter 3 gives a detailed summary of the objectives for this dissertation. Section 4 provides a detailed description of the weakest-link method as it applies to multiparameter studies in terms of searching for significant combinations of variables, and cell-aggregation techniques for cell-based studies. It also provides a description of the various search methods used in this dissertation, and the simulation procedures used to assess their effectiveness. Chapter 5 provides a detailed summary of the study plan of this dissertation, for each of the aforementioned objectives. Chapter 6 describes an application of these methods in screening a high-throughput data set of breast cancer patients for combinations of covariates related to lymph node positivity. This chapter also compares the effectiveness of these procedures in a simulation study across a wide variety of scenarios. Chapter 7 assesses the methods in a data set with hierarchical data structure, by constructing models for predicting recurrence-

free survival in a cytometry data set of lung cancer patients. Chapter 8 provides a brief introduction to the *R* package *weakestLink* in development.

2.0 LITERATURE REVIEW

The following sections will summarize current statistical methods presented in this dissertation and their applications in assessing joint effects and interaction. Later sections provide relevant articles pertaining to cytometry techniques and some of the biomarkers of interest.

2.1 STATISTICAL METHODS USED IN DETECTING AND ASSESSING JOINT EFFECTS

The following procedures are common approaches for finding similar statistical patterns across multiple covariates, which can help in identifying a subset of features of interest to the researcher.

Regression techniques Genome-wide searches using logistic regression models with interaction terms have been implemented for case-control studies (Marchini et al., 2005), however they can miss interactions that include markers with weak marginal effects (Zhang and Liu, 2007) or be computationally intensive.

Discriminant function analysis In discriminant function analysis, the results of a fitted model predict the group membership for each of the individual observations. One application of discriminant function analysis is to diagnose types of disease in data sets consisting of acute leukemia patients (Ratei et al., 2007).

Multifactor dimensionality reduction With genomics in mind, another approach to associating binary outcomes with combinations of categorical attributes is multifactor dimensionality reduction (MDR) (Ritchie et al., 2001). This supervised procedure commonly separates patients into high and low-risk groups (or the outcome Y classified as 1 or 0) by

analyzing observations consisting of multilocus genotypes. The goal of MDR is to find a set of covariates, and a corresponding class prediction (into $Y = 1$ and $Y = 0$) for each possible combination of these covariates. The class predictions are chosen to maximize the ratio of observed $Y = 1$ and $Y = 0$ between two the prediction groups. The initial implementation of this method was for balanced case-control studies. The algorithm is exhaustive and not feasible when there are more than 100 attributes in the data set. Proposed implementations for improving computational time include an initial stage for filtering genes, and a simulated annealing algorithm.

Penalized logistic regression Penalized logistic regression (PLR) deals with the problem of overfitting in logistic regression models. Practical applications of PLR are in microarray (Zhu and Hastie, 2004) and genetics (Park and Hastie, 2008) studies. PLR's advantages over other methods include more stable coefficient fits, thus reducing overfitting and multicollinearity; it provides an advantage over SVM in that it estimates group probabilities rather than a strict classification. PLR stabilizes the fitted curve by adding a "penalty" to large fluctuations of the estimate parameters. This penalty consists of adding a term $\lambda_k J(\beta_k)$ to the log-likelihood of the regular logistic log-likelihood. Typically, the penalty term $J(\beta)$ is set to a quadratic $\|\beta\|$, enabling a model fit similar to ridge regression. However, model selection is often not straightforward. The choice of λ may be complicated and require cross-validation procedures tested at different values of λ. Additionally, depending on the chosen λ and the penalty term $J(\beta)$, the coefficients β_k may be precluded from shrinking to zero, thus requiring forward and backward selection procedures. In simulations, Zhu and Hastie (2004) found penalized logistic regression as effective as SVM for classification problems in terms of cross-validation, while selecting fewer genes than SVM.

Adaptive learning techniques Another adaptive machine learning method is boosting, which consists of repeatedly classifying observations, and in subsequent iterations, giving more weight to those observations that were misclassified. Two of the more effective adaptive methods are AdaBoost and LogitBoost, respectively using exponential and linear-scale re-weighting.

Classification and regression trees CART (classification and regression trees) is a non-parametric method for building a decision tree (Breiman et al., 1984), consisting of bi-

nary covariates. For each covariate, CART finds an optimal cutpoint to split the observation set into two branches (or groups). The splitting then repeats until a stopping rule is satisfied, whereby each terminal node represents a subgroup of ideally homogeneous observations where either a class outcome (classification) or continuous response (regression) can be well-predicted. The general implementation of CART uses a greedy algorithm for determining the local ideal split, and then removing branches (pruning) to reduce overfitting. A clear disadvantage of CART is that dichotomizing the predictor covariates typically results in a loss of information. The random forests procedure (Breiman, 2001) is an extension of the CART method, consisting of several decision trees obtained from CART from several resamplings. The mode of these numerous decision trees is the final accepted model.

Logic regression Logic regression (Ruczinski, 2000; Ruczinski et al., 2003) obtains a linear model consisting of one or more decision trees, to signify separate combinations of covariates linked by Boolean operators. As with CART algorithms, the covariates must be binary. CART produces trees consisting of splits of individual categorical covariates, whereas a logic regression tree consists of splits of combinations of covariates. A typical application of logic regression is in genomics studies for identifying combinations of single nucleotide polymorphisms (SNPs) indicative of an increased risk of disease (Kooperberg et al., 2001; Schwender and Ickstadt, 2007). Logic regression minimizes a scoring function, chosen according to the outcome type; some of these scores include residual sums of squares for linear regression, misclassification error for classification, and partial log-likelihood for Cox regression. Current *R* implementations of logic regression include simulated annealing and a faster greedy search to deal with data sets with a large number of covariates.

Other approaches to detecting joint associations One Bayesian approach to detecting epistatic genetic interactions is Bayesian epistatic association mapping (BEAM), introduced by Zhang and Liu (2007). BEAM was designed for genome-wide studies, where the data consists of thousands of markers. A typical data set consists of binary outcome data, and alleles represented by categorical covariates such as SNPs. A Markov-Chain Monte Carlo approach, the Metropolis-Hastings procedure, simulates the posterior probability that each marker, or combination of markers, is involved in the epistasis.

The restricted partitioning method (RPM) is another method to detect epistatic interactions in the possible absence of main effects (Culverhouse et al., 2004). It reduces computation time by merging combinations of similar genotype groups, as assessed by mean continuous trait values. Thus, it does not require all possible genotype combinations to be investigated.

Greedy searching algorithms Greedy algorithms are a common approach to identify useful variables out of hundreds (or thousands). These types of procedures find the best local solution at each individual step. In a variable selection problem, each step pertains to the inclusion or exclusion of a single variable, or analyzing their marginal effect. In a data set with many covariates, a greedy algorithm finds a locally optimal solution. As such, they may not detect all associations, but are still necessary to keep the computational time within feasible limits. Implementations of variable selection algorithms such as CART (Breiman et al., 1984) and logic regression (Ruczinski et al., 2003) generally include an option for greedy searches in problems where exhaustive searches are not feasible due to combinatorial explosion.

Stochastic filtering approaches Stochastic methods are other approaches for identifying interesting variable combinations when computational time is prohibitive. One such filtering approach is SNPHarvester (Yang et al., 2009), specifically designed for obtaining epistatic combinations from binary SNP data. SNPHarvester randomly first obtains an initial set of markers. A greedy search then finds the locally optimum path of variable combinations within this subset, while also recording other significant combinations obtained on the way. This procedure is repeated during many iterations. SNPHarvester may find several different locally optimal combinations due to the initial randomization step.

2.1.1 Model fitting, assessment, and selection

A selected model and its constituent parameters summarize the important properties of a data set. The likelihood function quantifies the model and its parameters, for model fit and assessment. Parameter estimation is usually accomplished through likelihood maximization procedures such as the Newton-Raphson method, Nelder-Mead Simplex methods, or through more computationally intensive stochastic methods such as simulated annealing (Belisle,

1992). These methods can also build decision trees, such as those generated in CART and logic regression.

Most model fitting procedures include some sort of adjustment to -2 times the log-likelihood, specifically between two models that are nested. The closely related Akaike Information Criterion (AIC) (Akaike, 1974) approximates the relative Kullback-Leibler distance between the observed and fitted models. Wald or Likelihood Ratio tests are two possible approaches for testing individual parameters within a model.

Cross-validation is a common technique for assessing the predictive ability of a model that accounts for possible over-fitting. It consists of fitting the predictive model within a subset of observations (training set), then using this model to predict the outcome in another exclusive subset of observations (test set).

2.2 WEAKEST-LINK METHODS

This dissertation is as an extension of the study into the weakest-link algorithm proposed by Richards and Day (Richards, 2002). In a simulation study and analysis of a microarray data set, Richards showed that the weakest-link method was often more powerful in identifying pairwise combinations of covariates predictive of response than conventional linear regression methods. Weakest-link methods were applied in a later study of fibrosis in lung disease (Pardo et al., 2005).

This dissertation extends the practical applications of weakest-link methodology for screening a high-throughput data consisting of a large number of covariates, comparing methods in simulated high-throughput data sets, and reducing the dimensionality of data sets with hierarchical structures.

The weakest-link model was the next step of earlier doctoral work, the conjunctive split model (CosMo) (Shannon, 1995). For binary outcomes, CosMo partitions the covariate space into two different subsets or nodes. Similar to the weakest-link method, partitions the data set according to a locus, with each subset identifying a separate covariate that, alone, affects

the outcome. The partitioning minimizes the node impurity, assessed by the Gini diversity index between each of these subsets.

2.3 CYTOMETRY

Cytometry is a technique to obtain multiplexed data on individual cells, with simultaneous measurement of multiple chemical and physical characteristics for each cell. This technique consists of using cell-based assays to quantify specific proteins or DNA within each cell, in addition to other properties such as size and texture. In cancer, cytometry places cancerous cells stained with a light-sensitive dye under a laser beam for analyzing these biomarkers according to their fluorescence (i.e. the ability of parts of the cells to absorb and re-emit light) of associated antibodies. The data consists of multiparametric vectors of physical and chemical characteristics for each individual cell. Cytometry measures these characteristics by taking a sample of numerous cells (often in the tens or hundreds of thousands per patient). The data from these cell-based studies can be summarized across numerous markers and cells in order to classify subpopulations . The proportions of subpopulations of cells within a patient can detect an interesting mechanism, or predict risk in an individual patient.

The most widely used cytometry method is flow cytometry, consisting of analyzing cells while they are suspended in a stream of fluid. Laser-scanning cytometry is an emerging technology that also uses aspects of fluorescence imaging. The cells are positioned on microscopic slides, where the fluorescence measurements are taken. It has several advantages over the flow cytometry method (Darzynkiewicz et al., 1999). The cells do not need to be immersed in fluid; they can be in tissue, cultured cells, and others. Similarly to flow cytometry, it allows for light scatter and fluorescence measurements, but also records the position of each measurement. Laser-scanning cytometry also records each of the cells' physical characteristics, such as their perimeter and texture. Another advantage is that laser-scanning cytometry requires fewer cells for analysis than flow cytometry, as significantly fewer cells are lost during the process of sample preparation and cell staining.

Gating Gating is the usual procedure for analyzing cytometry data. It is a guided procedure where subpopulations of the cells from a particular sample or patient are visually classified using multidimensional plots. Statistical applications of multiparametric cytometry data include predictions of patient survival (Kern et al., 2004; Nowakowski et al., 2005), response to treatment (Emlet et al., 2007), and risk of progression (Perez-Persona et al., 2007). A vast amount of literature pertaining to visual techniques in cytometry is currently available (Donnenberg and Donnenberg, 2007).

Cluster analysis Clustering consists of unsupervised methods to assign observations into groups (clusters) so that the observations in a cluster share a common trait. A predefined distance measure usually determines the similarity of the observations. Their main application in cell-based studies is to classify cells; these classification results can in turn predict outcomes such as survival or death, or assess a stage of disease or type of disease. Clustering methods such as k-means and hierarchical clustering can detect subpopulations of cells in cytometry (Lo et al., 2008).

Mixture models Mixture models are probability distributions typically modeled as a convex combination of several probability distributions. Identification of cell subpopulations (Boedigheimer and Ferbas, 2008; Wang and Huang, 2007) is one application of mixture models in cancer data. This parametric method has advantages over conventional clustering techniques. Each component parameter may represent a qualitative trait within the cell subpopulation. However, fitting these types of models often requires more complex and computationally-intensive methods.

Support vector machines Machine learning is a branch of artificial intelligence aimed at extracting relevant information from numerous features by finding patterns within the data. The method of support vector machines (SVM) maximizes the margin, or the distance between similarly classified data and a decision hyperplane in non-linear space. After a kernel-based transformation of the data, data classification occurs through linear techniques. Toedling et al. (2006) discussed applications of machine learning for cytometry data, where SVM automates the visual gating process ordinarily used for identifying and labeling cell subpopulations.

2.3.1 Multiplex biomarker data for predicting survival

Intracellular combinations of two of the markers later analyzed in this dissertation, Human Epidermal growth factor Receptor 2 (her-2/neu) and Ras, along with DNA ploidy were predictive of progression-free survival in breast cancer patients (Shackney et al., 2004).

The study demonstrated the importance of preserving the multiplex information within each cell. Patients with an overabundance of *triple-positive* cells, defined as a cell with simultaneous overexpression of all three markers, had significantly lower survival than patients with an overexpression of all the three markers considered separately for each patient, but not an overexpression of triple-positive cells. In other words, tumors with an overabundance ($> 5\%$) of triple-positive cells, defined as cells with simultaneous overexpression of her-2/neu, Ras, and aneuploidy, were associated with significantly shorter recurrence-free survival than tumors with abnormalities of all these quantities considered separately, and not in the same cell: $> 5\%$ cells with abnormal her-2/neu, $> 5\%$ cells with abnormal Ras, and $> 5\%$ cells with aneuploidy.

In other studies, the combination of other proteins, Epidermal growth factor (EGFR) and Vascular endothelial growth factor (VEGF) were significantly related to survival in carcinoma (Gaffney et al., 2003). In another study, cells with VEGF overexpression also had Ras overexpression, but not p53 (Konishi et al., 2000), or Ras and p53 (Fukuyama et al., 1997). The lung cancer study included three other proteins previously been found to be prognostic of survival in lung cancer and other types of cancer, the gene p16, the tumor suppressor protein Rb3, and the protein Cyclin E (Niklinski et al., 2001; Muller-Tidow et al., 2001).

2.4 LIMITATIONS OF CURRENT METHODS

Interaction in conventional linear models is usually represented by a product term between variables $X_1 * X_2$, and is usually only considered if both marginal variables are already included in the model. In addition, overfitting is often an issue in linear models, as each additional main or interaction effects parameter uses up additional degrees of freedom.

Typically, prediction methods such as linear regression or CART have limitations in modeling the effects of higher-order interactions, or combinations of 2 or more variables. This is especially true when available outcome data is limited. The binary predictor covariates required for decision tree methods such as CART and logic regression also rely on cutpoints that could vary greatly due to the data-driven or arbitrary nature of obtaining the best cutpoint. Other methods are applicable to only specific types of data sets; an example is multifactor dimensionality reduction, which relates only to binary outcomes. Many of these methods, such as hierarchical clustering and boosting, have been found to be not robust to noise and outliers.

Logic regression is an example of a decision tree method that identifies variable combinations, but requires the predictor covariates to be dichotomous or be dichotomized. Clustering of classified observations are other common approaches to reducing the dimensionality of data sets with many covariates. However, the result of placing data into discrete categories may also lead to a loss of information. Other methods useful for classification include machine learning techniques such as support vector machines; however, these methods tend to be computationally intensive and novice statisticians may find them difficult to use.

In cytometry, the conventional procedure is to classify each cell into a specific, discrete category, using the information recorded for each cell. This is usually done through visual "gating", where an expert visually classifies cells. Automated methods alternately use clustering or other complicated computational procedures. In immunology, where cell types are well-defined, and other applications, these types of classifications are usually clear. However, in cancer studies, categorizations of cells within a tumor may not be well-defined. In such situations, replacing continuous measurements by category labels reduces the dimensionality of the data, but also results in discarding quantitation and information.

3.0 OBJECTIVES

The primary goal of this dissertation is to provide new practical applications of the weakest-link method for modeling biological interactions, in view of limitations of conventional statistical interaction techniques. In certain types of scenarios, the method should be efficient and easy to implement, while also being as effective as current methods in identifying combinations of variables related with various types of outcomes. As listed in the literature review, there are limitations to current methods in use for assessing biological interaction. We adapt the methodology based on the previously introduced weakest-link paradigm in view of these limitations.

The findings in this dissertation further extend the applications of the weakest-link paradigm. We develop an efficient method to search for combinations of variables, out of hundreds (or thousands) of variables, whose joint effects are associated with an outcome. We adapt the weakest-link methodology to deal with more complex data structures, such as those found in cell-based studies with many observations of variables for each subject and outcome. We evaluate these weakest-link methods by assessing their performance in data sets with evidence of weakest-link properties and in simulation studies.

Objective 1. Large-scale interaction screening: to develop methodology for screening through a set which can consist of thousands of variables, for identifying combinations of two or more variables whose joint effect is associated with an outcome.

Results. An algorithm that screens a large number of variables for subsets of two or more variables, whose joint effects are associated with an outcome variable.

Motivation. Limitations in current methods in screening large data sets for joint effects of variables.

- Additive models take main and interaction effects into account, marginally, and may not represent the true effects of joint combinations of variables.
- Logic regression is limited to detecting combinations of binary covariates.
- Multifactor dimensionality reduction is limited to binary outcomes.

Objective 2. Hierarchical data weakest-link: to expand weakest-link methodology to perform analysis of more complex data structures currently found in biological studies. Specifically, use the weakest-link method to analyze hierarchical data sets consisting of numerous observations of covariate vectors for each single outcome, or patient.

Results. An algorithm which produces derived covariates that simultaneously:

- Uses the weakest-link method to model the joint effect of two or more covariates.
- Aggregates the information from each cell, across all cells of a patient.
- Optionally, dichotomizes patients or cells.

Motivation. Conventional analyses of cell-based studies, such as cytometry, typically require time-consuming gating procedures that require visual input from an expert user. Machine learning techniques such as support vector machines, automate the process of analyzing these types of hierarchical data structures, but are computationally intensive.

4.0 WEAKEST-LINK METHODOLOGY

This chapter describes the weakest-link method and its applications in searching high-throughput data sets for combinations of data.

4.1 WEAKEST-LINK NOTATION

Consider a data set with n subjects, $i = 1, \ldots, n$, each having one outcome labeled, respectively, $Y_i; i = 1, \ldots, n$. The data set consists of p covariates, $k = 1, \ldots, p$, each with a vector of observations of length n: $\mathbf{x}_k = (x_{1k}, \ldots, x_{nk})$ for the k^{th} covariate. This set of vectors comprises the $n \times p$ matrix of covariates $\mathbf{X}$, consisting of observations $x_{ik}; i = 1, \ldots, n; k = 1, \ldots, p$.

Consider a data set with $p = 2$ covariates, $k = 1, 2$. According to the weakest-link model, the expected value of Y given both covariates is:

$$E(Y|X_1, X_2) = \min\left\{\phi_1(X_1), \phi_2(X_2)\right\},$$

where ϕ_1 and ϕ_2 are unknown continuous monotone functions, defined as "weakest-link" functions. We define the locus of optimal use for the weakest-link model as

$$\left\{(a_1, a_2) : \phi_1(a_1) = \phi_2(a_2)\right\}.$$

A curve of optimal use (COU) plot (Figure 1) illustrates the locus of optimal use. The condition $\phi_1(X_1) = \phi_2(X_2)$ is true for all points on the COU locus and plotted as a diagonal

line. Given any point (a_1, a_2) on the COU, the relationship $\phi_1(a_1) = \phi_2(a_2)$=$E(Y|X_1 = a_1, X_2 = a_2)$ holds. If $\phi_2^{-1}(\phi_1(x_1))$ were known, the COU locus is:

$$\left\{x_1, \phi_2^{-1}(\phi_1(x_1)) : x_1 \in \text{ range of } \mathbf{x}_1\right\}.$$

The inverse function on the expected values obtains points on the curve of optimal use (a_1, a_2) :

$$\left\{(a_1, a_2) = \phi_1^{-1}(t), \phi_2^{-1}(t) : t \in \text{ range of } E(Y)\right\}$$

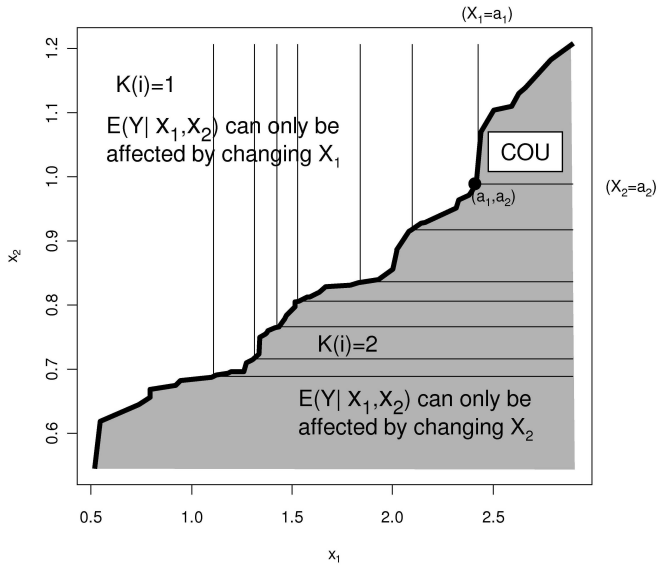

Figure 1: Curve of optimal use (COU) plot for 2 covariates

Curve of optimal use (COU) plot for 2 covariates. The curve of optimal use (COU) consists of points (a_1, a_2) at which the expected value $E(Y|X_1, X_2) = \min\{\phi_1(X_1), \phi_2(X_2)\}$ can change by decreasing either X_1 or X_2. At all points in the white area, $E(Y|X_1, X_2)$ can only change by altering X_1. At all points in the shaded area, $E(Y|X_1, X_2)$ can only change by altering X_2. Each contour line plots equal values of $E(Y|X_1, X_2)$.

On all points along this COU, the expected value of the outcome $E(Y|X_1, X_2)$ changes if either X_1 or X_2 decreases, but not if either increases. The COU locus, furthermore, partitions the data space into two subsets. At all points to the left, or above the COU locus, $E(Y|X_1, X_2)$ can only change by altering X_1; altering X_2 has no effect. At all points to the

right, or below this COU locus, $E(Y|X_1, X_2)$ can only change by altering X_2; changing X_1 has no effect.

Let $K(i) = \arg\min \{\phi_1(x_{i1}), \phi_2(x_{i2}); k = 1, 2\}$ for $i = 1, \ldots, n$, for points not along the COU. Define $K(i)$ as the index of the weakest-link covariate for observation i, and $x_{iK(i)}$ as the weakest-link covariate. $K(i)$ identifies the subset of the i^{th} observation, or its location relative to the COU. When point (x_{i1}, x_{i2}) is on the locus of optimal use, $K(i)$ is not uniquely defined because $\phi_1(x_{i1}) = \phi_2(x_{i2})$.

According to the weakest-link model:

$$E(Y_i|X_{i1} = x_{i1}, X_{i2} = x_{i2}) = \phi_{K(i)}(x_{iK(i)}),$$

given continuous monotone functions ϕ_1 and ϕ_2, with inverse functions ϕ_1^{-1} and ϕ_2^{-1}, a corresponding weakest-link covariate X^* can be expressed as:

- $\min\left\{X_1, \phi_1^{-1}(\phi_2(X_2))\right\}$, if $\phi_1(x)$ is monotone increasing with respect to x
- $\max\left\{X_1, \phi_1^{-1}(\phi_2(X_2))\right\}$, if $\phi_1(x)$ is monotone decreasing with respect to x.

In either case, the expected value $E(Y|X_1, X_2)$ will be monotone with respect to this weakest-link covariate X^*.

We prove the previous statement as follows:

Theorem Suppose ϕ_1 is a monotone increasing function. We define

$$X^* = \min\left\{X_1, \phi_1^{-1}(\phi_2(X_2))\right\}.$$

The expected value $E(Y|X_1, X_2) = \min\{\phi_1(X_1), \phi_2(X_2)\}$ is monotone increasing with respect to X^*.

Proof

$$\begin{aligned} E(Y|X_1, X_2) &= \min\{\phi_1(X_1), \phi_2(X_2)\} \\ &= \min\left\{\phi_1(X_1), \phi_1(\phi_1^{-1}(\phi_2(X_2)))\right\} \end{aligned}$$

Because ϕ_1 is monotone increasing, its inverse ϕ_1^{-1} is also monotone increasing. Therefore, $\phi_1(X_1) < \phi_1(\phi_1^{-1}(\phi_2(X_2)))$ is true if and only if $X_1 < \phi_1^{-1}(\phi_2(X_2))$.

$$\begin{aligned}\min\left\{\phi_1(X_1), \phi_1(\phi_1^{-1}(\phi_2(X_2)))\right\} &= \phi_1(\min\left\{X_1, \phi_1^{-1}(\phi_2(X_2))\right\}) \\ &= \phi_1(X^*).\end{aligned}$$

Because ϕ_1 is a monotone increasing function, $E(Y|X_1, X_2)$ is monotone increasing with respect to X^*.[**QED**]

Theorem Suppose ϕ_1 is a monotone decreasing function, and we define

$$X^* = \max\left\{X_1, \phi_1^{-1}(\phi_2(X_2))\right\}.$$

The expected value $E(Y|X_1, X_2) = \max\left\{\phi_1(X_1), \phi_2(X_2)\right\}$ is monotone decreasing with respect to X^*.

Proof

$$\begin{aligned}E(Y|X_1, X_2) &= \min\left\{\phi_1(X_1), \phi_2(X_2)\right\} \\ &= \min\left\{\phi_1(X_1), \phi_1(\phi_1^{-1}(\phi_2(X_2)))\right\}\end{aligned}$$

Because ϕ_1 is monotone decreasing, its inverse ϕ_1^{-1} is also monotone decreasing. Therefore, $\phi_1(X_1) < \phi_1(\phi_1^{-1}(\phi_2(X_2)))$ is true if and only if $X_1 > \phi_1^{-1}(\phi_2(X_2))$.

$$\begin{aligned}\min\left\{\phi_1(X_1), \phi_1(\phi_1^{-1}(\phi_2(X_2)))\right\} &= \phi_1(\max\left\{X_1, \phi_1^{-1}(\phi_2(X_2))\right\}) \\ &= \phi_1(X^*).\end{aligned}$$

Because ϕ_1 is a monotone decreasing function, $E(Y|X_1, X_2)$ is monotone decreasing with respect to X^*.[**QED**]

Generally, for $p \geq 2$, consider a subset of the covariate matrix $\mathbf{X}$, consisting of $p^* \leq p$ covariates. Let $\kappa = (\kappa_1, \ldots, \kappa_{p^*})$ be a set of column indices of $\mathbf{X}$ in the weakest-link model. According to the weakest-link model, the expected value of outcome Y given the subset of covariates with indices κ is:

$$\begin{aligned}E(Y_i|X_{i\kappa_1} = x_{i\kappa_1}, \ldots, X_{i\kappa_{p^*}} = x_{i\kappa_{p^*}}) &= \min\left\{\phi_{\kappa_1}(x_{i\kappa_1}), \ldots, \phi_{\kappa_{p^*}}(x_{i\kappa_{p^*}})\right\} \\ &= \min_{k\in\kappa}\left\{\phi_k(x_{ik})\right\}\end{aligned}$$

where $\min_{k \in \kappa}$ indicates the minimum across covariates with indices k in the subset κ.

We can also summarize the weakest-link model through the expression:

$$E(Y_i | X_{i\kappa_1} = x_{i\kappa_1}, \ldots, X_{i\kappa_{p^*}} = x_{i\kappa_{p^*}}) = \phi_{K(i)}(x_{iK(i)}),$$

where $K(i) = \arg\min \{\phi_{\kappa_{k^*}}(x_{i\kappa_{k^*}}); k^* = 1, \ldots, p^*\}$ for $i = 1, \ldots, n$.

The result of the function $\phi_{K(i)}(x_{iK(i)})$ is one covariate, taken from column $K(i)$ for each row (or patient) $i = 1, \ldots, n$ of the matrix $\mathbf{X}$.

Replacing the *minimum* function in the weakest-link expression with the *maximum* function results in another family of functions similar to the weakest-link model, expressed as: $E(Y|X_1, X_2) = \max\{\phi_1(X_1), \phi_2(X_2)\}$. Section 4.2.2 contains a discussion of this general family of weakest-link models.

4.2 STITCHED WEAKEST-LINK MODELS

If the true COU locus were known for a particular set of covariates, fitting the weakest-link model would be extremely efficient. Any parametrization of the curve would provide a derived covariate for fitting a weakest-link model. With the curve parameter as the lone covariate, a univariate regression would then be the only step required for model fitting. However, in practical situations the true COU locus is always unknown.

The following non-parametric approach accelerates the weakest-link model fits in screening a higher-throughput data set. We can postulate a specific COU locus, which helps in quickly screening through a large data set for combinations of variables that are strongly associated with the outcome through the weakest-link relationship.

We can parametrize the COU locus by "stitching" values of the two (or more) covariates included in the weakest-link relationship. On the COU, covariates are stitched together in the sense that they all contribute to the outcome. We choose a set of convenient monotone functions $(\omega_1, \ldots, \omega_p)$ to postulate a COU locus in a systematic fashion, where

$$\omega_k : \mathbb{X}_k \mapsto \Re$$

and $\mathbb{X}_k$ is the sample space for covariates X_k.

We refer to this set of functions ω_k as stitching functions; they determine a specific COU locus. We postulate the COU locus:

$$t \mapsto (\omega_1^{-1}(t), \ldots, \omega_p^{-1}(t)).$$

At any point $(a_1, \ldots, a_p)$ on the COU:

$$\omega_1(a_1) = \ldots = \omega_p(a_p).$$

In this sense, according to the weakest-link model $X_1 = a_1$ is equivalent to $X_2 = a_2$; we think of the X_1 and X_2 scales being stitched together, at points where a_1 is stitched to a_2.

The ϕ_k themselves still need to be estimated. For p^* covariates with indices $\kappa = (\kappa_1, \ldots, \kappa_{p^*})$, define t_i for patient i to be the derived covariate:

$$\begin{aligned} t_i &= \min_{k \in \kappa} \{\omega_k(x_{ik})\} \\ &= \min \left\{\omega_{\kappa_1}(x_{i\kappa_1}), \ldots, \omega_{\kappa_{p^*}}(x_{i\kappa_{p^*}})\right\}. \end{aligned}$$

Two straightforward ways to define ω are:

- Quantile stitching: $\omega_k(x_{ik}) = \hat{F}_k(x_{ik})$
- Probit stitching: $\omega_k(x_{ik}) = \hat{F}_{norm,k}(x_{ik}) = \Phi(\frac{x_{ik} - \hat{\mu}_k}{\hat{\sigma}_k})$

We describe these procedures below.

4.2.0.1 Quantile stitching Richards (2002) chose the empirical cumulative distribution function (CDF) as the stitching function, $\omega_k(x_{ik}) = \underset{i=1:n}{\hat{F}_k}(x_{ik})$. The k subscript indicates that the empirical distribution function is being taken for the k^{th} covariate, across all $i = 1, \ldots, n$ patients. For conciseness, we can define $\hat{F}_k(x_{ik}) = \underset{i=1:n}{\hat{F}_k}(x_{ik})$ if the empirical cdf is relative to all n patients. We define the empirical cdf as: $\underset{i=1:n}{\hat{F}_k}(x_{ik}) = \frac{rank(x_{ik})-1}{n-1}$, where $rank(x_{ik})$ is the rank (in ascending order) of x_{ik} within all patients $i = 1, \ldots, n$ for the k^{th} covariate.

The Curve of Optimal Use (COU) plot (Figure 2) illustrates the quantile stitching weakest-link model for a bivariate data set (X_1, X_2). For patient i, the quantile stitched derived covariate is $t_i = \min\left\{\hat{F}_1(x_{i1}), \hat{F}_2(x_{i2})\right\}$. Points on the same contour line have equal values of t_i. The lines going from the lower-left to the upper-right of both plots are the curve of optimal use, where $\hat{F}_1(x_{i1}) = \hat{F}_2(x_{i2})$. On the $\mathbf{x}$ plot the COU is a jagged line, while onto the plot of covariate set $\mathbf{t}$ the COU is a straight line. At all points along this COU, $t_i = \omega_1(x_{i1}) = \omega_2(x_{i2})$; decreasing either X_1 or X_2 results in changing the derived covariate t. At points above or to the right of this COU, the derived covariate t can only be changed by altering X_1, while at points below or left of this COU, only X_2 can alter t.

For $p = 2$ covariates, the points of the COU are:

$$\left\{(\hat{F}_1^{-1}(t), \hat{F}_2^{-1}(t)) : t \in [0,1]\right\}.$$

4.2.0.2 Probit stitching We introduce another approach to stitching the covariates. Probit stitching normalizes the covariates to the standard normal distribution $N(0,1)$, and then takes the probit of these standard normal variables to have the range on the unit interval, as they are for quantile stitching. To accomplish this, we define the stitched weakest-link function as $\hat{F}_{norm,k}(x_k) = \Phi\left[\frac{x_k-\hat{\mu}_k}{\hat{\sigma}_k}\right] = \Phi\left[\frac{x_k-\overline{x}_k}{s(x_k)}\right]$, where $\overline{x}_k$ and $s(x_k)$ are, respectively, the empirical mean and standard deviation over all m observations for covariate k, and Φ is the probit function. Define the function $\hat{F}^{(-)}_{norm,k}(x_k) = 1 - \hat{F}_{norm,k}(x_k)$ in a similar manner to quantile stitching. Notation and likelihood maximization procedures are similar to those for quantile stitching.

For two dimensions, we obtain the following curve of optimal use locus:

$$\left\{(\hat{\mu}_1 + \hat{\sigma}_1 \cdot \Phi^{-1}(t), \hat{\mu}_2 + \hat{\sigma}_2 \cdot \Phi^{-1}(t)) : t \in [0,1]\right\}.$$

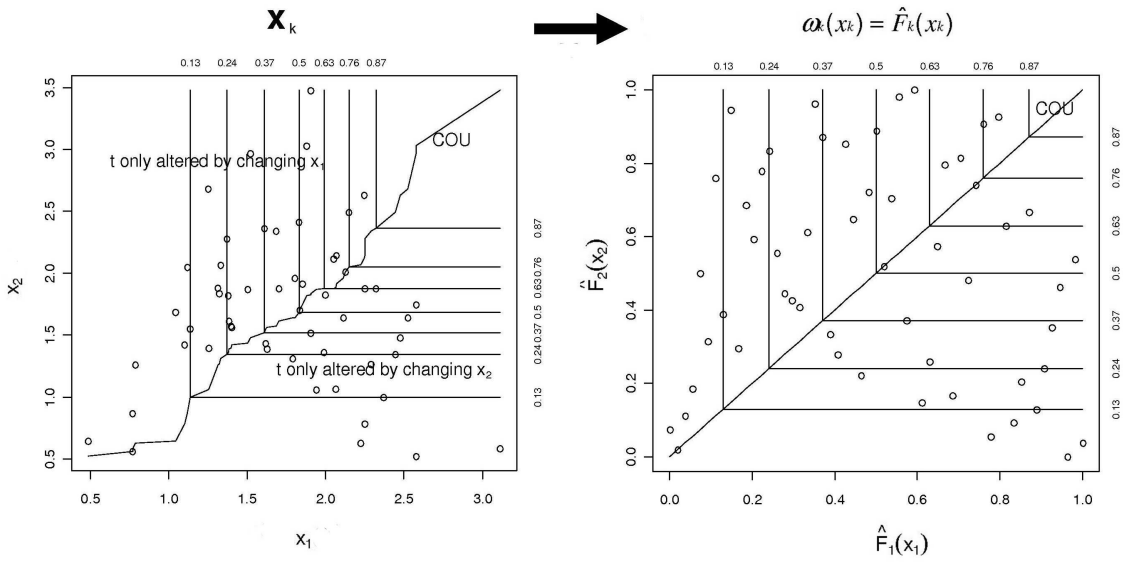

Figure 2: Curve of optimal use (COU) plot for quantile stitching

Curve of optimal use (COU) plot for quantile stitching. Plot of original data ($\mathbf{x}$) (left), mapping onto $t_i = \min\left\{\hat{F}_1(x_{i1}), \hat{F}_2(x_{i2})\right\}$ (right). Contour lines plot equal values of $t_i = \min\left\{\hat{F}_1(x_{i1}), \hat{F}_2(x_{i2})\right\}$.

Unlike the COU locus from quantile stitching, the COU from probit stitching is a straight line (Figure 3):

$$\left\{(x_1, (\bar{x}_2 - \tfrac{\hat{\sigma}_2}{\hat{\sigma}_1}\bar{x}_1) + (\tfrac{\hat{\sigma}_2}{\hat{\sigma}_1})x_1) : x_1 \in \mathbb{X}_1\right\}$$

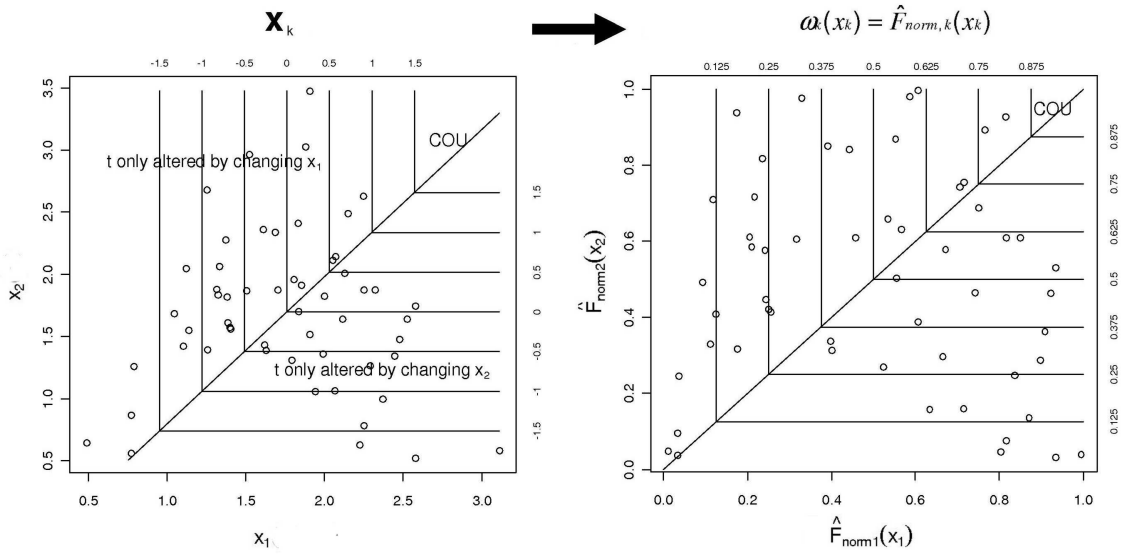

Figure 3: Curve of optimal use (COU) plot for probit stitching

Curve of optimal use (COU) plot for probit stitching. Plot of original data ($\mathbf{x}$) (left), with the data after transformation onto the probit space $t_i = \min\left\{\hat{F}_{norm,1}(x_{i1}), \hat{F}_{norm,2}(x_{i2})\right\}$ (right). The diagonal line is the COU locus. Points on the same contour line have equal values of $t_i = \min\left\{\hat{F}_{norm,1}(x_{i1}), \hat{F}_{norm,2}(x_{i2})\right\}$.

Probit stitching, with computational complexity of $O(n)$ is more efficient than quantile stitching, which requires a sorting algorithm of complexity $O(n^2)$. However, a non-linear transformation of the data more easily changes the locus from probit stitching compared to quantile stitching.

The following sections describe the procedures used in optimizing the quantile stitching weakest-link, though we can easily generalize them to probit stitching or to any other definitions of the stitching functions ω_k.

4.2.1 Maximum likelihood estimation for weakest-link models

The stitched weakest-link procedure described in the previous section obtains a single covariate vector $t_i = \min_{k \in \kappa} \{\omega_k(x_{ik})\}; i = 1, \ldots, n$. It is important to note that the stitching functions ω_k are chosen for postulating a COU locus, rather than to model expected values $E(Y)$, which are directly obtained from the functions ϕ_k.

With this in mind, we can define a set of parameters, θ, to relate the derived covariate observations $t_1, \ldots, t_n$ to the outcomes $Y_1, \ldots, Y_n$. A simple approach is to use a conventional procedure such as linear regression to estimate a parameter set $\theta = (\alpha, \beta)$, with α being an intercept parameter and β being a slope parameter. For example, a generalized linear model with parameter set θ can estimate the relationship between the expected value $g[E(Y_i)]$ and the derived covariates t_i:

$$\begin{aligned} g\left[E(Y_i | X_{i\kappa_1} = x_{i\kappa_1}, \ldots, X_{i\kappa_{p^*}} = x_{i\kappa_{p^*}})\right] &= (1 \min_{k \in \kappa} \{\omega_k(x_{ik})\}) \cdot \theta^T \\ &= \alpha + \beta t_i \end{aligned}$$

where $g(u)$ is a suitable link function, for patients $i = 1, \ldots, n$.

Define $\phi(\mathbf{x}, \omega; \theta) = (1 \min_{k \in \kappa} \{\omega_k(x_k)\}) \cdot \theta^T = \alpha + \beta t$ as the dot product of the derived covariates obtained from quantile stitching, and regression parameters θ. This result is the vector of expected outcome values $g[E(Y)] = (g[E(Y_1)], \ldots, g[E(Y_n)])$.

Define $L(x, y, \omega; \theta)$ as the likelihood function from a set of derived covariates and the corresponding regression parameters.

For binary outcomes, we obtain maximum likelihood estimates θ for the original derived covariates $t = \min_{k \in \kappa} \{\omega_k(x_k)\}$ by maximizing the following expression:

$$L(x, y, \omega; \theta) = \prod_{i=1}^{n} \left[\phi(x_i, \omega; \theta)\right]^{y_i} \left[1 - \phi(x_i, \omega; \theta)\right]^{1 - y_i}.$$

For survival analysis, two approximations of the partial likelihood corresponding to the Cox proportional hazards model are Breslow's model (Breslow, 1974), and Efron's model (Efron, 1977). The expression of Efron's approximation is:

$$L(x, y, \omega; \theta) = \prod_{h=1}^{E} \frac{\exp\left(\phi(\mathbf{x}_h, \omega; \theta)\right)}{\prod_{j=1}^{e_h} \left[\sum_{i \in R_h} \exp\left(\phi(x_i, \omega; \theta)\right) - \frac{j-1}{e_h} \sum_{i \in \mathtt{E}_h} \exp\left(\phi(x_i, \omega; \theta)\right)\right]},$$

where $y_1, y_2, \ldots, y_E$ are the E distinct event times in the data set, h is the index of distinct event times, e_h is the number of events at time y_h, $\mathtt{E}_h$ is the set of individuals with an event at time y_h, R_h is the set of all individuals at risk prior to y_h, θ is the vector of coefficients, $\phi(x_i, \omega; \theta) = \theta^T \min_{k \in \kappa} \{\omega_k(x_{ik})\}$, and $\phi(\mathbf{x}_h, \omega; \theta) = \theta^T \mathbf{s_h}$, where $\mathbf{s}_h$ is the sum of the covariate vectors $\min_{k \in \kappa} \{\omega_k(x_{ik})\}$ over all individuals who had an event at y_h.

4.2.2 Directionality in stitched weakest-link methods

An assumption of the weakest-link model is that the unknown component functions ϕ_k are monotone. However, also unknown is whether or not these functions are monotone increasing or decreasing. The following procedures allow us to consider a general family of stitched weakest-link models.

We use the following labeling convention for each stitched weakest-link model and the directionality of their corresponding component functions. Let $\omega_k^{(+)}(x_k)$ and $\omega_k^{(-)}(x_k)$ be two stitching functions for hypothesizing the COU, when ϕ_k is unknown. Let $\omega_k^{(+)}(x_k)$ be monotone increasing, to account for the possibility that $\phi_k(x_k)$ is monotone increasing. Similarly, let $\omega_k^{(-)}(x_k)$ be a monotone decreasing function. If the weakest-link model includes $\omega_k^{(+)}(x_k)$, the derived covariate t_i is monotone increasing with respect to x_k; if the model includes $\omega_k^{(-)}(x_k)$, t_i is monotone decreasing with respect to x_k.

A set of two covariates result in two different loci of optimal use. The points (x_1, x_2) of the two possible COU loci are:

$$\text{(Locus 1)} : \left(x_1, \omega_2^{-1(+)}(\omega_1^{(+)}(x_1))\right)$$
$$\text{(Locus 2)} : \left(x_1, \omega_2^{-1(+)}(\omega_1^{(-)}(x_1))\right)$$

Let D be a sequence of length p^* consisting of symbols '+' or '-'. Let $D(k^*)$ be the symbol in the k^*th position of D, denoting the inclusion of either $\omega_k^{(+)}(x_k)$ or $\omega_k^{(-)}(x_k)$ in the weakest-link model. Define $\text{wl}^{(D)}(X) = \min(\omega_1^{(D(1))}(X_1), \omega_2^{(D(2))}(X_2))$ when $D_1, D_2 \in \{+, -\}$ as the stitched weakest-link function between two covariates.

For example, for the weakest-link between two covariates and $D = +-$, $D(1)$ is '+' and $D(2)$ is '-'. The notation for the weakest-link model is:

$$\begin{aligned}\text{wl}^{(D)}\left\{\omega_1(x_1), \omega_2(x_2)\right\} &= \min\left\{\omega_1^{(D(1))}(x_1), \omega_2^{(D(2))}(x_2)\right\} \\ &= \min\left\{\omega_1^{(+)}(x_1), \omega_2^{(-)}(x_2)\right\}.\end{aligned}$$

The family of weakest-link models accounting for directionality are illustrated in the COU graphs in Figure 4, for the weakest-link between $p^* = 2$ covariates taken from quantile stitching: $t_i = \left(\hat{F}_1(x_{i1}), \hat{F}_2(x_{i2})\right)$. We define $\omega_k^{(+)}(x_k) = \hat{F}_k^{(+)}(x_k) = \hat{F}_k(x_k)$ and $\omega_k^{(-)}(x_k) = \hat{F}_k^{(-)}(x_k) = 1 - \hat{F}_k(x_k)$. We can also think of $\hat{F}^{(+)}(x)$ as taking the CDF of x in ascending order, with $\hat{F}^{(-)}(x)$ taken in descending order. Define $\text{wl}^{(D)}\left\{\hat{F}_1(x_1), \hat{F}_2(x_2)\right\}$ as the weakest-link function, between 2 covariates.

For 2 covariates, there are 4 different directions of quantile stitched covariates. The first two sets of derived covariates $t^{(++)}$ and $t^{(--)}$ are for the first COU $\left(x_1, \omega_2^{-1(+)}(\omega_1(x_1)^{(+)})\right)$, the third and fourth sets $t^{(-+)}$ and $t^{(+-)}$ for the other COU $\left(x_1, \omega_2^{-1(+)}(\omega_1(x_1)^{(-)})\right)$.

$$\begin{aligned}(1): t^{(++)} &= \text{wl}^{(++)}\left\{\hat{F}(x)\right\} = \min\left\{\hat{F}_1(x_1), \hat{F}_2(x_2)\right\} \\ (2): t^{(--)} &= \text{wl}^{(--)}\left\{\hat{F}(x)\right\} = \min\left\{1 - \hat{F}_1(x_1), 1 - \hat{F}_2(x_2)\right\} \\ (3): t^{(-+)} &= \text{wl}^{(-+)}\left\{\hat{F}(x)\right\} = \min\left\{1 - \hat{F}_1(x_1), \hat{F}_2(x_2)\right\} \\ (4): t^{(+-)} &= \text{wl}^{(+-)}\left\{\hat{F}(x)\right\} = \min\left\{\hat{F}_1(x_1), 1 - \hat{F}_2(x_2)\right\}.\end{aligned}$$

For example, quantile stitching satisfies the property $\hat{F}^{(-)}(X) = 1 - \hat{F}(X)$. This accounts for the maximum relationship, $\max\left\{\hat{F}_1(X_1), \hat{F}_2(X_2)\right\}$, between both covariates:

$$\begin{aligned}\min\left\{F_1^{(-)}(X_1), F_2^{(-)}(X_2)\right\} &= \min\left\{1 - F_1(X_1), 1 - F_2(X_2)\right\} \\ &= 1 + \min\left\{-F_1(X_1), -F_2(X_2)\right\} \\ &= 1 - \max\left\{F_1(X_1), F_2(X_2)\right\} \\ &= 1 - \max\left\{F_1^{(+)}(X_1), F_2^{(+)}(X_2)\right\}\end{aligned}$$

Choosing the optimal directionality of stitched weakest-link models The stitched weakest-link procedure described in this section reduces the predictors to a single covariate

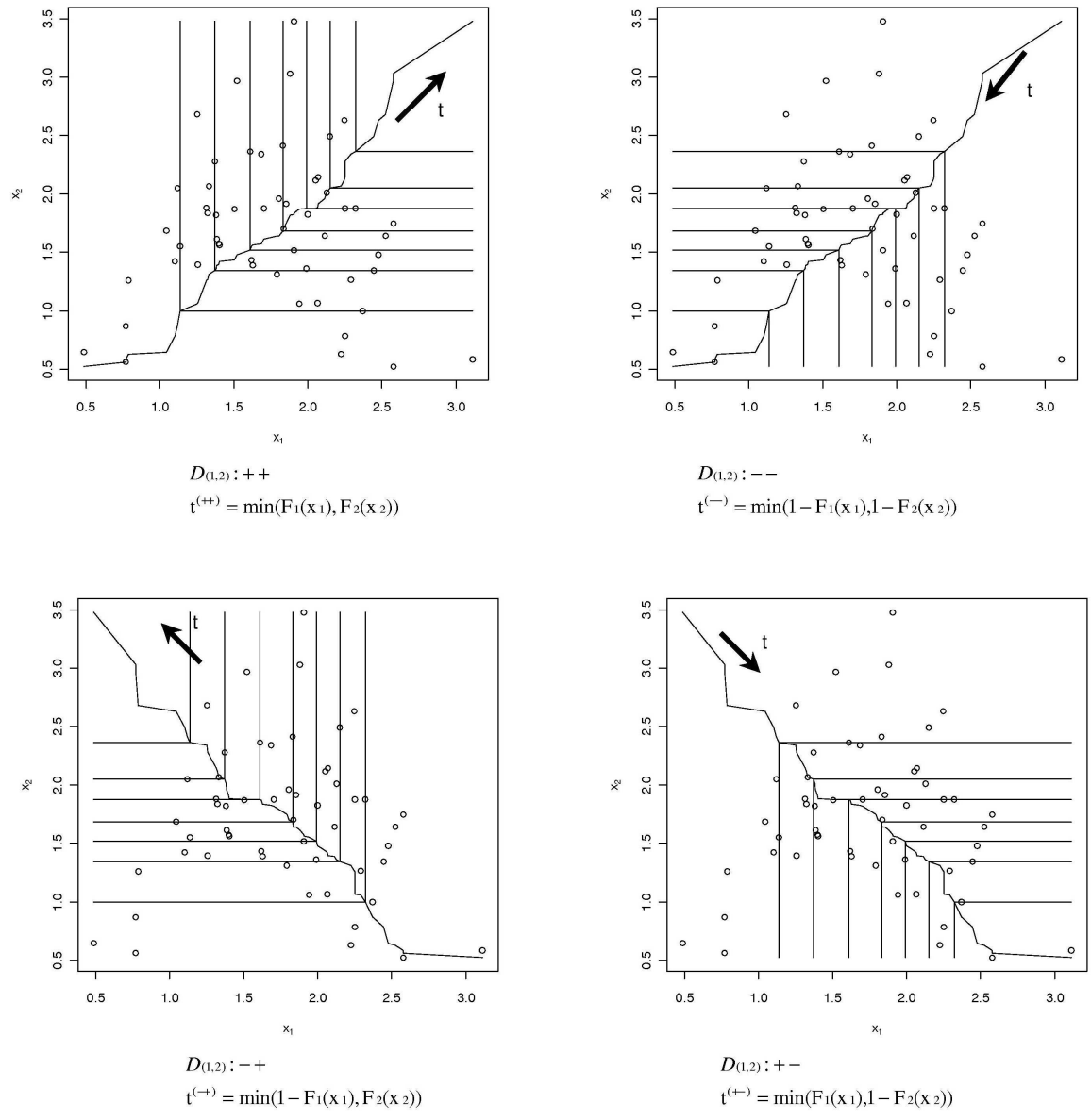

Figure 4: Curve of optimal use (COU) plots for weakest-link directionalities

Curve of optimal use (COU) plots showing four possible combinations of pairwise derived covariates. The two top panels have the same COU $\left(x_1, \hat{F}_2^{-1}(\hat{F}_1(x_1))\right)$; reversing the sign of both wl functions changes the direction in which t increases. The bottom figure corresponds to another COU $\left(x_1, \hat{F}_2^{-1}(1-\hat{F}_1(x_1))\right)$ generated from the same data set by changing the sign of one of the wl functions, opposing directions of increasing t obtains two different sets of covariates.

$t_i = \text{wl}^{(D)}(\omega_i(x_i))$ for $i = 1, \ldots, n$ which we call the *weakest-link* derived covariate across a set of covariates k.

Suppose we assess the weakest-link between p^* covariates though n observations. In general, there are 2^{p^*-1} different COUs for p^* covariates, and two different sets of directions for each COU. For p^* covariates, there are 2^{p^*} sets of stitched covariates. For $p^* > 2$ covariates, we express the weakest-link model for each combination (D):

$$\underset{k\in\kappa}{\text{wl}}^{(D)} \{\omega_k(x_{ik})\} = \min \left\{ \omega_{\kappa_1}^{(D(1))}(x_{i\kappa_1}), \ldots, \omega_{\kappa_{p^*}}^{(D(p^*))}(x_{i\kappa_{p^*}}) \right\}.$$

We denote the derived covariate $\underset{k\in\kappa}{\text{wl}}^{(D)} \{\omega_k(x_{ik})\}$ as the *weakest-link* across a set of covariates $k \in \kappa$.

For each of the 2^{p^*} stitched covariates $t^{(D)}$, there is a corresponding likelihood function $L(x, y, \omega; \theta, D)$ and a set of maximum likelihood estimates, $\hat{\theta}^{(D)}$.

We can use any parametric or non-parametric univariate regression method to relate $t^{(D)}$ to the outcomes Y. The final chosen direction (D) provides the best fit, amongst the 2^{p^*} models, according to a conventional model assessment criterion. Typical criteria include minimizing the negative log-likelihood for regression models or partial log-likelihood for proportional hazards models.

4.3 MODEL ASSESSMENT AND SELECTION THROUGH CROSS-VALIDATION

Cross-validation is a method to assess the predictive value of a model. Cross-validation partitions the set of observations into two exclusive sets: the training set and the test set. The model is initially fit using the outcomes and corresponding covariates in the training set. This fitted model then incorporates covariates from observations in the test set to obtain a corresponding set of predicted outcomes. These predicted outcomes are compared to the observed outcomes and quantified according to a scoring criterion such as residual sum of squares, for linear regression, or misclassification rate, for classification problems of binary or other categorical data. Model selection proceeds by minimizing these scoring criteria.

Leave one-out cross-validation uses a single observation as the test set, with the remaining observations being the training set, and repeats the procedure once for all observations in the original data set.

Verweij and Houwelingen (1993) developed a version of cross-validation for survival analysis. Let $l_i(\beta)$ be the individual contribution of the i^{th} observation to the partial log-likelihood:

$$l_i(\beta) = l_{(-i)}(\beta) - l(\beta),$$

where $l_{(-i)}(\beta)$ is the partial log-likelihood when the i^{th} observation is excluded, and $\hat{\beta}_{(-i)}$ is the corresponding maximum likelihood estimate. The leave one-out cross-validated partial log-likelihood cvl is the sum of the contributions from all observations:

$$cvl = \sum_{i=1}^{n} l_i(\hat{\beta}_{(-i)}).$$

4.4 VARIABLE COMBINATION SCREENING FOR HIGH-THROUGHPUT DATA SETS

For high-throughput data sets with hundreds or thousands of covariates, the combinatorial explosion of possible covariate pairs makes exhaustive searches not feasible. A simple method of reducing this combinatorial explosion is to simply concentrate on a subset of covariates in the search. Alternately, stochastic methods are other possible approaches for searching through very large data sets.

4.4.1 Filtered subsets of covariates

When computational time is prohibitive due to the number of covariates in some types of data sets (more than 100 covariates), a simple procedure is to obtain a reduced subset, after removing covariates with no or low marginal associations with the outcomes.

However, we risk ignoring an interesting combination if we do not consider covariates that are not marginally associated with the outcome. In fact, it is possible for a covariate of zero correlation with an outcome be included in a weakest-link pair that is perfectly associated

with the same outcome. Consider the following simple example, where the weakest-link between X_1 and X_2 is perfectly correlated with the outcome Y, even though the respective correlations between these individual covariates and the outcome are zero (Table 1):

Table 1: Weakest-link without main effects example

i	1	2	3	4
X_1	1	-1	0	0
X_2	-1	1	0	0
$\min(X_1, X_2)$	-1	-1	0	0
Y	0	0	1	1

Example: weakest-link of two covariates significantly associated with outcome; both covariates have no marginal association.
$\text{corr}(X_1, Y) = 0$ and $\text{corr}(X_2, Y) = 0$
$\text{corr}\left(\min\left\{X_1, X_2\right\}, Y\right) = 1$

From a practical point of view, filtering out covariates that show no marginal association may still improve the detection of interesting covariate combinations. Simulation studies in Section 6.2.2 look for situations where filtering improves detection and computational time.

4.4.2 Greedy search algorithm for locally optimal combinations

Many methods, such as CART or stepwise regression, use greedy algorithms to add or delete individual covariates out of many to a model. We apply a greedy approach to variable screening in this section.

4.4.2.1 Finding a locally optimal combination We introduce a modification of the PathSeeker algorithm used in SNPHarvester (Yang et al., 2009) to find a locally optimal combination of covariates associated with the outcome. The algorithm switches out covariates within a subset, one at a time, until convergence is attained.

Suppose we have a matrix of covariates, were X_k is the k^{th} column vector in the data set. Let A define a subset of covariates which we refer to as the *active* set, consisting of p^* covariates $A = \left\{X_{\kappa_1}, \ldots, X_{\kappa_{p^*}}\right\}$ denoted by column labels $\kappa = \{\kappa_1, \ldots, \kappa_{p^*}\}$.

Let C be a locally optimal combination of covariates. We obtain C by minimizing a typical scoring criterion $Score(A)$, for example -log LL for a generalized linear model

or misclassification for categorical outcomes. This is accomplished by replacing only one covariate in active set A at a time.

Algorithm LocalOne

Step 1 We first initialize the active set A, by generating a random subset of column indices, κ, of length p^*: $\kappa = \{\kappa_1, \ldots, \kappa_{p^*}\}$.

Step 2 We apply the following procedure:

For k in 1 to $p; k \notin \kappa$:

For j in 1 to p^*:

Form a proposal set $A^* = \left\{X_{\kappa_1}, \ldots, X_k, X_{\kappa_{j+1}}, \ldots, X_{\kappa_{p^*}}\right\}$, and calculate the criterion $Score(A^*)$.

If the criterion from any of these proposal sets A^* improves relative to $Score(A)$ then set $A \leftarrow A^*$, otherwise leave A unchanged.

Separately, if we are also interested in finding many interesting combinations instead of just the locally optimal one, we also may take note of proposal set A^* if it meets a simple criterion, such as a Bonferroni-adjusted p-value below a threshold.

Step 3 We repeat **Step 2** until the active set A does not change within that step. When this occurs, we determine that convergence is attained and mark the active subset A to be a locally optimal combination of covariates with respect to the outcome.

4.4.2.2 Finding multiple combinations

For detecting multiple local minima, we use the procedure **LocalOne** outlined in subsection 4.4.2.1 repeatedly, while looping over multiple random starting points, to find each local minimum.

Algorithm LocalMany Let C be a set of locally optimal combinations of p^* covariates, and let $\kappa_{LO}(l)$ be one of these combinations, where l is an index denoting the l^{th} locally optimal combination in C. Use the greedy algorithm from 4.4.2.1 to find $\kappa_{LO}(l)$, or the active combination of covariates after convergence.

We can truncate the search by removing combinations that already have been tested to be locally optimal. After finding each locally optimal combination, we redo the greedy search and remove covariates with indices $k = \{\kappa_{LO}(l)_1, \ldots, \kappa_{LO}(l)_{p^*}\}$ from further consideration. In other words, for the next locally optimal combination $\kappa_{LO}(l{+}1)$ we search within covariates

$k \notin (\kappa_{LO}(1), \ldots, \kappa_{LO}(l))$. If $p^* = 2$, this increases efficiency without any reduction in search scope; each locally combination $\kappa_{LO}(l)$ cannot be improved by switching out one covariate in the combination. However if $p^* > 2$, there is a possible reduction in search scope that may or may not be acceptable.

We repeat the process, finding multiple locally optimal combinations, until either l_{max} locally optimal combinations, or l_{cons} consecutive locally optimal combinations not passing a statistical threshold, such as a Bonferroni-adjusted p-value < 0.05.

4.4.3 Simulated annealing procedures

There will likely be numerous locally optimal combinations of covariates in large-scale data sets found in microarray and proteomics studies. Thus, it would be difficult to determine if a given locally optimal combination is truly a global optimum. Likewise, it would be computationally intensive to compute all possible combinations. Simulated annealing (Aarts et al., 2005) procedures are heuristics whose aim is usually to find globally optimal solutions over functions with rough surface plots.

We apply a procedure similar to the simulated annealing procedure described by Belisle (1992) in order to find a globally optimal combination of covariates in this large-combinatorial space. We use a similar procedure to the previously described greedy procedure **LocalOne**. Suppose our goal is to find a globally optimal combination of p^* covariates out of p covariates, by minimizing a scoring function $Score(X)$.

Step 1 We randomly sample, without replacement, all indices from $k = 1, \ldots, p$, forming a new vector of indices of length p $(K_1, \ldots, K_p)$ with a different order than the original vector.

Step 2 We initialize the active set A by obtaining a subset of p^* covariates: $A = X_{\kappa_1}, \ldots, X_{\kappa_{p^*}}$. Record $Score(A)$.

Step 3 For k in K_1 to $K_p; k \notin \kappa$ we repeat the following procedure.

Randomly select j from $1 : p^*$. Form proposal set $A^* = \left\{X_{\kappa_1}, \ldots, X_k, X_{\kappa_{j+1}}, \ldots, X_{\kappa_{p^*}}\right\}$, and calculate the criterion $Score(A^*)$. If $Score(A^*) < Score(A)$, then let $A \leftarrow A^*$. Otherwise, if $Score(A^*) \geq Score(A)$, set $A \leftarrow A^*$ according to the acceptance probability from the Metropolis function $\min\{\exp(\Delta/temp), 1\}$, where Δ is the difference between the scores and $temp$ is the temperature.

The temperature follows the logarithmic cooling schedule suggested by Belisle (1992). Temperatures decrease according to the logarithmic cooling schedule $temp = temp_0/\log(((t-1)\backslash tmax) * tmax + \exp(1))$, where the backslash $\backslash$ signifies integer division, while $temp_0$ is the starting temperature, $tmax$ is the number of iterations at each temperature, and t is the iteration index for the current temperature. The defaults for the starting temperature $temp$ for the runs is 10, the number of iterations at each temperature $tmax$ is 10. In practice, these parameters often need to be tweaked to improve optimization.

Step 4 We repeat **Steps 2 and 3** until the total number of iterations is reached. If at any point we reach the end of the ordering of indices K_p, we then generate a new random order of indices, without replacement, and proceed as before. We set the total number of iterations as $\max\{10000, 10 * p\}$, to ensure that each covariate is in a proposal set at least 10 times.

This procedure is similar to the described greedy algorithm, except it permits non-greedy movements, or proposal sets that do not improve the score. This helps to avoid getting stuck in a local optimum, specifically in earlier iterations. At later iterations, the temperature is much lower and the procedure behaves closer to a greedy algorithm. This approach has a greater chance of finding a global optimum than greedy methods, and while not assured to find one, is much faster than exhaustive methods for high throughput data sets.

The reordering of indices every p iterations forces each covariate to be considered as part of a proposal set at least once every p iterations. If candidate covariates are chosen with replacement, there is a risk that some will not be included in the proposal sets at all. This is specifically true for high-dimension data sets of thousands of covariates, where many thousands of iteration steps are already required.

4.5 GENERAL NOTATION FOR HIERARCHICAL DATA

In this section we adapt the weakest-link model for analyzing data sets of hierarchical structure, under the motivating setting of multiple measurements within each of multiple cells per patient.

Consider a set of n patients with one outcome each, $Y_i : i = 1, \ldots, n$. There are m_i observed cells for each patient i, and $m = \sum_{i=1}^{n} m_i$ total cells. There are thus p covariate vectors $\mathbf{x}_k \in R^m; k = 1, \ldots, p$. The matrix $\mathbf{X}$ consists of p covariate vectors, consisting of observations $x_{j(i)k} : i = 1, \ldots, n; j = 1, \ldots, m_i; k = 1, \ldots, p$, with the measurements of each cell $j(i)$ nested within patient i.

We introduce additional notation for finding the weakest-link within an individual cell. Let $\hat{F}_k(x_{j(i)k})$ denote the empirical CDF of the k^{th} covariate of cell $j(i)$, relative to all cells, $j = 1, \ldots, m_i$, across all patients, $i = 1, \ldots, n$.

Denote $\underset{k\in\kappa}{\text{wl}}^{(D)} \left\{\hat{F}_k(x_{j(i)k})\right\}$=min $\left\{\hat{F}_{\kappa_1}^{(D(1))}(x_{j(i)\kappa_1}), \ldots, \hat{F}_{\kappa_{p^*}}^{(D(p^*))}(x_{j(i)\kappa_{p^*}})\right\}$ as the within-cell weakest-link from quantile stitching across covariates $k = \kappa_1, \ldots, \kappa_{p^*}$. The sequence D is the set of associated cdf directions for the weakest-link function across p^* parameters.

To aggregate information across all m_i cells of patient, let $\underset{j\in 1:m_i}{\text{mean}} \left[\underset{k\in\kappa}{\text{wl}}^{(D)} \left\{\hat{F}_k(x_{j(i)k})\right\}\right]$ denote the mean of the derived covariates for all cells $j = 1, \ldots, m_i$ cells obtained from the sample from a patient i. We can think of the Wilcoxon rank sum test statistic as being analogous to these within-patient sums of nonparametric empirical cdfs.

Table 2 provides a summary of these procedures and notation.

Table 2: Steps to obtain two different quantile stitched covariates

Level of weakest-link	Original covariate	Step 1	Step 2	Step 3 Final covariate
Cell $(j(i))$	$x_{j(i)k}$	$\hat{F}_k(x_{j(i)k})$ Take empirical cdf relative to all m cells	$\underset{k\in\kappa}{\text{wl}}^{(D)} \left\{\hat{F}_k(x_{j(i)k})\right\}$ Take weakest-link within cell $j(i)$ across covariates $k \in \kappa$	$\underset{j\in 1:m_i}{\text{mean}} \left[\underset{k\in\kappa}{\text{wl}}^{(D)} \left\{\hat{F}_k(x_{j(i)k})\right\}\right]$ Aggregate across all cells $j(i) = 1$ to m_i of patient i
Patient (i)	$x_{j(i)k}$	$\hat{F}_k(x_{j(i)k})$ Take empirical cdf relative to all m cells	$\underset{j\in 1:m_i}{\text{mean}} \left[\hat{F}_k(x_{j(i)k})\right]$ Aggregate across all cells $j(i) = 1$ to m_i of patient i	$\underset{k\in\kappa}{\text{wl}}^{(D)} \left\{\underset{j\in 1:m_i}{\text{mean}} \left[\hat{F}_k(x_{j(i)k})\right]\right\}$ Take weakest-link at level of patient i across covariates $k \in \kappa$

Consider an experiment with three covariates measured on each cell. We use the quantile stitching procedure as described above. The empirical cdf $\hat{F}_k^{(D)}(x_{j(i)k})$ is computed across all cells $j(i) = 1, \ldots, m_i; i = 1, \ldots, n$ for each of the 3 covariates $k = 1, 2, 3$. For a direction vector (D), the weakest-link for each cell $j(i)$ is $\underset{k\in 1:3}{\text{wl}}^{(D)} \left\{\hat{F}_k(x_{j(i)k})\right\}$. We then take the mean of the resulting covariates across all m_i cells of patient i. This operation results in the within-cell weakest-link derived covariate for patient i:

$$\underset{j\in 1:m_i}{\text{mean}} \left[\underset{k\in 1:3}{\text{wl}}^{(D)} \left\{\hat{F}_k(x_{j(i)k})\right\}\right] = \frac{1}{m_i} \sum_{j=1}^{m_i} \underset{k\in 1:3}{\text{wl}}^{(D)} \left\{\hat{F}_k(x_{j(i)k})\right\},$$

where $\underset{k\in 1:3}{\text{wl}}^{(D)} \left\{\hat{F}_k(x_{j(i)k})\right\} = \min \left\{\hat{F}_1^{(D(1))}(x_{j(i)1}), \hat{F}_2^{(D(2))}(x_{j(i)2}), \hat{F}_3^{(D(3))}(x_{j(i)3})\right\}$. For readability purposes, the notation can use the name of the k^{th} covariate instead of the index. For example, if $k = 1$ refers to protein p53, we reference the stitched covariates $\hat{F}_1(x_{j(i)1})$ simply as $\hat{F}(\text{p53})$.

Theorem Covariates using weakest-link taken at the patient level are greater or equal to those taken at the cell level for all patients i, or

$$\underset{k\in\kappa}{\text{wl}} \left\{\underset{j\in 1:m_i}{\text{mean}} \left[\hat{F}_k(x_{j(i)k})\right]\right\} \geq \underset{j\in 1:m_i}{\text{mean}} \left[\underset{k\in\kappa}{\text{wl}} \left\{\hat{F}_k(x_{j(i)k})\right\}\right].$$

Proof This property holds in general. Without loss of generality, assume we are taking the weakest-link as the minimum over all covariates. Define $K(i)$ as the weakest-link at the level of patient i, as described in section 4.1. Define for a cell j, the index of the derived covariate $K^*(j)$ to be:

$$K^*(j) = \arg\min \left\{\hat{F}_k(x_{j(i)k}); k = \kappa_1, \ldots, \kappa_{p^*}\right\} \text{ for } j = 1, \ldots, m_i.$$

The index $K^*(j)$ is the covariate index k at which the minimum of the empirical CDFs is attained within each cell j. By definition, $\hat{F}_k(x_{j(i)K(i)}) \geq \hat{F}_k(x_{j(i)K^*(j)})$ for all cells $j = 1, \ldots, m_i$; the quantities are equal when $K(i)$ corresponds to $K^*(j)$, the index of the minimum empirical CDF for cell j.

$$
\begin{aligned}
\underset{k\in\kappa}{\text{wl}}\left\{\underset{j\in 1:m_i}{\text{mean}}\left[\hat{F}_k(x_{j(i)k})\right]\right\} &= \min\left(\underset{j\in 1:m_i}{\text{mean}}\left[\hat{F}_k(x_{j(i)1})\right],\ldots,\underset{j\in 1:m_i}{\text{mean}}\left[\hat{F}_k(x_{j(i)p})\right]\right) \\
&= \underset{j\in 1:m_i}{\text{mean}}\left[\hat{F}_k(x_{j(i)K(i)})\right] \\
&= \frac{1}{m_i}\sum_{j=1}^{m_i}\left[\hat{F}_k(x_{j(i)K(i)})\right] \\
&= \frac{1}{m_i}\left\{\hat{F}_k(x_{1K(i)}) + \hat{F}_k(x_{2K(i)}) + \ldots + \hat{F}_k(x_{m_iK(i)})\right\} \\
&\geq \frac{1}{m_i}\left\{\hat{F}_k(x_{1K^*(j)}) + \hat{F}_k(x_{2K^*(j)}) + \ldots + \hat{F}_k(x_{m_iK^*(j)})\right\} \\
&= \frac{1}{m_i}\sum_{j=1}^{m_i}\hat{F}_k(x_{j(i)K^*(j)}) \\
&= \underset{j\in 1:m_i}{\text{mean}}\left[\underset{k\in\kappa}{\text{wl}}\left\{\hat{F}_k(x_{j(i)k})\right\}\right]
\end{aligned}
$$

This property is thus true for derived covariates from all subjects $i = 1,\ldots,n$.[**QED**]

4.6 LOGIC REGRESSION METHODS FOR DETECTING JOINT COMBINATIONS OF BINARY VARIABLES

Logic regression is another variable selection technique for constructing trees that consist of combinations of variables, where the variables are binary. The *R* implementation *logreg* of logic regression includes options for survival analysis, and require the predictor covariates to be dichotomous.

We use logic regression to obtain models, with dichotomous covariates, for the cytometry data set. We denote $I(\geq \text{median}(x))$ to be a dichotomizing function, equal to 1 if $x \geq$ the median of x, while $I(< \text{median}(x))$ is equal to 1 if $x <$ the median of x. More generally, $I(\geq \hat{F}_q^{-1}(x))$ is equal to 1 if x is larger than the $100{*}q^{th}$ percentile of x. We can thus compare the efficiency of models with binary covariates as predictors to models with continuous covariates. Within each panel of 4 covariates, we used logic regression to build the best model containing from $p = 1$ to 4 covariates.

The *logreg* function in the *LogicReg* of *R* assesses model fit through a scoring function such as sum of squares, for linear models, or partial likelihoods, for proportional hazards models. The cross-validation option of *logreg* assesses the number of covariates p^* that minimizes the average of this score, in the set of possible test sets. The function *logreg* then builds the best model consisting of boolean combinations *(and, or)* of exactly p^* binary covariates.

5.0 STUDY PLAN

For each of the objectives mentioned in Section 3, a short description is given, along with proposed data sets and methods for evaluating the practicality of weakest-link procedures compared to other algorithms.

Objective 1. To develop methods for screening of joint effects in high-throughput data using weakest-link approaches.

Description. The non-parametric (stitched) weakest-link function models the joint effects of variables during the variable selection process.

- We examine different transformation functions for stitching together multiple covariates. Previous weakest-link methods used the empirical cumulative distribution function (quantile stitching), while we propose the probit function of z-scores (probit stitching) in this dissertation to potentially reduce computation time.
- To handle cases where computational time is prohibitive due to the number of covariates in some types of data sets (more than 100 variables), we explore greedy procedures that obtain locally optimal results. We examine the effects of using a filtering step that removes covariates with low marginal effects on the outcome.
- We examine simulated annealing procedures as an alternative approach for variable screening high-throughput data sets.

Breast cancer microarray data set. To evaluate the methods' ability in identifying combinations of variables useful for predicting outcomes in a real data set of gene and protein data, that may be from the following sources:

- *Luminex*: Consists of at most 100 covariates, where it is possible to investigate all combinations of covariates.
- *Affymetrix*: Consist of considerably more covariates, sometimes over 1000. The variable selection procedures may require the use of greedy searching algorithms or filtering steps. However, these types of procedures may not adequately detect purely epistatic effects.

Evaluation. To evaluate the performance of these methods, we assess their ability in identifying combinations of variables predictive of binary lymph node status in a high-throughput Affymetrix breast cancer data set, with the following criterion:

- Computational time.
- Cross-validation for binary classification.

Objective 2. Compare efficiency between weakest-link methods and additive linear models using simulation.

Description. Use simulation to generate outcome vectors and feature matrices from known multivariable distribution under different scenarios, and compare how often methods identify true underlying model.

- Feature matrices consist of covariates related to the outcome, and random noise.
- We generate data from a known multivariable distribution, for both normally and non-normally distributed data:
- Use screening methods to compare how often methods identify covariate pairs truly associated with the outcome.

Simulated data sets. We investigate the efficiency of these methods by comparing their estimates with the underlying distributions used to generate the data.

- We compare the methods when the models are correct, and when they are misspecified, when:
  - An additive model, consisting of main effects and/or interaction effects, generates the data.
  - An epistatic model, consisting of weakest-link effects generate the data.

Evaluation. To evaluate the performance of these methods, we assess their ability in identifying predictive combinations of variables in the simulated data sets, with the following criterion:

- Computational time.
- Power, or ability to detect combinations of variables truly associated with outcomes in simulations, for data with known distributions.

Objective 3. To adapt and apply weakest-link methods for fitting hierarchical data.

Description. We combine weakest-link procedures and aggregation procedures to summarize information from many covariates and different cells within a patient. The aforementioned data simplification procedures can be performed in any order; changing the order of the operation results in different models and as a result, possibly different inferences and predictions.

- We present notation that concisely summarizes all steps in a logical sequence.
  - We present all weakest-link and aggregating functions, and the order they were used to obtain the derived covariate.
  - We choose notation to facilitate the communication and replication of these methods, either in text form or in code for implementation in software such as *R*.
- We use simple functions to aggregate data within all cells of a patient, by using:
  - Means.
  - Proportion above cutoff.

Lung cancer cytometry data set. We implement methods to model hierarchical data from a laser-scanning cytometry study of Pennsylvania lung cancer patients.

- Raw data: contains measurements of specific proteins and DNA content for each individual cell.
- Processed data: after gating procedures, consists of one value to estimate the corresponding protein or DNA content for each patient.

Evaluation. Compare the results of an analysis of a cytometry data of lung cancer patients by weakest-link method to currently used methods.

- Compare results from weakest-link methods to those from:
  - A conventional linear model.
  - Covariates that have been optimized visually, using a procedure such as gating.
  - Logic regression models with covariates dichotomized by:
    - From *a priori* cutoff points.
    - Covariates with data-driven cutoff points (such as the median).
    - Covariates optimized with respect to the outcome variable, using a procedure such as maximally selected statistics.
- Compare the aforementioned methods in their ability to model recurrence-free survival.

6.0 COVARIATE PAIR DETECTION FOR LARGE SCALE DATA SETS

In this chapter, we implement searches for interesting pairs of covariates in high-throughput data sets using both additive linear models consisting of main and/or interaction effects, and weakest-link models. To compare the effectiveness of the models in terms of their overall feasibility for examining these types of high-throughput data sets, we use the methods on an available data set of breast cancer patients, and also on high-throughput sets of simulated data.

6.1 SCREENING COVARIATE PAIRS FROM A DATA SET OF BREAST CANCER PATIENTS

As a test case, we use data from the Duke Breast Cancer SPORE study (West et al., 2001) tissue bank, consisting of samples of frozen tumors with clinical and pathologic information. Data were obtained from `http://data.genome.duke.edu/west.php`. This data set consisted of tumors from 49 patients. All tumors were diagnosed as invasive ductal carcinoma between 1.5 and 5 cm in maximal dimension, most being Stage II or worse. Data was obtained from Affymetrix Human Gene FL GENECHIP DNA arrays. Clinical annotations included estrogen receptor positivity and lymph node positivity, which are considered to be important as predictors of progression and survival.

6.1.1 Combinations of biomarkers associated with lymph node status

A diagnostic axillary lymph node dissection was performed on each of the 49 tumors involved in this set of patients. The investigators identified samples with at least one positive lymph

node as being lymph node positive (LN+) and lymph node negative (LN-) otherwise. Lymph node status is the classification of interest for the algorithms described in this section. In this study, lymph node positivity was assessed in all 49 patients; 24 were LN+ and 25 were LN-. The data set consisted of $7,129$ covariates of gene expression measurements; a description of each gene was provided. The online catalog OMIM (Online Mendelian Inheritance in Man) http://www.ncbi.nlm.nih.gov/omim served as a reference for the genes that are mentioned in this section. This webpage is updated daily and provides full-text descriptions and links to all known mendelian disorders and over 12,000 genes.

Previous studies (Shek and Godolphin, 1988) determined lymph node status to be one of the most useful single factors in predicting overall survival. According to West et al. (2001), these proteins may generate highly useful predictors of metastatic spread of the tumor. They may indicate patients in which positive lymph nodes are missed by the surgeon or pathologist. They may also be helpful in finding if the primary cancer is on the verge of metastasizing to an axillary lymph node at the time of diagnosis. West et al. (2001) further noted the potential practical benefits of lymph nodal classification by gene expression: it may preclude the need for an invasive and potentially harmful procedure such as dissection in diagnosing lymph node status.

We implement the greedy search algorithm **LocalOne** described in Section 4.4.2 to find a locally optimal pair of variables, and **LocalMany** described in Section 4.4.2.2 to find multiple locally optimal combinations in this breast cancer data. We compare the results of using the likelihoods from these models as the scoring function: a logistic regression model with either a quantile- or probit-stitching weakest-link covariate, and a logistic regression model with main effects only and one with both main and interaction effects.

To account for the number of possible pairs of covariates, p-values were Bonferroni-adjusted: p-values from the logistic regression model were multiplied by $\binom{7129}{2} = 25,407,756$ to account for the number of possible pairwise comparisons within an exhaustive search. Weakest-link models were also optimized over 4 different directionalities, with p-values also being multiplied by 4. These extremely conservative Bonferroni adjustments are acceptable as formal hypothesis testing is not the primary goal of these algorithms, rather, the objective

is to account for the huge amount of pairs being screened in as quick and simple a way as possible.

We used leave-one-out cross-validation (LOOCV) to assess the classification ability of the models. A single observation from the original sample is the test set, with the remaining observations constituting the training set. We first fit a logistic regression model in the training set, with the resulting coefficients $\hat{\beta}$ from the fitted model subsequently used to predict the binary lymph node status LN+ in the independent observation in the test set. The predicted outcome for the tumor in the test set is LN+ if $\hat{pr} > 0.5$, and LN-, otherwise. We thus obtain a prediction of lymph node status for all tumors by repeating this procedure with each observation used as a test set.

Before investigating interactions, we fit univariate logistic regression to assess the association between each of the biomarkers and lymph node positivity. The 20 genes with the highest marginal association with LN+ are listed in Table 3, along with their respective misclassification rates for predicting lymph node positivity by leave one out cross-validation.

6.1.1.1 Exhaustive filtered search for pairwise combinations Of the 7129 markers, 2936 had univariate associations with LN+ with unadjusted $p < 0.05$, and 1607 had associations with unadjusted $p < 0.01$. To make the screening more feasible in terms of computation time, we needed a more conservative criteria to filter the covariates. Of the 7,129 fitted univariable models, 376 had a $p < 0.001$ from the likelihood ratio test, while only 14 had a Bonferroni-adjusted $p < 0.05/7129$.

As a subset of 14 covariates is too limited, an exhaustive search was performed on all $\binom{376}{2} = 70,500$ possible pairwise combinations of biomarkers within this subset. We used a logistic regression model with a single probit-stitching weakest-link covariate, and also a logistic regression model with main and interaction effects. The respective p-values from these models were obtained from the likelihood ratio test, and Bonferroni adjusted. The weakest-link model identified 5 covariate pairs significantly associated with lymph node status after Bonferroni-adjustment, while the interaction model found 3 pairs (Table 4). Interestingly, these two sets of pairs did not overlap even though both models searched

Table 3: Genes with highest marginal association with LN+

Index	Unigene cluster	Deviance	p	mis(CV)
132	Bacteriophage P1 cre recombinase	38.1	4.7×10^{-8}	0.14
2921	Heparin binding binding protein (HBp17)	41.4	2.6×10^{-7}	0.22
2955	Uroporphyrinogen decarboxylase (URO-D)	41.8	3.1×10^{-7}	0.27
7072	mRNA for ORF (clone ICRFp507G2490)	43.4	7.5×10^{-7}	0.18
116	Telomerase reverse transcriptase (hTRT)	44.4	1.2×10^{-6}	0.18
5788	mRNA for spi-1 proto-oncogene	44.8	1.6×10^{-6}	0.29
4881	MutY gene (homolog of E.coli)	45.8	2.6×10^{-6}	0.24
5354	Paternally expressed gene 3	46.2	3.1×10^{-6}	0.29
773	mRNA for transactivator HSM-1	46.2	3.2×10^{-6}	0.35
5061	Cosmid clone LUCA14 from 3p21.3	46.7	4.2×10^{-6}	0.24
3059	Surfactant, pulmonary-associated protein SP-A	46.8	4.3×10^{-6}	0.29
6263	Mal gene exon 4	47.2	5.2×10^{-6}	0.24
4413	Bloom's syndrome protein	47.2	5.3×10^{-6}	0.22
4411	Butyrophilin precursor mRNA	47.3	5.5×10^{-6}	0.20
3430	Retinoblastoma protein 1 isoform I	47.7	7.1×10^{-6}	0.24
5793	Leukocyte tyrosine kinase mRNA	47.7	7.1×10^{-6}	0.27
1347	Gastric mucin	47.8	7.2×10^{-6}	0.24
6962	mRNA for chloride channel (putative) 2163bp	47.9	7.9×10^{-6}	0.22
2189	Ephrin receptor EphB2	48.3	9.7×10^{-6}	0.27
2037	Calcium channel, L type, alpha-1S	48.4	9.8×10^{-6}	0.33

20 genes with highest marginal associations with LN+. Deviance is the -log-likelihood with the binomial link function, p is the unadjusted p-value from univariate logistic regression.

through the same subset of genes. Furthermore, only one gene, Sp-A1 appeared in both lists even though, individually, 65 other genes had stronger associations.

Table 4: Covariate pairs associated with lymph node status, after exhaustive filtered search within the subset of genes with marginal $p < 0.001$

Pair	κ	Gene	marginal		weakest-link	
			p(unadj)	mis(cv)	p(Bonf)	mis(cv)
1	1264	Surfacant protein Sp-A1 delta	9.1×10^{-5}	0.18	0.015	0.14
	2955	Uroporphyrinogen decarboxylase (URO-D)	3.1×10^{-7}	0.27		
2	132	Bacteriophage P1 cre recombinase	4.7×10^{-8}	0.14	0.040	0.12
	4172	TPA-inducible gene-1 (TIG1) mRNA	4.1×10^{-4}	0.29		
3	118	Biotin synthetase	1.1×10^{-5}	0.24	0.033	0.20
	4413	Bloom's syndrome protein	5.3×10^{-6}	0.22		
4	132	Bacteriophage P1 cre recombinase	4.7×10^{-8}	0.14	0.001	0.12
	4413	Bloom's syndrome protein	5.3×10^{-6}	0.22		
5	1264	Surfacant Protein Sp-A1 Delta	9.1×10^{-5}	0.18	0.040	0.12
	4413	Bloom syndrome	5.3×10^{-6}	0.22		
Pair	κ	Gene	marginal		interaction	
			p(unadj)	mis(cv)	p(Bonf)	mis(cv)
6	1164	Tyrosine Phosphatase 1, Non-Receptor, Alt Splice 3	2.7×10^{-4}	0.37	0.001	0.10
	1264	Surfacant Protein Sp-A1 Delta	9.1×10^{-5}	0.18		
7	956	Modulator Recognition Factor 2	5.3×10^{-4}	0.29	0.027	0.14
	6962	mRNA for chloride channel (putative) 2163bp	7.9×10^{-6}	0.22		
8	4072	Lymphocyte dihydropyrimidine dehydrogenase	3.1×10^{-4}	0.27	0.008	0.18
	6962	mRNA for chloride channel (putative) 2163bp	7.9×10^{-6}	0.22		

Covariate pairs associated with lymph node status, after an exhaustive search within a subset of genes with marginal $p < 0.001$. Pairs 1 to 5 were detected by a logistic regression model with a probit-stitching weakest-link covariate, and pairs 6 to 8 by a logistic regression model with main and interaction effects. P(unadj) is the unadjusted p-value from a univariate logistic regression model with the gene as the only covariate with the misclassification rate from leave-one-out cross-validation. The Bonferroni-adjusted p-values for the pairwise models are listed with the misclassification rate, for the weakest-link model (WL) and linear model (LM), respectively.

6.1.1.2 Greedy search for locally optimal pairwise combinations

We found multiple local minima using the greedy algorithm described in Section 4.4.2.2. Logistic regression was used to model the data, with lymph node positivity as an outcome, and two different models for the joint effects of a covariate pair: a logistic regression model with a single probit-stitching weakest-link covariate, and a logistic regression model with main and interaction effects. The algorithm stops if $l_{cons} = 5$ consecutive local minima do not meet the Bonferroni-adjusted threshold $p < 0.05$. Additionally, due to computational time constraints the algorithm stops after finding $l_{max} = 25$ local minima.

Probit-stitching weakest-link The weakest-link model identified 21 pairs with Bonferroni-adjusted p-values < 0.05, out of $212,814$ pairs that were searched (Table 5). 13 of these combinations included two covariates each of which had Bonferroni-adjusted p of greater than 0.1; 12 of them involved the human protein hBrm protein, which is a homolog of the drosophila brm gene (Muchardt et al., 1996), with an unadjusted $p = 0.002$ ($p = 1.000$ after Bonferroni-adjustment). The greedy function found 10 local minima, 3 having Bonferroni-adjusted $p < 0.05$. The iterations stopped after the last 5 minima failed to satisfy $p < 0.05$.

As an example, covariate pair 18 consists of X_{5886} (Oncogene JUN-D) and X_{6199} hBrm, both of whose marginal association with lymph node positivity were not statistically significant after Bonferroni adjustment (unadjusted p of 0.005 and 0.002, respectively). These genes were not in the filtered data set in Table 3. Figure 5 is a dot plot of these two biomarkers by LN+. The dotted line on both of these plots indicates where $\hat{pr} = 0.5$, or the value of the covariate that distinguishes between LN+ and LN-.

On other hand, the probit stitching weakest-link covariate between these two biomarkers was very highly associated with LN+ ($p = 1.1 \times 10^{-9}$, unadjusted), ($p = 0.028$, after Bonferroni adjustment). A dot plot of the probit weakest-link covariate by LN+ is in Figure (left). A scatterplot between the two biomarkers is in Figure 6 (right). The COU is the diagonal line, while points with the same probit stitching weakest-link, i.e.

$$\text{wl}^{(+-)}\left\{\hat{F}_{norm}(X_{5886}), \hat{F}_{norm}(X_{6199})\right\} = \min\left\{\hat{F}_{norm}(X_{5886}), 1 - \hat{F}_{norm}(X_{6199})\right\}$$

lie on the same contour line. The dotted line from the weakest-link clearly distinguishes cases of LN+ and LN- better than those from the univariate analyses, from Figure 5.

Linear model with main and interaction effects The interaction model identified 26 pairs with Bonferroni-adjusted p-values < 0.05 (Table 6) out of $523,840$ pairs searched. 12 of these combinations involved two covariates where both had a Bonferroni-adjusted p of greater than 0.1; 5 of them involved the protein, Octamer binding transcription factor 1, with an unadjusted marginal $p = 2.2 \times 10^{-5}$ ($p = 0.156$ after Bonferroni-adjustment) and 4 of them involved the protein, Surfacant protein Sp-A1 delta, with an unadjusted marginal $p = 9.1 \times 10^{-5}$ ($p = 0.650$ after Bonferroni-adjustment). The greedy search from this model found many more local minima than that from the weakest-link model. Due to time and

Table 5: Covariate pairs associated with lymph node status, after greedy search for multiple local minima using probit stitching weakest-link model

Pair	κ	Gene	marginal		weakest-link	
			p(unadj)	mis(cv)	p(Bonf)	mis(cv)
1	132	Bacteriophage P1 cre recombinase	4.7×10^{-8}	0.14	0.036	0.20
	824	Immunoglobulin lambda gene locus DNA, clone:31F3	0.405	0.94		
2	132	Bacteriophage P1 cre recombinase	4.7×10^{-8}	0.14	0.040	0.12
	4172	TPA-inducible gene-1 (TIG1) mRNA	1.4×10^{-4}	0.29		
3*	132	Bacteriophage P1 cre recombinase	4.7×10^{-8}	0.14	0.001	0.12
	4413	Bloom syndrome	5.3×10^{-6}	0.22		
4	4413	Bloom syndrome	5.3×10^{-6}	0.22	0.004	0.12
	6199	hBrm	0.002	0.43		
5	11	mRNA for SH3 binding protein	0.004	0.41	0.046	0.14
	4413	Bloom syndrome	5.3×10^{-6}	0.22		
6	118	Biotin synthetase	1.1×10^{-5}	0.24	0.033	0.20
	4413	Bloom syndrome	5.3×10^{-6}	0.22		
7	1264	Surfacant Protein Sp-A1 Delta	9.1×10^{-5}	0.18	0.040	0.12
	4413	Bloom syndrome	5.3×10^{-6}	0.22		
8	6199	hBrm	0.002	0.43	0.048	0.16
	6385	mRNA for glutamate receptor subunit GluRC	0.002	0.39		
9	6199	hBrm	0.002	0.43	0.037	0.18
	6581	TGN46 (Trans-golgi network protein)	0.007	0.39		
10	6199	hBrm	0.002	0.43	0.029	0.16
	7091	Brain fetus mRNA (clone 1D8)	0.001	0.35		
11	6199	hBrm	0.002	0.43	0.038	0.16
	7112	mRNA axonemal dynein heavy chain (partial, ID hdhc4)	1.1×10^{-4}	0.29		
12	897	Acetyl-coenzyme A transporter	0.002	0.29	0.030	0.16
	6199	hBrm	0.002	0.43		
13	1190	K Channel, Voltage-Gated, Isk-Related Family, Member 1	1.4×10^{-4}	0.27	0.049	0.14
	6199	hBrm	0.002	0.43		
14	2146	Thyroid hormone receptor 13	8.7×10^{-5}	0.37	0.041	0.16
	6199	hBrm	0.002	0.43		
15	2338	Ribosomal protein S14	7.9×10^{-5}	0.31	0.005	0.12
	6199	hBrm	0.002	0.43		
16	3787	Folate receptor 3	1.6×10^{-5}	0.29	0.041	0.14
	6199	hBrm	0.002	0.43		
17	5870	Mahlavu hepatocellular carcinoma hhc	9.1×10^{-4}	0.31	0.021	0.16
	6199	hBrm	0.002	0.43		
18	5886	Oncogene JUN-D	0.005	0.33	0.028	0.12
	6199	hBrm	0.002	0.43		
19	6007	Beta-galactoside alpha-2,6-sialyltransferase	2.7×10^{-4}	0.24	0.017	0.14
	6199	hBrm	0.002	0.43		
20	1264	Surfacant protein Sp-A1 delta	9.1×10^{-5}	0.18	0.015	0.14
	2955	Uroporphyrinogen decarboxylase (URO-D)	3.1×10^{-7}	0.27		
21	1264	Surfacant protein Sp-A1 delta	9.1×10^{-5}	0.18	0.004	0.10
	3191	I-rel mRNA	0.004	0.41		

Covariate pairs associated with lymph node status using greedy search for local minima. P(unadj) is the unadjusted p-value from likelihood ratio test, and mis(cv) misclassification rate from leave one out cross validation for the univariate covariates with lymph node positivity; P(Bonf) is the Bonferroni-adjusted p-value with the misclassification rate from cross-validation, using the probit-stitching model. The pair numbers of locally optimal pairs are in **bold**, while the globally optimal pair is marked with an asterisk (*).

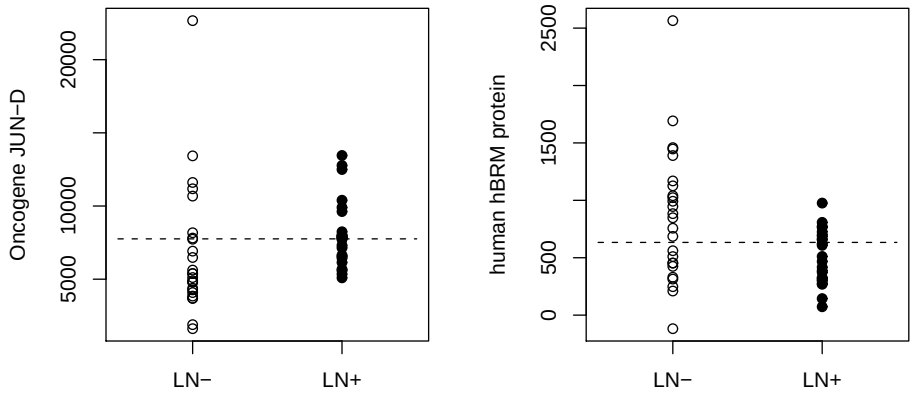

Figure 5: Dot plots of biomarkers Oncogene JUN-D and hBrm

Dot plots of marginal levels of biomarkers X_{5886} (Oncogene JUN-D, left) and X_{6199} (hBrm, right). Observations from lymph node positive samples (LN-) are in white, from lymph node positive (LN+) in black. The horizontal dotted lines lie along the biomarker value that provides a $\hat{pr} = 0.5$ from univariate logistic regression.

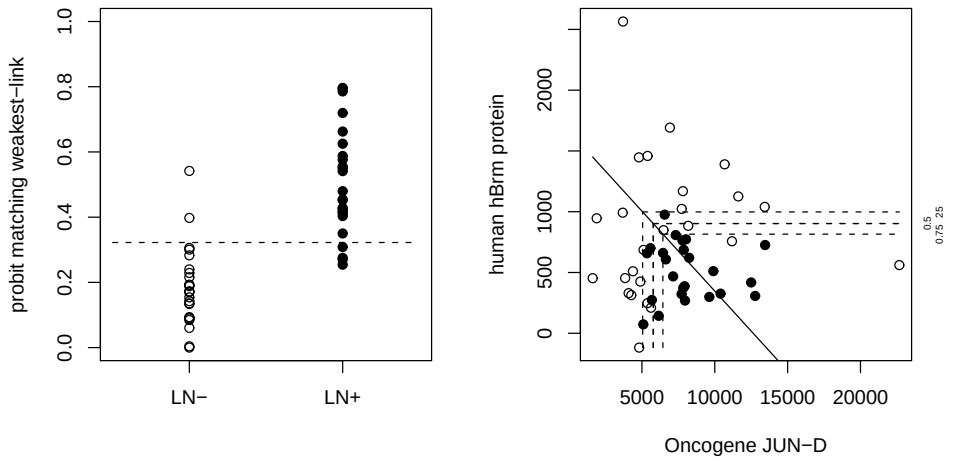

Figure 6: Probit stitching weakest-link between Oncogene Jun-D and hBrm protein

Dot plot of probit stitching weakest-link between biomarkers X_{5886} (Oncogene JUN-D) and X_{6199} (hBrm); the dotted line lies along the biomarker value that provides a $\hat{pr} = 0.5$ from univariate logistic regression (left). Scatterplot between biomarkers X_{5886}, (Oncogene JUN-D) and X_{6199} (hBrm); points with predicted $\hat{p}$ of LN+=0.25, 0.5 and 0.75 (right) are plotted on dotted contour lines. Observations from lymph node positive samples (LN-) are in white, from lymph node positive (LN+) in black.

computation constraints, with the search taking over 12 hours to run on a dual-core laptop, the algorithm was stopped after the user-input maximum of 25 local minima were found; 6 of these had a Bonferroni-adjusted $p < 0.05$.

Table 6: Covariate pairs associated with lymph node status, after greedy search for multiple local minima using main and interaction effects model

Pair	κ	Gene	marginal		interaction	
			p(unadj)	mis(cv)	p(Bonf)	mis(cv)
1	2955	Uroporphyrinogen decarboxylase (URO-D)	3.1×10^{-7}	0.27	0.022	0.14
	4202	Putative RNA binding protein RNPL	0.049	0.43		
2	773	mRNA for transactivator HSM-1	3.2×10^{-6}	0.35	0.009	0.12
	3271	Cell surface glycoprotein (IGB) mRNA	0.003	0.31		
3	3271	Cell surface glycoprotein (IGB) mRNA	0.003	0.31	0.028	0.14
	3327	Loricrin gene exons 1 and 2	0.004	0.31		
4	3079	Glycophorin Sta (type A) exons 3 and 4	0.215	0.69	0.003	0.14
	4413	Bloom syndrome	5.3×10^{-6}	0.22		
5	4413	Bloom syndrome	5.3×10^{-6}	0.22	0.012	0.14
	5731	Cytochrome P450HP	0.887	0.98		
6	4413	Bloom syndrome	5.3×10^{-6}	0.22	0.002	0.08
	6199	hBrm	0.002	0.43		
7	4413	Bloom syndrome	5.3×10^{-6}	0.22	0.026	0.18
	6432	Prostaglandin E receptor 3f	0.359	0.57		
8	4413	Bloom syndrome	5.3×10^{-6}	0.22	0.048	0.14
	6775	DAN26 protein	0.016	0.39		
9	4413	Bloom syndrome	5.3×10^{-6}	0.22	0.005	0.16
	6898	HMG2B	0.370	0.71		
10	4413	Bloom syndrome	5.3×10^{-6}	0.22	0.001	0.16
	7088	mRNA (clone 1A7)	0.623	0.86		
11	1919	Octamer binding transcription factor 1	2.2×10^{-5}	0.24	0.024	0.12
	7088	mRNA (clone 1A7)	0.623	0.86		
12	2235	Retinoblastoma susceptibility protein E413K	0.983	1.00	0.002	0.14
	4413	Bloom syndrome	5.3×10^{-6}	0.22		
13	2656	Melanoma-associated glycoprotein MUC18	0.408	0.53	0.024	0.12
	4413	Bloom syndrome	5.3×10^{-6}	0.22		
14	1919	Octamer binding transcription factor 1	2.2×10^{-5}	0.24	0.013	0.20
	2350	Human endogenous retrovirus HERV-K10	0.031	0.43		
15	1919	Octamer binding transcription factor 1	2.2×10^{-5}	0.31	0.024	0.16
	4205	High-mobility group phosphoprotein isoform I-C	0.580	0.59		
16	1919	Octamer binding transcription factor 1	2.2×10^{-5}	0.24	0.023	0.14
	4708	Cell cycle checkpoint control protein	0.878	0.86		
17	1153	Myelin basic protein	0.080	0.45	0.039	0.12
	1919	Octamer binding transcription factor 1	2.2×10^{-5}	0.24		
18	2607	Renin gene exon 9	0.032	0.37	0.044	0.14
	6424	Membrane-type matrix metalloproteinase	6.5×10^{-5}	0.27		
19*	1164	Tyrosine Phosphatase 1, Non-Receptor, Alt Splice 3	2.7×10^{-4}	0.37	0.001	0.10
	1264	Surfacant protein Sp-A1 delta	9.1×10^{-5}	0.18		
20	1264	Surfacant protein Sp-A1 delta	9.1×10^{-5}	0.18	0.009	0.14
	1713	Replication protein A3, 14-KD	0.068	0.47		
21	1264	Surfacant protein Sp-A1 delta	9.1×10^{-5}	0.18	0.038	0.12
	3191	I-rel (transcription factor RelB)	0.004	0.41		

... Table 6 continued

Pair	κ	Gene	marginal		interaction	
			p(unadj)	mis(cv)	p(Bonf)	mis(cv)
22	1264	Surfacant protein Sp-A1 delta	9.1×10^{-5}	0.18	0.040	0.16
	3434	Infertility-related sperm protein	0.001	0.35		
23	2065	Mitogen-activated protein kinase 14	0.795	0.78	0.016	0.16
	4296	K+ channel beta 2 subunit	0.016	0.24		
24	754	KIAA0198 gene	0.784	0.51	0.019	0.12
	7072	mRNA for ORF (clone ICRFp507G2490)	7.5×10^{-7}	0.18		
25	3969	Na,K-ATPase beta-1 subunit	0.714	0.80	0.025	0.12
	7072	mRNA for ORF (clone ICRFp507G2490)	7.5×10^{-7}	0.18		
26	4241	Human (memc) RNA	0.260	0.45	0.047	0.10
	7072	mRNA for ORF (clone ICRFp507G2490)	7.5×10^{-7}	0.18		

Covariate pairs associated with lymph node status using greedy search for local minima. P(unadj) is the unadjusted p-value from likelihood ratio test, and mis(cv) misclassification rate from leave one out cross validation for the univariate covariates with lymph node positivity; P(Bonf) is the Bonferroni-adjusted p-value with the misclassification rate from cross-validation, using the main and interaction effects model. The pair numbers of locally optimal pairs are in **bold**, while the globally optimal pair is marker with an asterisk (*).

There were only two pairs that had a Bonferroni-adjusted $p < 0.05$ that were detected by the greedy searches using both models: (1) Bloom syndrome and hBrm and (2) Surfacant protein Sp-A1 delta and I-rel mRNA. Figure 7 shows scatterplots of these pairs with corresponding contours from both types of models. Each pair included one biomarker not in the subset of genes in Table 3 and would have been missed by the filtered search: hBrm (p=0.002) and I-rel (p=0.004), respectively.

6.1.1.3 Linear combinations of pairs As in logic regression (Ruczinski, 2000), we can combine several derived covariates from weakest-link pairs into a same additive model. As each pair modeled from weakest-link is reduced to only one derived covariate, combining several weakest-link terms would only take up one degree of freedom each, compared to three degrees for each pair in additive interaction models.

Consider the following pairs detected by the greedy searches in this section (Table 7), which were found to have the lowest misclassification rates from each model fitting type (weakest-link and interaction).

We implemented additive linear models of the above pairs; (Table 8) shows the misclassification rate for each linear model of sets of pairs. While combining two derived weakest-link pairs slightly improved the misclassification rate relative to the best single derived weakest-

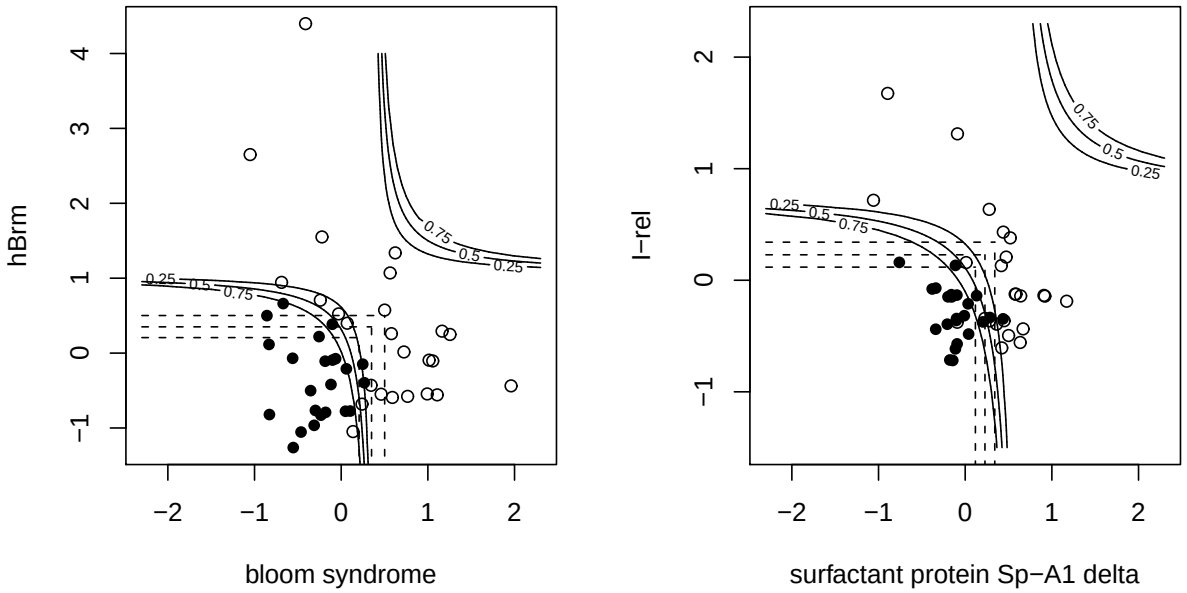

Figure 7: Scatterplots of pairs associated with LN+ detected by both models

Scatterplots of two sets of covariates, that were found to be significantly associated with lymph node status LN+ by both probit weakest-link and interaction models. Lymph node positive tumors (LN+) are in black, negative (LN-) in white. Equivalent values of the probabilities $\hat{pr} = 0.25, 0.5, 0.75$ of a positive outcome, predicted by the interaction model, are on the solid curved contour lines. Equivalent weakest-link values, being the true probability for generating a positive outcome, pr, are on the straight dotted contour lines. Values are normalized to the standard normal for the plots. Left: Bloom syndrome and hBrm and Right: Surfacant protein Sp-A1 delta and I-rel.

Table 7: Covariate pairs with lowest misclassification rates for predicting lymph node status by cross-validation

Pair	Covariate	Gene	marginal		combination	
			p(unadj)	mis(cv)	p(Bonf)	mis(cv)
		weakest-link				
i	X_{1264}	Surfacant protein Sp-A1 delta	9.1×10^{-5}	0.18	0.004	0.10
	X_{3191}	I-rel mRNA	0.004	0.41		
ii	X_{132}	Bacteriophage P1 cre recombinase	4.7×10^{-8}	0.14	0.001	0.12
	X_{4413}	Bloom syndrome	5.3×10^{-6}	0.22		
iii	X_{2338}	Ribosomal protein S14	7.9×10^{-5}	0.31	0.005	0.12
	X_{6199}	hBrm	0.002	0.43		
		main and interaction effects				
iv	X_{4413}	Bloom syndrome	5.3×10^{-6}	0.22	0.002	0.08
	X_{6199}	hBrm	0.002	0.43		
v	X_{1164}	Tyrosine Phosphatase 1, Non-Receptor, Alt Splice 3	2.7×10^{-4}	0.37	0.001	0.10
	X_{1264}	Surfacant protein Sp-A1 delta	9.1×10^{-5}	0.18		
vi	X_{4241}	Human (memc) RNA	0.260	0.45	0.047	0.10
	X_{7072}	mRNA for ORF (clone ICRFp507G2490)	7.5×10^{-7}	0.18		

Covariate pairs with lowest misclassification rates for predicting lymph node positivity by misclassification. P(unadj) is the unadjusted p-value from likelihood ratio test, mis(cv) misclassification rate from leave one out cross validation for the univariate covariates with lymph node positivity; P(Bonf) is the Bonferroni-adjusted p-value with the misclassification rate from cross-validation, using the probit-stitching model.

link covariate (from 0.10 to 0.08), the rates worsened when accounting for two or more multiple interaction pairs (from 0.08 to 0.14). This is most likely due to the 6 degrees of freedom needed to account for 2 sets of pairs by the interaction model relative to the 2 degrees of freedom needed by the weakest-link model.

Table 8: Lymph node status constructed by linear combinations of covariate pairs

Model	df	mis(cv)
weakest-link		
$LN^+ \sim \text{wl}\{X_{1264}, X_{3191}\}$	1	0.10
$LN^+ \sim \text{wl}\{X_{1264}, X_{3191}\} + \text{wl}\{X_{132}, X_{4413}\}$	2	0.08
$LN^+ \sim \text{wl}\{X_{1264}, X_{3191}\} + \text{wl}\{X_{132}, X_{4413}\} + \text{wl}\{X_{2338}, X_{6199}\}$	3	0.14
main + interaction effects		
$LN^+ \sim X_{4413} * X_{6199}$	3	0.08
$LN^+ \sim X_{4413} * X_{6199} + X_{1164} * X_{1264}$	6	0.12
$LN^+ \sim X_{4413} * X_{6199} + X_{1164} * X_{1264} + X_{4241} * X_{7072}$	9	0.14

Misclassification rate by cross-validation for predicting lymph-node positivity (LN+), for additive linear models consisting of pairs found by greedy methods. df is the degrees of freedom of the model.

This type of model is an example where the lower degrees of freedom required by the weakest-link models poses a clear advantage.

6.1.2 Discussion of analysis of breast cancer data set

Both models included in the greedy searches found numerous pairs associated with lymph node status, even after an extremely large Bonferroni adjustment for over 25 million possible combinations. The large number of potentially interesting combinations found in this analysis points to the weakest-link model as being a practical tool for screening through high-throughput data sets.

The greedy search using the weakest-link model took less computation time, requiring approximately 40% of the time to run as that of the interaction model. For this high-throughput data set, the weakest-link model was found to be more efficient, considering it found almost as many significant pairs, and these significant pairs had approximately the same predictive value (around 0.10-0.14 misclassification rate from LOOCV) as the interaction model.

The exhaustive search within the filtered set of genes with marginal $p < 0.001$ fewer results; only 5 pairs significantly associated with LN+ according to the weakest-link model, and 3 pairs from the interaction model. As such it is interesting to note that most of the pairs found by either greedy method would have been missed in an exhaustive filtered search. A search in a subset with a less stringent filtering criteria (say, $p < 0.01$) would have detected most of the pairs found by the greedy methods, but it would have required a large amount of computation time for fitting over 1 million models.

More pairs also would have likely been found by using a less conservative multiple comparisons procedure such as false discovery rate or Q-value. However, given the large number of pairs associated with LN+ after Bonferroni adjustment of p-values, this was not an issue in this data set.

As expected, we found that combining two or more weakest-link covariates, representing two or more pairs, in a single linear model was more feasible than combining two or more pairs of main and interaction effects terms. The misclassification rate slightly decreased when combining the weakest-link information from two pairs, but increased when combining pairs in an interaction model. This is most likely due to each pair needing only 1 coefficient in a weakest-link model, but 3 in a main and interaction effects model.

6.2 COVARIATE PAIR DETECTION COMPARISONS THROUGH SIMULATION

In this section, we implement the previously described methods on simulated data sets generated from underlying distributions with known parameters. In microarray data sets, researchers are often interesting in finding a set of genes indicative of a phenotype and in turn, disease risk. We assess the relative efficiency of each of the weakest link and additive models in these high-throughput data sets typical in these types of studies, by comparing how often they are able to screen through the massive amounts of data and successfully detect genes that are truly of interest.

6.2.1 Simulation procedures

The simulations proceed as follows. We generate a matrix $\mathbf{X} = x_{ij}; i = 1, \ldots, n; j = 1, \ldots, p$, consisting of covariates with index $j = 1, \ldots, p$ with elements $i = 1, \ldots, n$. In the simulation study, we first generate three pairs of bivariate normal distributed covariates, with randomly chosen indices $k \in 1, \ldots, p$; these index pairs are $k = (\kappa_1, \kappa_2)$, $k = (\kappa_3, \kappa_4)$ and $k = (\kappa_5, \kappa_6)$. Using the *rmvnorm* function from the *mvtnorm* package in *R*, we generate these 3 pairs of covariates in an identical fashion differing only in their respective correlations. Covariate pairs with indices of (κ_1, κ_2), (κ_3, κ_4) and (κ_5, κ_6) have respective pairwise correlations of -0.5, 0 and 0.5 between them. The rest of the $p - 6$ covariates have independent and identical standard normal distributions; these covariates serve as additional noise that the model fitting methods must work around to find the true outcome-generating pair.

We use these simulated feature data sets to compare the relative efficiency and precision of the weakest-link model and additive linear models in modeling joint effects, for feature sets consisting of $p = 10$ and $p = 100$ covariates. For conciseness, we refer to an additive linear model with main effects only as a main effects model, and a linear model with main and interaction effects as an interaction model. Each model generates a separate set of outcomes Y_{wl}, Y_{main} and Y_{int} from the same set of covariates. The data matrix $\mathbf{X}$ generates 3 different sets of n outcomes, directly from each of the 3 covariate pairs:

- *generated from weakest-link model:* $Y_{wl}^{(1)}$, $Y_{wl}^{(2)}$ and $Y_{wl}^{(3)}$
- *generated from additive model with main effects only:* $Y_{main}^{(1)}$, $Y_{main}^{(2)}$ and $Y_{main}^{(3)}$
- *generated from additive model with main and interaction effects:* $Y_{int}^{(1)}$, $Y_{int}^{(2)}$ and $Y_{int}^{(3)}$.

For units of observation $i = 1, \ldots, n$ and covariate pairs in model $j = (x_{i\kappa_1}, x_{i\kappa_2})$, we generate binary outcomes $y_i^{(1)} \sim \text{Bernoulli}(pr_i)$. Each outcome $y_i^{(1)}$ is a Bernoulli trial of probability $pr_i = logit^{-1}(log(d) * \min\{x_{i\kappa_1}, x_{i\kappa_2}\})$. pr_i is the anti-logit of the derived weakest-link covariate $log(d) * \min\{x_{i\kappa_1}, x_{i\kappa_2}\}$, and we multiply by $\log(d)$ to reflect an odds ratio of d for each unit change in the weakest-link covariate.

Similarly, to generate outcomes from the additive models perform similar procedures on the sums of the terms in these models. For the main effects model, set $pr_i = logit^{-1}(log(d) * (x_{i\kappa_1} + x_{i\kappa_2})/2)$ and the interaction model, set $pr_i = logit^{-1}(log(d) * (x_{i\kappa_1} + x_{i\kappa_2} + x_{i\kappa_1} * x_{i\kappa_2})/2)$.

This corresponds to using a conventional linear model with $E(Y|X) = \alpha + \sum_{k=1}^{p} \beta_k X_k$, with $\alpha = 0$ and $\beta_k = 1$ for all k. For any exponential family distribution we can use generalized linear models with the appropriate link function to generate the outcome.

The 2 in the denominator of the $Y_{main}^{(1)}$ and $Y_{int}^{(1)}$ expressions accounts for the greater effect sizes produced by the interaction model, relative to the weakest-link model, from summing two random variables. This produces more comparable effect sizes and results between the two generating models.

We then similarly sets 3 sets of weakest-link, main effects, and interaction effects outcome for each of the other two covariate pairs $j = (x_{i\kappa_3}, x_{i\kappa_4})$ and $j = (x_{i\kappa_5}, x_{i\kappa_6})$.

We then compare the weakest-link and additive models as follows. We fit $\binom{p}{2}$ regression models, corresponding to each possible pairwise combination of covariates. The program fits the data according to each of the following models:

- a quantile stitched weakest-link model
- a probit stitched weakest-link model
- an additive linear model with only main effects
- an additive linear model with both main and interaction effects.

A simple way to identify specific pairs as being of interest with respect to outcome is to apply a simple hypothesis test. A simple test is a likelihood ratio test and the corresponding p-value, after Bonferroni adjustment for number of the pairs. This applies to a hypothesis test of at least one pair associated with the outcome.

$$\begin{aligned} H_0 &: \text{no pairs associated with outcome} \\ H_1 &: \text{at least one pair associated with outcome} \end{aligned}$$

We can also use these p-values to detect more than one covariate pair associated with the outcome. With the extremely high number of possible combinations in a high-throughput data set, it may be difficult to detect the globally optimal pair, especially when certain covariates are correlated with each other. We thus are also interested to see how successful these methods detect these covariate pairs with these significance tests. Another question is seeing how often covariate pairs are falsely identified as being associated with the outcome.

Thus, to summarize the coverage probabilities of the methods in each of the 100 replicates in a concise manner, we record:

- if the methods correctly detect the generating covariate pair as the global optimum, or the pair with the highest association with the outcome according to a scoring function such as a log-likelihood statistic, out of $\binom{p}{2}$ possible combinations.
- if the correct covariate pair passes a hypothesis test. For this study, we test the hypothesis by using a likelihood ratio test with p-value < 0.05, after Bonferroni adjustment for $\binom{p}{2}$ comparisons.
- to assess the precision of the hypothesis test, we also record the number of covariate pairs deemed significant according to the same likelihood ratio test. The identified covariate pairs other than the true outcome-generating pair are thus false positives.

The simulations also recorded other quantities, such as the frequency that of detecting at least one of the covariates from the true outcome-generating pair, or the frequency of detecting a pair containing at least one of the covariates in the list of pairs passing the hypothesis test. However, they are not reported here due to space constraints, and they did not provide any additional insights to the quantities included in this dissertation.

We assess the coverage probabilities of the methods in 100 replicates, each with a new randomly generated data matrix $\mathbf{X}$. We quantify the ability of the methods to detect a globally optimal covariate pair by the estimated proportion of replicates that correctly detect the outcome-generating pair. This enables us to use a paired design, with each of the models testing the same feature set and outcomes within the same replicate.

We use the appropriate McNemar tests to tests paired differences in proportions. In practice, we see that 100 replicates is enough to reject the null hypothesis with an observed difference of at least 0.10 in most of the paired tests for difference in proportions. Given two independent covariates X_1 and X_2 of size 100, with respective $p_1 = 0.65$ and $p_2 = 0.80$ to generate a 'yes', a two-sided McNemar's test with $\alpha = 0.05$ detects this difference 60% of the time. However given two covariates of size 100, with respective $p_1 = 0.70$ and $p_2 = 0.80$ to generate a 'yes', and X_2 constrained to only be 'Yes' if X_1 is no, a two-sided McNemar's test with $\alpha = 0.05$ detects this difference 79% of the time.

The following four methods screen through the data set for covariate pairs, when appropriate:

- an exhaustive search through all possible pairs of all covariates
- a filtered search through all pairs within a subset of covariates
- a greedy search for a locally optimal pair of covariates
- a simulated annealing algorithm for a globally optimal pair of covariates.

The exhaustive search fits a model associating the outcomes with all possible covariate pairs. This approach is feasible for lower dimension data sets, with few features. For higher-dimension data sets, we need to investigate other approaches which require fewer model fits and computations.

A simple approach to reduce computation time consists of searching only within a subset of covariates more likely to be part of the true outcome-generating combination. Filtering out covariates with poor marginal associations is a simple way to accomplish this. We first perform a univariable test for the marginal associations between each of the covariates and the outcome. In this study, we form a reduced data set consisting only of covariates that have Bonferroni-adjusted p-values of < 0.1 from univariable likelihood ratio tests from a filtered set of covariates. This not only reduces computation time by reducing the number of combinations screened through, it also lessens the Bonferroni adjustment. However, this approach may fail to detect purely epistatic effects. As shown in Section 4.4.1, there is no bound to the association of a weakest-link covariate with the outcome, with respect to the marginal association between each of the component covariates of the pair and the outcome.

For higher-dimensional data sets where exhaustive searches are not feasible due to combinatorial explosion, we study two other algorithms: a greedy search for local minima (Section 4.4.2) and a simulated annealing algorithm (Section 4.4.3).

6.2.2 Simulation results

The following tables summarize the detection of covariate pairs in several situations.

Simulations for variable screening for high-throughput data sets, of 100 covariates or more, were performed at the Pittsburgh Supercomputing Center. These simulations were

performed on the POPLE cluster, which is a SGI Altix 4700 distributed shared machine consisting of 768 processing cores. Computations for these simulations were parallelized to run on 16 cores. An R version of MPI (Message Passing Interface) was used to run each replication separately on each core, with the package *snowfall* used as a wraparound for R commands in MPI.

For a data set of $p = 10$ covariates, of size $n = 50$ and 100 respectively, Tables 9 and 11 display the proportion of times, in 100 replications, that each of the models and methods successfully identified the outcome-generating pair as the pair that best fit the data. In each replication, there were 3 independent sets of binary outcomes, of size n, generated by 3 separate covariate pairs within a randomly generated n by 10 data matrix, as described in Section 6.2.1. For each covariate pair, with pairwise correlation of -0.5, 0 and 0.5, respectively, the model generates a separate set of outcomes according to probabilities by the anti-logit of the outcome-generating model. We multiply by $log(10)$ to obtain an odds ratio of 10 for each incremental increase in weakest-link covariate, or sum of terms in the main or main+interaction effects models.

The exhaustive search was successful if, for a given replication, the true outcome generating pair minimized the negative log-likelihood, out of all $\binom{10}{2} = 45$ possible pairs. The filtered search was successful if both components of the outcome-generating pair had a univariate association with the outcomes of $p < 0.05$ (to pass the filtering stage), and also minimized the negative log-likelihood out of $\binom{p^*}{2}$ possible pairs, where p^* is the number of covariates marginally significantly associated with the outcome. The McNemar test for paired binary data was used to compare the success of the test models in finding the true outcome-generating covariate pair as the globally optimal solution.

Table 10 displays results in terms of significance testing, for sample size of 50. The numerator is the proportion of replications where the true outcome-generating covariate pairs had Bonferroni adjusted $p < 0.05$, after taking into account 45 possible comparisons for the exhaustive method, and $\binom{p^*}{2}$ possible pairs for the filtering method. The denominator is the mean number of pairs, per replication, with Bonferroni adjusted $p < 0.05$, after taking into account 45 possible comparisons for the exhaustive method, and $\binom{p^*}{2}$ possible pairs for the filtering method. The denominator thus illustrates the tendency for false positives and

possible overfitting from each of the fitting models. We also used the McNemar test to compare the models' relative ability to detect the true outcome-generating pair as 'statistically significant'.

Table 12 displays the above finding for a single set of binary outcomes generated simultaneously by three different covariate pairs. We obtain the sums of the 3 expressions corresponding to each covariate pair, and in turn determine the probabilities for the binary outcomes from the anti-logit of this sum. For example, for the weakest-link model the probabilities are obtained from:

$$pr_i = logit^{-1}\left\{\min(x_{i\kappa_1}, x_{i\kappa_2}) + \min(x_{i\kappa_3}, x_{i\kappa_4}) + \min(x_{i\kappa_5}, x_{i\kappa_6})\right\}.$$

Tables 13 and 14 repeat the above procedures, in data sets of $p = 100$ covariates of size 100. We also performed a filtered search on a subset of covariates with univariate association with outcome of $p < 0.001$, or $p < 0.1/7129$ corresponding to Bonferroni adjustment. Due to the additional computational time needed for screening through more combinations, we also implemented a greedy and simulated annealing search. For the latter two methods, p-values were adjusted for $\binom{100}{2} = 4950$ comparisons. Finally, Table 15 displays the effectiveness of the methods for right-skewed covariates. To generate the right-skewed covariates, we took the exponential of each of the randomly generated normal covariates.

Example As an example, take Table 13, from the simulation of 100 covariates of length 100. For positively correlated covariates $\rho = 0.5$, both types of stitched weakest-link models, quantile- and probit-, detected the true pair generating the outcomes from the weakest-link model in 95 of the 100 replicates, compared to 58 by the main effects model and 51 from the interaction model. The estimated difference in detection rates between the weakest link and the main effects models was 0.37 ± 0.05, and the difference between the weakest-link and interaction models was 0.44 ± 0.05; the differences in both were highly significant and the displayed proportions of the weakest-link models have a *cd* in the superscript. However, the estimated difference between the rates from the main effects and interaction models was 0.07 ± 0.05, so there are no superscripts with either proportion. The tables summarize information in this way due to the sheer amount of comparisons in these simulations.

Also, of note, are the result of outcomes generated by a main and interaction effects model. Both weakest-link models detected the outcome-generating pair in 79 of the 100 replicates; the main effects detected the pair in 34, and the interaction model in 81 replicates. The difference in detection rates between the interaction and main effects models was 0.47 ± 0.05, and between both weakest-link models and main effects models was 0.45 ± 0.06. However, there was virtually no difference between the detection rate of the interaction and weakest-link models (0.02 ± 0.05), even though the interaction model generated the data.

Section 6.2.5 provides a more general summary of the observations from these comparisons.

Table 9: Observed proportions of replicates detecting correct pair out of 10 simulated normally-distributed covariates of size 50, 3 binary outcomes

$\mathbf{X}$: 50×10 matrix, generated in 100 replicates as follows:

$$(X_{\kappa_1}, X_{\kappa_2}) \sim N\big(0, \left[\begin{smallmatrix} 1 & -0.5 \\ -0.5 & 1 \end{smallmatrix}\right]\big); (X_{\kappa_3}, X_{\kappa_4}) \sim N\big(0, \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]\big)$$
$$(X_{\kappa_5}, X_{\kappa_6}) \sim N\big(0, \left[\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right]\big); X_k \sim N(0,1); k \notin \kappa \text{ i.i.d.}$$

There are 3 outcomes from the pair of covariates $(X_{\kappa_1}, X_{\kappa_2})$:

$$Y_{wl}^{(1)} \sim Bernoulli(logit^{-1}(log(10) * \min(X_{\kappa_1}, X_{\kappa_2}))); Y_{main}^{(1)} \sim Bernoulli(logit^{-1}(log(10) * (X_{\kappa_1} + X_{\kappa_2})/2))$$
$$Y_{int}^{(1)} \sim Bernoulli(logit^{-1}(log(10) * (X_{\kappa_1} + X_{\kappa_2} + X_{\kappa_1} X_{\kappa_2})/2))$$

and similarly define Y_{wl}, Y_{main} and Y_{int} for the covariate pairs indexed by (κ_3, κ_4) and (κ_5, κ_6).

Data fitting model:	QWL	PWL	main	int
Exhaustive search through $\binom{10}{2}$ pairs				
Outcome generating model		Negative pairwise correlation $\rho = -0.5$		
Weakest-link	0.72^{c}	0.75^{c}	0.55	0.72^{c}
Main effects	0.45	0.47	0.78^{abd}	0.65^{ab}
All effects	0.74	0.80^{ac}	0.67	0.83^{c}
		No pairwise correlation $\rho = 0$		
Weakest-link	0.81^{d}	0.85^{cd}	0.71	0.68
Main effects	0.65	0.68	0.84^{abd}	0.69
All effects	0.76	0.79^{c}	0.69	0.86^{ac}
		Positive pairwise correlation $\rho = 0.5$		
Weakest-link	0.88^{cd}	0.87^{cd}	0.56^{d}	0.44
Main effects	0.77^{cd}	0.78^{cd}	0.61^{d}	0.44
All effects	0.57^{c}	0.59^{c}	0.27	0.64^{c}
Filtered search within subset of covariates with unadjusted $p < 0.05$				
Outcome generating model		Negative pairwise correlation $\rho = -0.5$		
Weakest-link	0.71^{c}	0.80^{ac}	0.52	0.76^{c}
Main effects	0.42^{b}	0.34	0.81^{abd}	0.65^{ab}
All effects	0.68	0.71	0.67	0.83^{abc}
		No pairwise correlation $\rho = 0$		
Weakest-link	0.82^{c}	0.84^{cd}	0.69	0.72
Main effects	0.59	0.58	0.81^{abd}	0.62
All effects	0.78^{c}	0.80^{c}	0.61	0.85^{c}
		Positive pairwise correlation $\rho = 0.5$		
Weakest-link	0.87^{cd}	0.85^{cd}	0.53	0.46
Main effects	0.80^{cd}	0.80^{cd}	0.60^{d}	0.43
All effects	0.67^{c}	0.68^{c}	0.31	0.65^{c}

Numbers: proportion of times that globally optimal pair detected correct outcome generating pair, within 100 replicates.
QWL: quantile-stitching weakest-link; PWL: probit-stitching weakest-link
main: linear model with main effects only; interaction: linear model with interaction effects
significant difference $p < 0.05$ according to McNemar's test
^{a}: significantly higher than QWL; ^{b}: significantly higher than PWL
^{c}: significantly higher than main; ^{d}: significantly higher than interaction

Table 10: Observed mean number of significant pairs detected per replication, from 10 simulated normally-distributed covariates of size 50, 3 binary outcomes

$\mathbf{X}$: 50×10 matrix, generated in 100 replicates as follows:

$$(X_{\kappa_1}, X_{\kappa_2}) \sim N\big(0, \left[\begin{smallmatrix} 1 & -0.5 \\ -0.5 & 1 \end{smallmatrix}\right]\big); (X_{\kappa_3}, X_{\kappa_4}) \sim N\big(0, \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]\big)$$
$$(X_{\kappa_5}, X_{\kappa_6}) \sim N\big(0, \left[\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right]\big); X_k \sim N(0,1); k \notin \kappa \text{ i.i.d.}$$

There are 3 outcomes from the pair of covariates $(X_{\kappa_1}, X_{\kappa_2})$:

$$Y_{wl}^{(1)} \sim Bernoulli(logit^{-1}(log(10) * \min(X_{\kappa_1}, X_{\kappa_2}))); Y_{main}^{(1)} \sim Bernoulli(logit^{-1}(log(10) * (X_{\kappa_1}) + X_{\kappa_2})/2)$$
$$Y_{int}^{(1)} \sim Bernoulli(logit^{-1}(log(10) * (X_{\kappa_1} + X_{\kappa_2} + X_{\kappa_1} X_{\kappa_2})/2))$$

and similarly define Y_{wl}, Y_{main} and Y_{int} for the covariate pairs indexed by (κ_3, κ_4) and (κ_5, κ_6).

Data fitting model:	QWL	PWL	main	int
Exhaustive search through $\binom{10}{2}$ pairs				
Outcome generating model		Negative pairwise correlation $\rho = -0.5$		
Weakest-link	0.51^{c}/0.70	0.55^{c}/0.77	0.25/0.45	0.51^{c}/0.65
Main effects	0.21/0.58	0.19/0.52	0.44^{ab}/1.08	0.41^{ab}/0.95
All effects	0.55^{c}/0.89	0.60^{c}/1.02	0.33/0.91	0.60^{c}/1.20
		No pairwise correlation $\rho = 0$		
Weakest-link	0.77^{cd}/1.73	0.81^{cd}/1.93	0.56/2.75	0.65/2.37
Main effects	0.69/1.97	0.72/2.26	0.78^{a}/4.71	0.77^{a}/3.89
All effects	0.80^{c}/2.00	0.82^{c}/2.16	0.68/4.18	0.87^{c}/3.49
		Positive pairwise correlation $\rho = 0.5$		
Weakest-link	0.95/4.74	0.96/5.32	0.95/11.66	0.92/10.48
Main effects	0.92/5.69	0.91/1.84	0.93/13.01	0.91/11.50
All effects	0.68/2.52	0.68/2.84	0.59/5.80	0.82^{abc}/4.86
Filtered search within subset of covariates with unadjusted $p < 0.05$				
Outcome generating model		Negative pairwise correlation $\rho = -0.5$		
Weakest-link	0.58^{cd}/0.78	0.62^{cd}/0.86	0.24/0.38	0.47^{c}/0.63
Main effects	0.18/0.42	0.21/0.49	0.52^{abd}/1.06	0.41^{ab}/0.82
All effects	0.56^{cd}/0.76	0.61^{cd}/0.81	0.31/0.57	0.70^{c}/1.00
		No pairwise correlation $\rho = 0$		
Weakest-link	0.77^{cd}/1.74	0.77^{cd}/1.64	0.62/3.34	0.63/2.82
Main effects	0.64/1.63	0.62/1.68	0.79^{abd}/4.27	0.66/3.30
All effects	0.76^{c}/2.01	0.76^{c}/2.13	0.62/3.57	0.83^{c}/3.29
		Positive pairwise correlation $\rho = 0.5$		
Weakest-link	0.90/4.66	0.95^{cd}/5.07	0.88/10.65	0.84/9.61
Main effects	0.95/5.62	0.97/6.21	0.97/12.94	0.94/11.83
All effects	0.68/2.87	0.75^{ac}/3.35	0.61/6.74	0.81^{ac}/6.06

Numerators: proportion of replicates where true outcome-generating pair had Bonferroni-adjusted p-value < 0.05.
Denominators: mean number of pairs per replicate with Bonferroni-adjusted p-value < 0.05, out of 100 replicates.
QWL: quantile-stitching weakest-link; PWL: probit-stitching weakest-link
main: linear model with main effects only; interaction: linear model with interaction effects
significant difference $p < 0.05$ according to McNemar's test
^{a}: significantly higher than QWL; ^{b}: significantly higher than PWL
^{c}: significantly higher than main; ^{d}: significantly higher than interaction

Table 11: Observed proportions of replicates detecting correct pair out of 10 simulated normally-distributed covariates of size 100, 3 binary outcomes

$\mathbf{X}$: 100×10 matrix, generated in 100 replicates as follows:

$$(X_{\kappa_1}, X_{\kappa_2}) \sim N\big(0, \left[\begin{smallmatrix} 1 & -0.5 \\ -0.5 & 1 \end{smallmatrix}\right]\big); (X_{\kappa_3}, X_{\kappa_4}) \sim N\big(0, \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]\big)$$
$$(X_{\kappa_5}, X_{\kappa_6}) \sim N\big(0, \left[\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right]\big); X_k \sim N(0,1); k \notin \kappa \text{ i.i.d.}$$

There are 3 outcomes from the pair of covariates $(X_{\kappa_1}, X_{\kappa_2})$:

$$Y_{wl}^{(1)} \sim Bernoulli(logit^{-1}(log(10) * \min(X_{\kappa_1}, X_{\kappa_2}))); Y_{main}^{(1)} \sim Bernoulli(logit^{-1}(log(10) * (X_{\kappa_1} + X_{\kappa_2})/2))$$
$$Y_{int}^{(1)} \sim Bernoulli(logit^{-1}(log(10) * (X_{\kappa_1} + X_{\kappa_2} + X_{\kappa_1} X_{\kappa_2})/2))$$

and similarly define Y_{wl}, Y_{main} and Y_{int} for the covariate pairs indexed by (κ_3, κ_4) and (κ_5, κ_6).

Data fitting model:	QWL	PWL	main	int
Exhaustive search through $\binom{10}{2}$ pairs				
Outcome generating model		Negative pairwise correlation $\rho = -0.5$		
Weakest-link	0.99^{c}	1.00^{c}	0.87	0.97^{c}
Main effects	0.69	0.74	0.98^{ab}	0.96^{ab}
All effects	0.96	0.98	0.95	0.99
		No pairwise correlation $\rho = 0$		
Weakest-link	1.00	1.00	0.96	0.96
Main effects	0.88	0.88	0.96^{ab}	0.94
All effects	1.00^{c}	1.00^{c}	0.94	0.99
		Positive pairwise correlation $\rho = 0.5$		
Weakest-link	0.98^{cd}	0.98^{cd}	0.82	0.78
Main effects	0.98^{cd}	0.96^{d}	0.90^{d}	0.81
All effects	0.94^{c}	0.95^{c}	0.70	0.95^{c}
Filtered search within subset of covariates with unadjusted $p < 0.05$				
Outcome generating model		Negative pairwise correlation $\rho = -0.5$		
Weakest-link	0.99^{c}	0.99^{c}	0.87	0.97^{c}
Main effects	0.70	0.71	0.97^{ab}	0.93^{ab}
All effects	0.98^{c}	0.98^{c}	0.92	0.99^{c}
		No pairwise correlation $\rho = 0$		
Weakest-link	1.00	1.00	0.97	0.99
Main effects	0.88	0.86	0.98^{ab}	0.96^{ab}
All effects	0.99^{c}	0.99^{c}	0.90	0.99^{c}
		Positive pairwise correlation $\rho = 0.5$		
Weakest-link	0.99^{cd}	0.99^{cd}	0.90	0.85
Main effects	0.98^{d}	0.97^{d}	0.95	0.89
All effects	0.95^{c}	0.92^{c}	0.72	0.98^{c}

Numbers: proportion of times that globally optimal pair detected correct outcome generating pair within 100 replicates.
QWL: quantile-stitching weakest-link; PWL: probit-stitching weakest-link
main: linear model with main effects only; interaction: linear model with interaction effects
significant difference $p < 0.05$ according to McNemar's test
^{a}: significantly higher than QWL; ^{b}: significantly higher than PWL
^{c}: significantly higher than main; ^{d}: significantly higher than interaction

Table 12: Observed proportions of replicates detecting correct pair out of 10 simulated normally-distributed covariates of size 100, one set of binary outcomes simultaneously generated from 3 covariate pairs

$\mathbf{X}$: 100×10 matrix, generated as follows:

$$(X_{\kappa_1}, X_{\kappa_2}) \sim N\big(0, \left[\begin{smallmatrix} 1 & -0.5 \\ -0.5 & 1 \end{smallmatrix}\right]\big); (X_{\kappa_3}, X_{\kappa_4}) \sim N\big(0, \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]\big)$$
$$(X_{\kappa_5}, X_{\kappa_6}) \sim N\big(0, \left[\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right]\big); X_k \sim N(0,1); k \notin \kappa \text{ i.i.d.}$$

There are a total of 3 outcomes generated by the sum of the 3 covariate pairs:

$$Y_{wl} \sim Bernoulli(logit^{-1}(log(10) * (\min(X_{\kappa_1}, X_{\kappa_2}) + \min(X_{\kappa_3}, X_{\kappa_4}) + \min(X_{\kappa_5}, X_{\kappa_6}))))$$
$$Y_{main} \sim Bernoulli(logit^{-1}(log(10) * (X_{\kappa_1} + X_{\kappa_2} + X_{\kappa_3} + X_{\kappa_4} + X_{\kappa_5} + X_{\kappa_6})/2))$$
$$Y_{int} \sim Bern(logit^{-1}(log(10) * (X_{\kappa_1} + X_{\kappa_2} + X_{\kappa_1}X_{\kappa_2} + X_{\kappa_3} + X_{\kappa_4} + X_{\kappa_3}X_{\kappa_4} + X_{\kappa_5} + X_{\kappa_6} + X_{\kappa_5}X_{\kappa_6})/2))$$

Data fitting model:	PWL	main	int	PWL	main	int
Exhaustive search through $\binom{10}{2}$ pairs						
Weakest-link generating model:	pair is globally optimal solution			pair has $p < 0.05$		
$(\kappa_1, \kappa_2); \rho = -0.5$):	0.02	0.00	0.04	0.23^{b}	0.09	0.17^{b}
$(\kappa_3, \kappa_4); \rho = 0$):	0.08	0.03	0.03	0.35^{bc}	0.22	0.25
$(\kappa_5, \kappa_6); \rho = 0.5$:	0.33	0.08	0.06	0.73^{c}	0.66^{c}	0.54
Avg pairs p-bonf < 0.05				6.77	10.2	8.06
Main effects generating model:	pair is globally optimal solution			pair has $p < 0.05$		
$(\kappa_1, \kappa_2); \rho = -0.5$:	0.0	0.0	0.0	0.07	0.21^{a}	0.17^{a}
$(\kappa_3, \kappa_4); \rho = 0$:	0.03	0.01	0.01	0.55	0.67^{ac}	0.60
$(\kappa_5, \kappa_6); \rho = 0.5$:	0.22^{bc}	0.05	0.05	0.93	0.93	0.91
Avg pairs p-bonf < 0.05				13.4	19.6	17.8
Interaction generating model:	pair is globally optimal solution			pair has $p < 0.05$		
$(\kappa_1, \kappa_2); \rho = -0.5$:	0.04	0.01	0.05	0.44^{b}	0.20	0.47^{b}
$(\kappa_3, \kappa_4); \rho = 0$:	0.11^{b}	0.04	0.13^{b}	0.65^{b}	0.52	0.76^{ab}
$(\kappa_5, \kappa_6); \rho = 0.5$:	0.23^{b}	0.06	0.32^{b}	0.90^{b}	0.80	0.96^{ab}
Avg pairs p-bonf < 0.05				10.2	14.1	13.2
Filtered search within subset of covariates with unadjusted $p < 0.05$						
Weakest-link generating model:	pair is globally optimal solution			pair has $p < 0.05$		
$(\kappa_1, \kappa_2); \rho = -0.5$:	0.05	0.04	0.07	0.19^{b}	0.09	0.16^{b}
$(\kappa_3, \kappa_4); \rho = 0$:	0.21^{bc}	0.05	0.09	0.46^{bc}	0.24	0.23
$(\kappa_5, \kappa_6); \rho = 0.5$:	0.20^{bc}	0.09	0.09	0.61^{c}	0.59^{c}	0.48
Avg pairs p-bonf< 0.05[3]				6.11	8.58	6.91
Main effects generating model:	pair is globally optimal solution			pair has $p < 0.05$		
$(\kappa_1, \kappa_2); \rho = -0.5$:	0.0	0.0	0.0	0.08	0.26^{a}	0.20^{a}
$(\kappa_3, \kappa_4); \rho = 0$:	0.01	0.01	0.02	0.63	0.75^{a}	0.70^{a}
$(\kappa_5, \kappa_6); \rho = 0.5$:	0.20^{bc}	0.02	0.02	0.93	0.96	0.90
Avg pairs p-bonf< 0.05[3]				13.1	19.7	18.0
Interaction generating model:	pair is globally optimal solution			pair has $p < 0.05$		
$(\kappa_1, \kappa_2); \rho = -0.5$:	0.05	0.02	0.05	0.39^{b}	0.17	0.42^{b}
$(\kappa_3, \kappa_4); \rho = 0$:	0.11^{b}	0.03	0.10^{b}	0.62^{b}	0.47	0.72^{ab}
$(\kappa_5, \kappa_6); \rho = 0.5$:	0.19^{b}	0.04	0.33^{ab}	0.83^{b}	0.75	0.94^{ab}
Avg pairs p-bonf< 0.05[3]				10.5	13.7	13.1

Numbers: proportion of replicates where the corresponding pair was identified as the true outcome-generating pair. PWL: probit-stitching weakest-link; main: linear model with main effects; all: linear model with main and interaction effects
[a]: significantly higher than PWL; [b]: significantly higher than main; [c]: significantly higher than interaction

Table 13: Observed proportions of replicates detecting correct pair out of 100 simulated normally-distributed covariates of size 100, 3 binary outcomes

$\mathbf{X}$: 100×100 matrix, generated in 100 replicates as follows:
$(X_{\kappa_1}, X_{\kappa_2}) \sim N\big(0, \left[\begin{smallmatrix} 1 & -0.5 \\ -0.5 & 1 \end{smallmatrix}\right]\big); (X_{\kappa_3}, X_{\kappa_4}) \sim N\big(0, \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]\big)$
$(X_{\kappa_5}, X_{\kappa_6}) \sim N\big(0, \left[\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right]\big); X_k \sim N(0,1); k \notin \kappa$ i.i.d.
There are 3 outcomes from the pair of covariates $(X_{\kappa_1}, X_{\kappa_2})$:
$Y^{(1)}_{wl} \sim Bernoulli(logit^{-1}(log(10) * \min(X_{\kappa_1}, X_{\kappa_2})))$;
$Y^{(1)}_{main} \sim Bernoulli(logit^{-1}(log(10) * (X_{\kappa_1} + X_{\kappa_2})/2))$;
$Y^{(1)}_{int} \sim Bernoulli(logit^{-1}(log(10) * (X_{\kappa_1} + X_{\kappa_2} + X_{\kappa_1} X_{\kappa_2})/2))$
and similarly define Y_{wl}, Y_{main} and Y_{int} for the covariate pairs indexed by (κ_3, κ_4) and (κ_5, κ_6).

Data fitting model:	QWL	PWL	main	int
Exhaustive search through $\binom{100}{2}$ pairs				
Outcome generating model			Negative pairwise correlation $\rho = -0.5$	
Weakest-link	0.86^{cd}	0.88^{cd}	0.48	0.77^{c}
Main effects	0.25	0.22	0.84^{abd}	0.71^{ab}
All effects	0.81^{c}	0.84^{c}	0.63	0.90^{ac}
			No pairwise correlation $\rho = 0$	
Weakest-link	0.92^{cd}	0.92^{cd}	0.76	0.84
Main effects	0.73	0.78	0.94^{abd}	0.86^{ab}
All effects	0.88^{c}	0.92^{c}	0.67	0.95^{ac}
			Positive pairwise correlation $\rho = 0.5$	
Weakest-link	0.95^{cd}	0.95^{cd}	0.58	0.51
Main effects	0.82^{cd}	0.79^{cd}	0.55^{d}	0.45
All effects	0.79^{c}	0.79^{c}	0.34	0.81^{c}
Filtered search within subset of covariates with Bonferroni $p < 0.1$				
Outcome generating model			Negative pairwise correlation $\rho = -0.5$	
Weakest-link	0.00	0.00	0.00	0.00
Main effects	0.07	0.06	0.09	0.09
All effects	0.02	0.02	0.02	0.02
			No pairwise correlation $\rho = 0$	
Weakest-link	0.71	0.71	0.68	0.71
Main effects	0.79	0.81	0.91^{ab}	0.90^{ab}
All effects	0.67^{c}	0.67^{c}	0.59	0.68^{c}
			Positive pairwise correlation $\rho = 0.5$	
Weakest-link	0.97^{cd}	0.97^{cd}	0.71	0.77
Main effects	0.89^{cd}	0.86^{cd}	0.76	0.77
All effects	0.88^{c}	0.88^{c}	0.61	0.92^{c}
Filtered search within subset of covariates with unadjusted $p < 0.05$				
Outcome generating model			Negative pairwise correlation $\rho = -0.5$	
Weakest-link	0.83^{cd}	0.87^{cd}	0.57	0.73^{c}
Main effects	0.34	0.31	0.88^{abd}	0.69^{ab}
All effects	0.88^{c}	0.90^{c}	0.68	0.93^{c}

... Table 13 continued

Data fitting model:	QWL	PWL	main	int
			No pairwise correlation $\rho = 0$	
Weakest-link	0.90^{cd}	0.93^{cd}	0.66	0.75
Main effects	0.68	0.69	0.90^{ab}	0.86^{ab}
All effects	0.93^{c}	0.93^{c}	0.80	0.98^{c}
			Positive pairwise correlation $\rho = 0.5$	
Weakest-link	0.91^{cd}	0.90^{cd}	0.50	0.45
Main effects	0.84^{cd}	0.85^{cd}	0.69^{d}	0.55
All effects	0.81^{c}	0.78^{c}	0.34	0.84^{c}
Greedy search				
Outcome generating model			Negative pairwise correlation $\rho = -0.5$	
Weakest-link	0.53^{cd}	0.55^{cd}	0.34^{d}	0.22
Main effects	0.17	0.13	0.69^{abd}	0.52^{ab}
All effects	0.51^{d}	0.56^{d}	0.48^{d}	0.33
			No pairwise correlation $\rho = 0$	
Weakest-link	0.89^{cd}	0.89^{cd}	0.73	0.79
Main effects	0.72	0.77	0.94^{abd}	0.86^{ab}
All effects	0.84^{c}	0.88^{c}	0.67	0.91^{c}
			Positive pairwise correlation $\rho = 0.5$	
Weakest-link	0.95^{cd}	0.95^{cd}	0.58	0.51
Main effects	0.82^{cd}	0.79^{cd}	0.55^{d}	0.45
All effects	0.79^{c}	0.79^{c}	0.34	0.81^{c}
Simulated annealing search (10,000 iterations)				
Outcome generating model			Negative pairwise correlation $\rho = -0.5$	
Weakest-link		0.75^{d}		0.56
All effects		0.86^{d}		0.74
			No pairwise correlation $\rho = 0$	
Weakest-link		0.91^{d}		0.83
All effects		0.93		0.84
			Positive pairwise correlation $\rho = 0.5$	
Weakest-link		0.96^{d}		0.41
All effects		0.75		0.75

Numbers: proportion of times that globally optimal pair detected correct outcome generating pair within 100 replicates.
QWL: quantile-stitching weakest-link; PWL: probit-stitching weakest-link
main: linear model with main effects only; interaction: linear model with main+interaction effects
significant difference $p < 0.05$ according to McNemar's test
^{a}: significantly higher than QWL; ^{b}: significantly higher than PWL
^{c}: significantly higher than main; ^{d}: significantly higher than interaction

Table 14: Observed mean number of significant pairs detected per replication, from 100 simulated normally-distributed covariates of size 100, 3 binary outcomes

$\mathbf{X}$: 100×100 matrix, generated in 100 replicates as follows:
$(X_{\kappa_1}, X_{\kappa_2}) \sim N\big(0, \left[\begin{smallmatrix} 1 & -0.5 \\ -0.5 & 1 \end{smallmatrix}\right]\big); (X_{\kappa_3}, X_{\kappa_4}) \sim N\big(0, \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]\big)$
$(X_{\kappa_5}, X_{\kappa_6}) \sim N\big(0, \left[\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right]\big); X_k \sim N(0,1); k \notin \kappa$ i.i.d.
There are 3 outcomes from the pair of covariates $(X_{\kappa_1}, X_{\kappa_2})$:
$Y^{(1)}_{wl} \sim Bernoulli(logit^{-1}(log(10) * \min(X_{\kappa_1}, X_{\kappa_2})))$;
$Y^{(1)}_{main} \sim Bernoulli(logit^{-1}(log(10) * (X_{\kappa_1} + X_{\kappa_2})/2))$;
$Y^{(1)}_{int} \sim Bernoulli(logit^{-1}(log(10) * (X_{\kappa_1} + X_{\kappa_2} + X_{\kappa_1} X_{\kappa_2})/2))$
and similarly define Y_{wl}, Y_{main} and Y_{int} for the covariate pairs indexed by (κ_3, κ_4) and (κ_5, κ_6).

Data fitting model:	QWL	PWL	main	int
Exhaustive search through $\binom{100}{2}$ pairs				
Outcome generating model		Negative pairwise correlation $\rho = -0.5$		
Weakest-link	0.73/1.03	0.77/1.07	0.19/0.57	0.57/0.91
Main effects	0.12/0.49	0.11/0.44	0.57^{abd}/2.30	0.47^{ab}/1.64
All effects	0.71^{c}/1.37	0.73^{c}/1.60	0.33/2.62	0.77^{c}/2.35
		No pairwise correlation $\rho = 0$		
Weakest-link	0.90^{cd}/3.83	0.94^{cd}/4.36	0.71/19.0	0.79^{c}/14.2
Main effects	0.84/6.49	0.85/8.03	0.96^{ab}/50.5	0.94^{ab}/42.5
All effects	0.91^{c}/3.75	0.95^{c}/4.32	0.68/25.1	0.95^{c}/19.3
		Positive pairwise correlation $\rho = 0.5$		
Weakest-link	0.98/36.1	0.98/42.7	0.97/159.6	0.97/148.9
Main effects	0.99/41.0	1.00/47.0	1.00/168.6	1.00/155.4
All effects	0.95^{c}/18.5	0.92^{c}/22.1	0.80/98.1	0.97^{c}/87.5
Filtered search within subset of covariates with Bonferroni $p < 0.1$				
Outcome generating model		Negative pairwise correlation $\rho = -0.5$		
Weakest-link	0.0/1.28	0.0/1.28	0.0/1.29	0.0/1.29
Main effects	0.09/2.21	0.09/2.21	0.09/2.24	0.02/2.18
All effects	0.02/1.39	0.02/1.40	0.02/1.42	0.02/1.41
		No pairwise correlation $\rho = 0$		
Weakest-link	0.72/2.85	0.72/2.85	0.72/2.85	0.72/2.86
Main effects	0.92/3.30	0.92/3.29	0.92/3.31	0.90/3.31
All effects	0.68/3.29	0.68/3.28	0.68/3.33	0.68/3.31
		Positive pairwise correlation $\rho = 0.5$		
Weakest-link	1.0/3.60	1.0/3.60	1.0/3.62	1.0/3.61
Main effects	0.99/3.50	0.99/3.52	0.99/3.52	0.99/3.51
All effects	0.99/3.65	0.99/3.69	0.99/3.72	0.99/3.65
Filtered search within subset of covariates with unadjusted $p < 0.05$				
Outcome generating model		Negative pairwise correlation $\rho = -0.5$		
Weakest-link	0.70^{cd}/0.94	0.74^{cd}/1.02	0.19/0.26	0.52^{c}/0.62
Main effects	0.13/0.34	0.15/0.36	0.58^{ab}/2.18	0.51^{ab}/1.22
All effects	0.75^{c}/1.20	0.77^{c}/1.37	0.38/0.82	0.86^{abc}/1.30

... Table 14 continued

Data fitting model:	QWL	PWL	main	int
		No pairwise correlation $\rho = 0$		
Weakest-link	0.92^{cd}/3.75	0.93^{cd}/4.29	0.72/20.9	0.78/16.1
Main effects	0.87/7.74	0.87/9.24	0.95^{ab}/54.2	0.95^{ab}/46.7
All effects	0.92^{c}/4.08	0.91^{c}/4.33	0.74/25.5	0.99^{abc}/19.2
		Positive pairwise correlation $\rho = 0.5$		
Weakest-link	0.98/35.9	0.98/40.7	0.97/153.8	0.96/137.8
Main effects	1.0/45.2	1.0/52.3	1.0/179.2	1.0/166.2
All effects	0.95^{c}/17.2	0.96^{c}/21.9	0.82/93.3	0.97^{c}/82.6
Greedy search				
Outcome generating model		Negative pairwise correlation $\rho = -0.5$		
Weakest-link	0.47^{cd}/0.76	0.50^{cd}/0.77	0.18/0.56	0.18/0.49
Main effects	0.09/0.39	0.09/0.24	0.50^{abd}/2.21	0.39^{ab}/1.52
All effects	0.49^{cd}/1.13	0.50^{cd}/1.32	0.29/2.58	0.33/1.89
		No pairwise correlation $\rho = 0$		
Weakest-link	0.87^{cd}/3.77	0.90^{cd}1/4.30	0.70/19.0	0.77^{c}/14.2
Main effects	0.83/6.42	0.83/7.93	0.96^{ab}/50.5	0.94^{ab}/42.4
All effects	0.86^{c}/3.65	0.90^{c}/4.22	0.68/25.1	0.90^{c}/19.3
		Positive pairwise correlation $\rho = 0.5$		
Weakest-link	0.98/36.0	0.98/42.4	0.97/146.3	0.97/134.2
Main effects	0.99/40.6	1.00/46.6	1.00/149.7	1.00/136.9
All effects	0.95^{c}/18.3	0.92^{c}/21.9	0.80/91.5	0.97^{c}/85.2
Simulated annealing search (10,000 iterations)				
Outcome generating model		Negative pairwise correlation $\rho = -0.5$		
Weakest-link		0.73^{d}/1.14		0.44/0.62
All effects		0.78/1.25		0.69/1.31
		No pairwise correlation $\rho = 0$		
Weakest-link		0.96^{d}/5.06		0.85/15.5
All effects		0.97/4.82		0.97/21.8
		Positive pairwise correlation $\rho = 0.5$		
Weakest-link		1.0/37.5		0.98/135.8
All effects		0.93/17.3		0.98/76.6

Numerators: proportion of replicates where outcome-generating pair had Bonferroni-adjusted p-value < 0.05. Denominators: mean number of pairs per replicate with Bonferroni-adjusted p-value < 0.05, out of 100 replicates.
QWL: quantile-stitching weakest-link; PWL: probit-stitching weakest-link
main: linear model with main effects only
all: linear model with main+interaction effects; significant difference $p < 0.05$ according to McNemar's test
^{a}: significantly higher than QWL; ^{b}: significantly higher than PWL
^{c}: significantly higher than main; ^{d}: significantly higher than interaction

Table 15: Observed proportions of replicates detecting correct pair out of 100 simulated right-skewed covariates of size 100, 3 binary outcomes

$\mathbf{X}$: 100×100 matrix, generated in 100 replicates as follows:
$(X^*_{\kappa_1}, X^*_{\kappa_2}) \sim N\big(0, \left[\begin{smallmatrix} 1 & -0.5 \\ -0.5 & 1 \end{smallmatrix}\right]\big); (X^*_{\kappa_3}, X^*_{\kappa_4}) \sim N\big(0, \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]\big)$
$(X^*_{\kappa_5}, X^*_{\kappa_6}) \sim N\big(0, \left[\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right]\big); X^*_k \sim N(0,1); k \notin \kappa$ i.i.d.
Use the transformation $X_k = e^{X^*_k}$ for all $k \in 1:100$,
to get positive-only covariates with a skewed right distribution.
There are 3 outcomes from the pair of covariates $(X_{\kappa_1}, X_{\kappa_2})$:
$Y^{(1)}_{wl} \sim Bernoulli(logit^{-1}(log(10) * log(\min\{X_{\kappa_1}, X_{\kappa_2}\})))$;
$Y^{(1)}_{main} \sim Bernoulli(logit^{-1}(log(10) * log(X_{\kappa_1} + X_{\kappa_2})/2))$;
$Y^{(1)}_{int} \sim Bernoulli(logit^{-1}(log(10) * log(X_{\kappa_1} + X_{\kappa_2} + X_{\kappa_1}X_{\kappa_2})/2))$
and similarly define Y_{wl}, Y_{main} and Y_{int} for the covariate pairs indexed by (κ_3, κ_4) and (κ_5, κ_6).

Data fitting model:	QWL	PWL	main	int
Exhaustive search through $\binom{100}{2}$ pairs				
Outcome generating model			Negative pairwise correlation $\rho = -0.5$	
Weakest-link	0.85^{cd}	0.82^{cd}	0.01	0.60^{c}
Main effects	0.77	0.72	0.86^{abd}	0.71
All effects	0.72	0.70	0.86^{abd}	0.76
			No pairwise correlation $\rho = 0$	
Weakest-link	0.94^{cd}	0.90^{cd}	0.21	0.54^{c}
Main effects	0.84	0.78	0.88^{bd}	0.78
All effects	0.92	0.86	0.93^{b}	0.89
			Positive pairwise correlation $\rho = 0.5$	
Weakest-link	0.94^{bcd}	0.76^{cd}	0.23	0.23
Main effects	0.90^{bcd}	0.75^{cd}	0.66^{d}	0.38
All effects	0.96^{bcd}	0.86^{cd}	0.72^{d}	0.53
Filtered search within subset of covariates with Bonferroni $p < 0.1$				
Outcome generating model			Negative pairwise correlation $\rho = -0.5$	
Weakest-link	0.01	0.01	0.01	0.01
Main effects	0.39	0.39	0.39	0.39
All effects	0.38	0.37	0.38	0.38
			No pairwise correlation $\rho = 0$	
Weakest-link	0.20^{c}	0.20^{c}	0.14	0.19
Main effects	0.87	0.85	0.86	0.85
All effects	0.93	0.88	0.94^{b}	0.95^{b}
			Positive pairwise correlation $\rho = 0.5$	
Weakest-link	0.85^{cd}	0.81^{cd}	0.49	0.56
Main effects	0.97^{bcd}	0.90^{c}	0.81	0.83
All effects	0.99^{cd}	0.98^{cd}	0.83	0.78
Filtered search within subset of covariates with unadjusted $p < 0.05$				
Outcome generating model			Negative pairwise correlation $\rho = -0.5$	
Weakest-link	0.79^{c}	0.75^{c}	0.00	0.49^{c}
Main effects	0.71	0.70	0.93^{abd}	0.77
All effects	0.59	0.58	0.85^{abd}	0.73^{ab}

... Table 15 continued

Data fitting model:	QWL	PWL	main	int
			No pairwise correlation $\rho = 0$	
Weakest-link	0.95^{bcd}	0.87^{cd}	0.14	0.51^{c}
Main effects	0.89	0.86	0.91	0.86
All effects	0.88	0.85	0.98^{abd}	0.90
			Positive pairwise correlation $\rho = 0.5$	
Weakest-link	0.96^{bcd}	0.77^{cd}	0.30	0.23
Main effects	0.88^{bcd}	0.75^{d}	0.69^{d}	0.56
All effects	0.97^{bcd}	0.84^{d}	0.78^{d}	0.61
Greedy search				
Outcome generating model			Negative pairwise correlation $\rho = -0.5$	
Weakest-link	0.58^{bcd}	0.38^{cd}	0.01	0.12^{c}
Main effects	0.58	0.61	0.83^{abd}	0.69
All effects	0.52	0.61	0.80^{abd}	0.61
			No pairwise correlation $\rho = 0$	
Weakest-link	0.89^{bcd}	0.78^{cd}	0.21	0.43^{c}
Main effects	0.82	0.78	0.88^{bd}	0.77
All effects	0.90	0.85	0.93^{b}	0.89
			Positive pairwise correlation $\rho = 0.5$	
Weakest-link	0.94^{bcd}	0.77^{cd}	0.23	0.23
Main effects	0.90^{bcd}	0.75^{d}	0.66^{d}	0.38
All effects	0.96^{bcd}	0.86^{cd}	0.72^{d}	0.53

Numbers: proportion of times that globally optimal pair detected correct outcome generating pair within 100 replicates.
QWL: quantile-stitching weakest-link; PWL: probit-stitching weakest-link
main: linear model with main effects only; interaction: linear model with main+interaction effects
significant difference $p < 0.05$ according to McNemar's test
^{a}: significantly higher than QWL; ^{b}: significantly higher than PWL
^{c}: significantly higher than main; ^{d}: significantly higher than interaction

6.2.3 Sample plots of covariates by outcome type

Figure 8 shows one of the randomly generated sets of binary outcomes generated from a weakest-link model $Y_{wl} \sim \text{Bernoulli}(pr)$. We obtain pr from the anti-logit of the weakest-link between (X_{20} and X_{95}). The pair (X_{20}, X_{95}) was generated from a bivariate standard normal distribution, with $\rho = 0.5$. The outcomes Y_{wl} were highly associated with both probit weakest-link ($p = 8 \times 10^{-5}$) and interaction models (p=2.8×10^{-3}). Due to a few more points located on the top-center portion of the plot, relative to the right-center portion, the overfitting from the interaction model tilt the corresponding contour in a slightly different direction than that of the true weakest-link generating model. This suggests that

the interaction model is not an ideal fit, leaving the variable combination search methods open to detecting false positives from other combinations. Thus, models with other pairs as predictors may fit the outcomes better simply by chance.

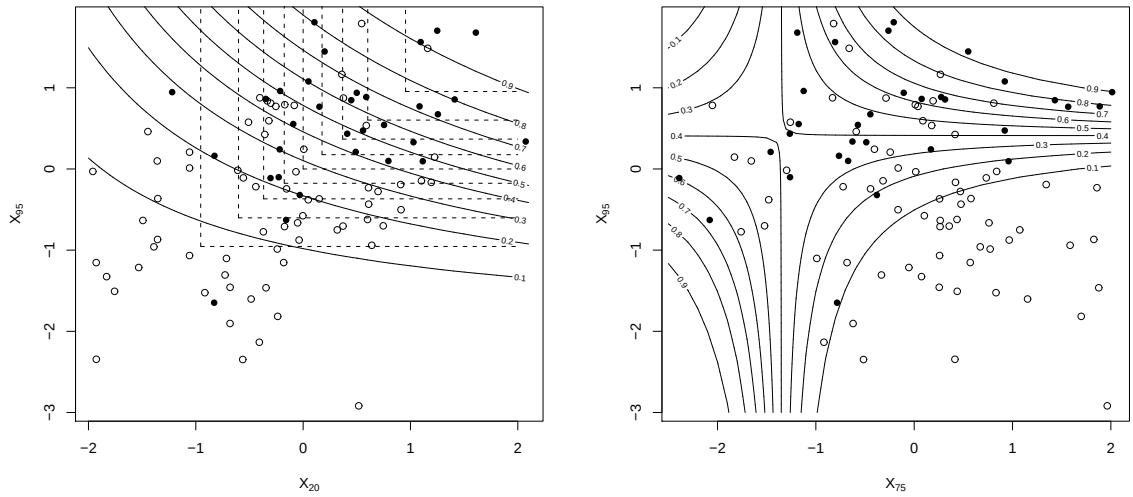

Figure 8: Sample scatterplot of weakest-link generated outcomes

Scatterplot of simulated weakest-link generated outcomes, with normally distributed covariates. Positive outcomes are in black, negative in white. Equivalent values of the probability $\hat{pr}$ of a positive outcome, predicted by the interaction model, are on the solid curved contour lines. Left: Plot of X_{20} vs X_{95}, the pair used to generate outcomes with correlation $\rho = 0.5$. Equivalent weakest-link values, being the true probability for generating a positive outcome, pr, are on the straight contour lines. Right: Plot of X_{75} vs X_{95}, which the interaction model identifies as having the greatest association with the outcomes even though it did not generate the outcomes.

A linear model with interaction term found another pair ($X_{75} * X_{95}$) to be more highly associated with Y_{wl} ($p = 9.5 \times 10^{-5}$). X_{75} followed a standard normal distribution independent from X_{95}. The closer fit of the latter plot appears to be due to a large number of negative outcomes in the lower right-hand corner. At lower values of X_{75}, the interaction model thus incorrectly concludes higher values of X_{95} to be negatively associated with positive outcomes. Evidently, the interaction model is overfitting. The simulations show that the overfitting scenario is a common problem when using an interaction model to fit data where the outcomes follow weakest-link behavior. Overfitting and the consequent false

positives from non-outcome generating combinations may be inevitable in high-throughput data sets with a huge number of covariate combinations.

Figure 9 shows one of the randomly generated sets of binary outcomes generated from a linear model with interaction term $Y_{int} \sim$ Bernoulli(pr). We obtain pr from the anti-logit of the interaction model $X_{20} + X_{95} + X_{20} * X_{95}$. The pair (X_{20}, X_{95}) was generated from a bivariate standard normal distribution, with $\rho = 0.5$. The outcomes Y_{int} were highly associated with both the correct probit weakest-link (p=8×10^{-3}) and interaction models (p=7.6×10^{-6}). However, the weakest-link model with interaction term found another pair ($X_{20} * X_{94}$) as being slightly more highly associated with Y_{int} (p=4.0×10^{-3}). X_{94} followed a standard normal distribution independent from X_{20}. The weakest-link model correctly identified higher values of X_{20} to be associated with positive outcomes in both models. However, the weakest-link model incorrectly found that (X_{20} and X_{94}) more closely resembled a weakest-link relationship with the outcomes. The simulations showed that this scenario, where the interaction model is correct and weakest-link is not, is not as common as that where the weakest-link model is correct and interaction model is wrong.

The advantages of the weakest-link model are very apparent when X is skewed. This situation can occur, for example, when using raw biomarker levels which are positive. In this situation, the contours all follow a similar shape to those in upper-right hand quadrant of Figure 10. In this quadrant, the contours of the weakest-link and interaction models more closely resemble each other. Thus, the weakest-link model should more easily fit an interaction model, while avoiding the pitfalls of the interaction model and possible overfitting.

In another situation, with covariates uniformly generated in the $(0, 1)$ interval, the weakest-link model tended to have better detection rates than the interaction model, even though the interaction model was used to generate the data. An example is Figure 11, with randomly generated sets of binary outcomes generated from an interaction model $Y_{int} \sim$ Bernoulli(pr). We obtain pr from the pair (X_{20}, X_{95}), which has a bivariate uniform distribution on the univariate interval, with $\rho = 0.5$. The outcomes Y_{int} were highly associated with both probit weakest-link (p=1.1×10^{-7}) and interaction models (p=7.3×10^{-8}). Regardless, the contours from the fitted interaction model do not entirely point in the same direction as those from the true outcome-generating interaction model.

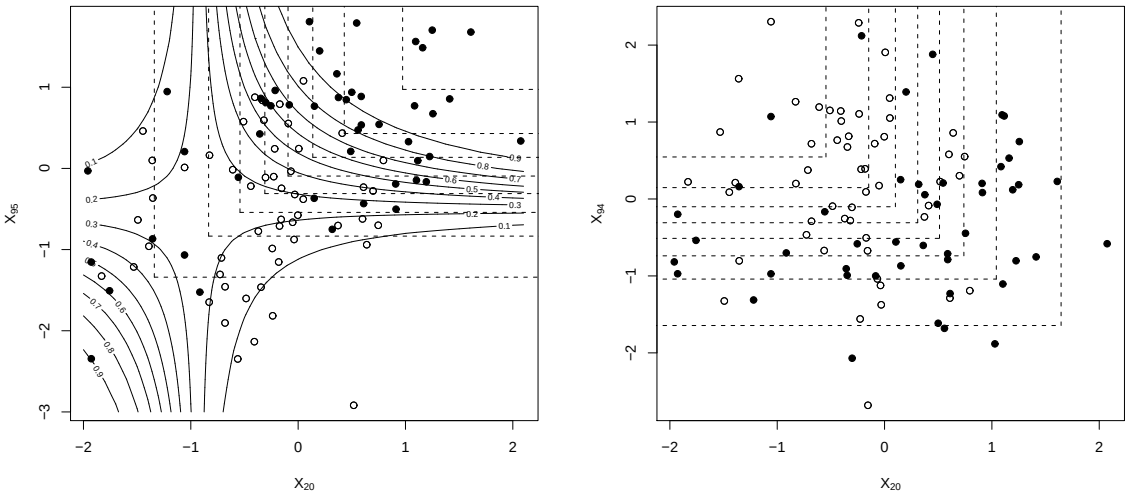

Figure 9: Sample scatterplot of interaction model generated outcomes

Scatterplot of simulated interaction model generated outcomes, with normally distributed covariates. Positive outcomes are in black, negative in white. Left: Plot of X_{20} vs X_{95}, the pair used to generate outcomes with correlation $\rho = 0.5$. Equivalent values of the probability $\hat{pr}$ of a positive outcome, predicted by the interaction model, are on the curved solid contour lines. Equivalent probabilities for generating a positive outcome by the weakest-link model are on the straight contour lines. Right: Plot of X_{20} vs X_{94}. Equivalent values of $\hat{pr}$ predicted by the weakest-link model are on the straight contour lines. The weakest-link model incorrectly identifies this pair as having the best association with the outcomes.

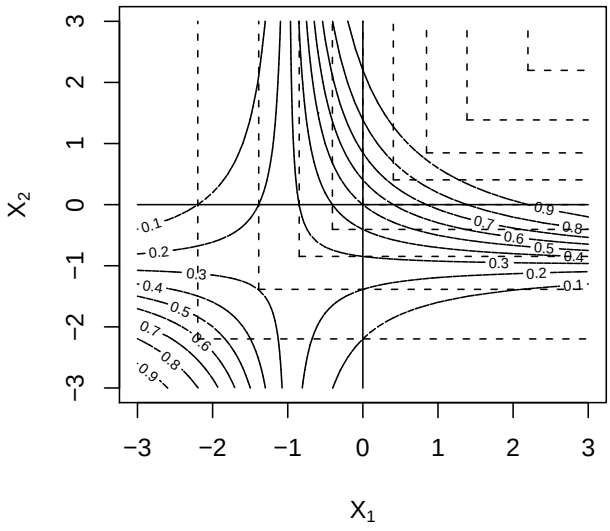

Figure 10: Comparison of contours from weakest-link and interaction models

Contour plots of pr, for generating outcomes $Y \sim \text{Bernoulli}(pr)$. Equivalent values from the weakest-link model $p = \{X_1, X_2\}$ are on the straight dotted contour lines and those from the interaction model $p = X_1 + X_2 + X_1X_2$ are on the curved contour lines.

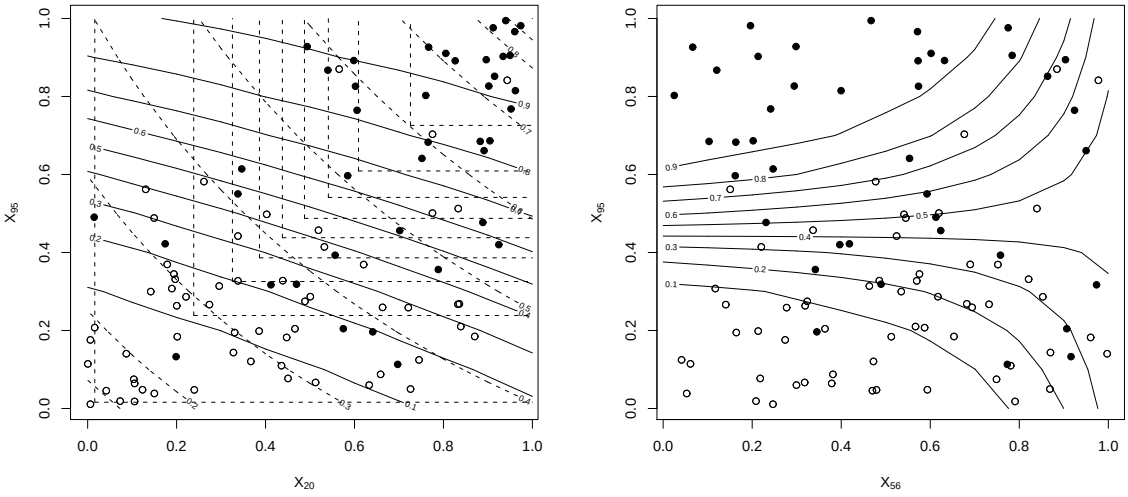

Figure 11: Sample scatterplot of interaction model generated outcomes, uniform covariates

Scatterplot of simulated interaction model generated outcomes, with covariates uniformly generated in the $(0, 1)$ interval. Positive outcomes are in black, negative in white. Equivalent values of the predicted $\hat{pr}$ of a positive outcome, from a fit linear model with interaction term, are plotted on the solid contour lines. Left: Plot of X_{20} vs X_{95}. Equivalent true probabilities for generating a positive outcome, pr by the interaction model are plotted on the dotted curved contour lines, while equivalent values of the weakest-link covariate are on the straight dotted contour lines. Right: Plot of X_{56} vs X_{95}. The interaction model incorrectly identifies this pair as having the best association with the outcomes, while the weakest-link model does not, even though the interaction model was used to generate the outcomes.

However, a linear model with interaction term found another pair (X_{56}, X_{95}) to be more highly associated with Y_{int} (p=3.6×10^{-9}). Similarly to Figure 8, the interaction model $(X_{56} * X_{95})$ provides a very strong fit to the data, that is nevertheless incorrect. It also does not illustrate the sort of relationship one usually pictures when using an interaction model. On the other hand, the weakest-link model correctly found (X_{20}, X_{95}) to be the pair most associated with Y_{int}, even though the interaction model actually generated the outcomes.

Figure 12 shows a typical plot when the outcome-generating covariates have negative correlation. The marginal associations between both covariates and the outcome are clearly not monotonic and would not be detected by main effects models. In these situations, there tended to be little difference in the efficiency of the weakest-link and interaction models in detecting the outcome-generating pair. Additionally, overfitting does not appear to be as much an issue due to the observations running from the lower right to the upper left of the plot. This provides more observations with similar $\hat{pr}$, and thus less variability and possible error.

6.2.4 Computation times

Even with modern fast computing, computation time is still an issue screening through high-throughput data, where there are many potential covariates of interest. This issue is more apparent when screening for combinations of 2 or more pairs, due to the sheer numbers of possible covariates. As such, we are interested in approaches which cut down on computation time.

With binary outcomes, we consider two different approaches to scoring the association of a joint effect with an outcome. For simplicity's sake, denote the binary outcome as being either 0 or 1.

Logistic regression: Use the likelihood ratio test from a logistic regression model, with the binary outcome as the dependent variable and the weakest-link covariate as the independent variable.

Two-sample t-test: Use the t-statistic from a two-sample t-test, comparing the mean of weakest-link covariates in those with outcome=1 to those with outcome=0. This approach does not require the fitting of linear models and, as such, should be much faster than the

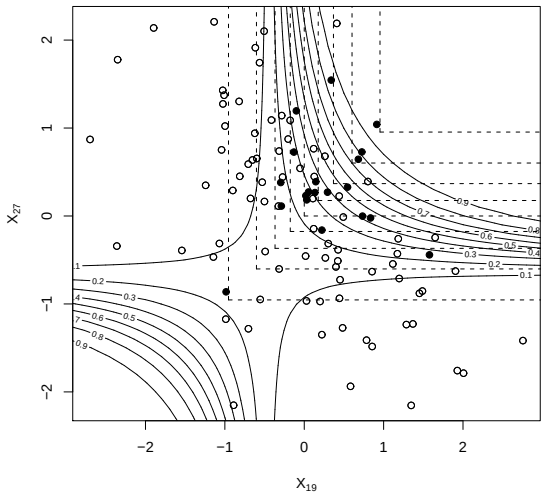

Figure 12: Sample scatterplot of outcomes generated by neg. correlated covariates

Scatterplot of simulated outcomes generated by a weakest-link of X_{19} vs X_{27}, with negative correlation $\rho = -0.5$. Positive outcomes are in black, negative in white. Equivalent values of the predicted $\hat{pr}$ of a positive outcome, from a fit linear model with interaction term, are plotted on the solid contour lines. Equivalent true probabilities of the weakest-link covariate are on the straight dotted contour lines. Both models correctly identified (X_{19}, X_{27}) as the pair most associated with the outcomes. Neither covariate is marginally associated with the outcomes.

logistic regression approach, but are appropriate only when the weakest-link covariates closely follow a normal distribution.

There are many situations where a logistic regression model is preferable to the t-test with respect to testing binary outcomes. For example, a two-sample test is usually not appropriate for a cohort study in relation to a binary disease outcome, especially if there is low prevalence in one group. Also, we do not estimate quantities of interest such as disease probability or relative risk when using a t-test. On the other hand, in a large-scale screening setting of microarray data with many thousands of covariates, the goal is usually to find covariate or covariate combinations that are potentially interesting for closer investigation, rather than for formal hypothesis testing. Additionally, in such a controlled setting we would expect to have a more even split of outcome, for example, normal vs diseased tissue or some sort of other outcome. Thus in large-scale screening, advantages in computational time for the t-test in quickly quantifying the association between a binary outcome and weakest-link covariate would be more important than finding precise models.

We can use other simple association procedure for other outcome types. For example, for continuous outcomes of interest we can use simple correlation instead of linear regression for scoring criteria.

In Table 16 we compare the use of these two methods of scoring the association between probit weakest-link covariates and binary outcomes: a t-stat from a two-sample t-test compared to the likelihood ratio test from logistic regression in terms of identifying the true outcome-generating pair. We use a similar setting to that in Section 6.2.2, with 100 covariates of size 100, and separate binary outcomes linked to 3 covariate pairs with correlation $\rho = -0.5, 0, 0.5$.

We also compare the computation times of quantile and probit stitching weakest-link methods. Probit stitching methods should provide some computation advantages as they do not require the estimation of empirical CDFs, and thus do not require the data to be ranked in a computationally-expensive sorting step.

To measure how much computation time saved by each of these approaches, we screened for pairs in a moderately size data set. We used a 50x100 matrix, or 100 covariates of size 50, and 50 binary outcomes. We carried out an exhaustive search of the best fitting pair, out

Table 16: T-test vs logistic regression: observed proportions of replicates detecting correct pair out of 100 simulated covariates of size 100, 3 binary outcomes.

$\mathbf{X}$: 100×100 matrix, generated in 100 replicates as follows:

$$(X_{\kappa_1}, X_{\kappa_2}) \sim N\big(0, \left[\begin{smallmatrix} 1 & -0.5 \\ -0.5 & 1 \end{smallmatrix}\right]\big); (X_{\kappa_3}, X_{\kappa_4}) \sim N\big(0, \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]\big)$$
$$(X_{\kappa_5}, X_{\kappa_6}) \sim N\big(0, \left[\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right]\big); X_k \sim N(0,1); k \notin \kappa \text{ i.i.d.}$$

There are 3 outcomes from the pair of covariates $(X_{\kappa_1}, X_{\kappa_2})$:

$$Y_{wl}^{(1)} \sim Bernoulli(logit^{-1}(log(10) * \min(X_{\kappa_1}, X_{\kappa_2})))$$
$$Y_{main}^{(1)} \sim Bernoulli(logit^{-1}(log(10) * (X_{\kappa_1} + X_{\kappa_2})/2))$$
$$Y_{int}^{(1)} \sim Bernoulli(logit^{-1}(log(10) * (X_{\kappa_1} + X_{\kappa_2} + X_{\kappa_1} X_{\kappa_2})/2))$$

and similarly define Y_{wl}, Y_{main} and Y_{int} for the covariate pairs indexed by (κ_3, κ_4) and (κ_5, κ_6).

Hypothesis testing model:	logistic regression	two-sample t-test
Exhaustive search through $\binom{100}{2}$ pairs		
Outcome generating model	Negative pairwise correlation $\rho = -0.5$	
Weakest-link	0.96	0.97
Main effects	0.73	0.73
All effects	0.86	0.88
	No pairwise correlation $\rho = 0$	
Weakest-link	0.96	0.97
Main effects	0.88	0.90
All effects	0.76	0.72
	Positive pairwise correlation $\rho = 0.5$	
Weakest-link	0.96	0.97
Main effects	0.88	0.90
All effects	0.76	0.72

Numbers: proportion of times, within 100 replications, that globally optimal pair detected correct outcome generating pair.
logistic regression: use logistic regression model to associate weakest-link derived covariate with binary outcomes, hypothesis testing by likelihood ratio test
two-sample t-test: obtain weakest-link derived covariates, use t-test to compare WL covariates in those with outcome= 1 with those with outcome= 0
significant difference $p < 0.05$ according to McNemar's test

of $\binom{100}{2} = 4950$ possible combinations using both t-test and logistic regression. We also used both approaches on quantile and probit-matching weakest-link derived covariates. We also repeated each search using logistic regression models, one with main effects only and another with interaction effects also. As the weakest-link models required the fitting of 4 univariate logistic regression models for each combination, to account for possible directions (Section 4.2.2), the use of t-test functions may help in reducing some of this added computational time.

Thus, we ran 4950 different model fits for each screening of pairs, and repeated each pair screening procedure 10 times to account for possible computing effects. These screening procedures were performed using the package R on a Dual Core 2.5 Ghz laptop, with computation times (Table 17) assessed using the *sys.time* command.

Table 17: Computation time of methods in screening for pairs

Computation time (in seconds)		
	mean±SD	mean±SD
Additive linear models		
main effects only	43.0±0.3	
main+int effects	44.0±0.2	
Weakest-link models		
Testing method:	Quantile-stitched WL	Probit-stitched WL
logistic regression (4 directionalities)	162.1±1.3	159.2±0.9
per direction	40.5±0.6	39.8±0.5
t-test (4 directionalities)	65.4±0.3	63.4±0.6
per direction	16.4±0.2	15.9±0.3

Reported are the mean ± standard deviation of the computation times for identical 10 searches using each method. For weakest-link searches, we also list the same information per direction.

The fitting of logistic regression models took most of the computational time of the searches. The searches which used weakest-link models took slightly less than 4 times to run

relative to the times of the additive model searches. As they required only one covariate to be fit, the weakest-link searches had a slight savings per directionality of 2-4 seconds.

As suspected, the probit stitching weakest-link was slightly faster than quantile stitching weakest-link. However, this computational advantage was small, averaging 2 to 3 seconds for 4950 model fits. For a search in a high-throughput data set that requires, for example, one million possible combinations, this advantage may grow to a few minutes. However, considering the large amount of time that fitting a million models would require, this computational advantage may not be relevant.

The computational advantages are greater in using the t-test for variable screening instead of the logistic regression model. Using the t-test cut the computational time by more than half. As there are no noticeable differences in detection rates between the t-test and logistic regression methods, using a t-test for large-scale screening may make weakest-link methods even more practical. Despite requiring the investigation of 4 directions for each combination, the t-test only required roughly 50% more computation time than the additive linear models. However, if we are able to specify an *a priori* direction for the weakest-link model that does not require searching through all directions, either of the weakest-link searches provide small, to great, benefits in computational time.

6.2.5 Discussion of simulation studies

This section reports and discusses general observations from the simulation studies. In general, we quantify the efficiency of each of the methods by the proportion of times that a given method correctly identified the true outcome-generating pair.

It is important to note the preceding observations apply to these situations where outcomes are positively correlated with each covariate. If outcomes are positively correlated with X_1 and negatively correlated with X_2, the implications of the sign of the pairwise correlation between covariates would be reversed. Regardless, the interpretations in terms of the relative efficiencies of the methods otherwise are unchanged.

6.2.5.1 Comparisons between models We focus on the exhaustive search through all possible pairs to compare the effectiveness of each of the models in finding the true

outcome-generating pair. This isolates the focus on the role of model misspecification rather than the role of search technique. Low detection rates can also be due to overfitting, which poses a greater problem in linear models that require more coefficients. Overfitting may find that another pair, not used to generate the outcomes, nevertheless has a higher statistical association with the outcome. The risk of this occurring is obviously higher in a high-throughput data set, as a result of the massive amount of possible combinations.

We were able to identify many situations where the weakest-link model was more effective in screening data sets than additive linear models for the outcome-generating pair.

For outcomes generated from positively correlated covariates, weakest-link methods perform better than additive models. For outcomes generated from a weakest-link model, both quantile and probit-stitching models detected the generating pair significantly better than the additive models. Weakest-link models performed better than main effects models even when the main effects model generated the outcome. When an additive model with an interaction term generated the outcome, the weakest-link models had success rates of finding the generating outcome that were slightly, but non-significantly, lower than those of the interaction models.

False positives are more frequent in additive models, specifically those with an interaction term. As shown in Tables 10 and 14, unless outcome-generating pairs were negatively correlated, additive models tended to detect many more other pairs that are significantly, and incorrectly, associated with the outcome. This suggests that overfitting is a main reason why weakest-link models were better than the others at finding pairs of positively correlated covariates truly associated with the outcome.

The probit-stitched weakest-link model is more appropriate for normally distributed data. For normally distributed covariates the probit-stitched weakest link models had detection rates slightly, and non-significantly higher, than the quantile-stitched models. However, when covariates were skewed, quantile-stitched models performed better, specifically when covariates were positively correlated. This follows the general usage of nonparametric rank-based methods, which are advisable for non-normally distributed data, with small loss of power for normally distributed data.

For outcomes generated from negatively correlated or non-correlated covariates, the model should be correctly specified. Weakest-link models performed significantly better than either additive models for weakest-link generated data. In particular, the main effects models had very poor detection rates for weakest-link generated outcomes.

When an additive model with only main effects generated the outcomes, the main effects model generally provided the best detection rates, significantly higher than the other models. This was specifically true when data was negatively correlated, and also for normally distributed data. The differences were less for the skewed data, mostly because the performance of weakest-link models improved in these situations.

When an additive model with interaction term generated the outcomes from pairs with normally distributed data, weakest-link models had detection rates that were slightly and non-significantly lower than the correct interaction fitting model. However, for skewed data, the weakest-link models had much higher detection rates than either of the additive models.

If weakest-link models are already considered, additive models with interaction term did not provide any practical advantages relative to an additive model with only main effects. When there was no interaction effect present, or in other words, the main effects model generated the outcomes, the main effects model had better detection rates than the interaction model. With normally distributed data, the interaction model had better detection rates than the main effects model in a few cases: either when the weakest-link model generated the outcome, or when the interaction model was the true model.

However, in all situations where the interaction model had higher detection rates than the main effects model, it did not have significantly higher rates than the weakest-link model. This was true even when the interaction model generated outcomes. The effectiveness of the interaction model was even lower in the skewed covariate examples, in which interaction models did not provide any advantage in any scenario.

When the relationship of one covariate with the outcome is altered by another covariate, weakest-link models may be more appropriate than the conventional additive model with interaction term. In cases where the interaction model truly

generates the data, the weakest-link model still provided detection rates that were only slightly lower, or in many cases better, than the interaction model.

When the main effects model generated the data, or in other words when the effect of one covariate on the outcome is not affected by the other covariate, the main effects model had better results than the weakest-link model. This finding is not surprising, because the weakest-link model is not an additive relationship at all; the identity of the weakest-link covariate explicitly requires that it only affects the outcome in a subset of covariate spaces, with the other covariates having no effect.

The distribution of the covariates in the feature set affected detection rates according to model type. When covariates where normally distributed the additive models tended to be less effective in finding the outcome generating pair, while the quantile weakest-link model tended to be slightly more effective when covariates were all positive and skewed.

None of the methods were able to detect multiple pairs of covariates that each were independently associated with the same outcome. Despite having relatively high sample size of $n = 100$ and only $p = 10$ covariates (Table 12), none of the methods identified any of the 3 true outcome-generating pairs in more than half of the replicates. Conversely, with the same 100 by 10 covariate matrix, but with 3 separately generated sets of outcomes each generated by the same pairs, the correct generating model detected the correct model close to 100% of the time.

This finding suggests that these methods may be more practical in finding one optimal combination of a larger subset of covariates, rather than multiple combinations of covariates.

6.2.5.2 Comparison of different search algorithms

For high-throughput data sets, exhaustive searches are often not feasible through all possible combinations. We investigated the effectiveness of 3 algorithms for reducing computation time: a filtered search, a greedy search and a simulated annealing-like method. These methods are designed to skip over covariates less likely to be part of the true outcome-generating pair, either by filtering out those with a poor marginal association with the outcome, or finding a locally optimal combination.

Filtered searches were mostly effective if the covariates were positively correlated, and may not be very practical for high-throughput data sets. When a search was performed within a subset of covariates with association with outcome of $p < 0.05$ filtered searches tended to provide detection rates that were close to those from the exhaustive searches. However, in high-throughput data sets many thousands of covariates may be correlated with the outcome of interest. In Section 6.1.1.1, there were 2,936 biomarkers with association with lymph node positivity of $p < 0.05$; a search within this subset would require more than 4 million fits. A more conservative p-value may thus be required to keep the subset and computational time within feasible limits for high-throughput data sets.

For Tables 13 and 15 that used a more stringent $p < 0.001$ for the filtering subset, which corresponds to a Bonferroni-adjusted $p < 0.1$, the detection rates were much lower than those from the exhaustive search. This makes sense, as covariates could not both have a high marginal association with the outcomes if they were negatively correlated with each other, and would not pass the filtering stages. Thus, in high-throughput data sets, filtering stages with these very conservative p-value thresholds may have poor detection rates.

Greedy searches were more effective if the covariates were positively correlated. For negatively correlated covariates, the greedy search was occasionally able to detect the true covariate pair, but this detection rate was still noticeably lower than the exhaustive search. This suggests the greedy search has a tendency to stay in a locally optimal state in these situations. As such, it may be necessary to use the time-consuming search for multiple locally optimal pairs. However, for positive- or non-correlated data, the greedy search performs almost as well as the exhaustive search.

Simulated annealing may be preferable to finding an optimal set of covariates. Like greedy searches, simulated annealing methods had detection rates very close to the exhaustive searches for non-correlated or positively correlated covariates. They performed slightly better than the greedy algorithm for negatively correlated data, though rates were still slightly lower than those from the exhaustive search. However, the computation-intensive nature of this procedure also requires fast computing.

7.0 ASSESSING JOINT EFFECTS IN A CYTOMETRY DATA SET

In this chapter, we adapt the previously introduced weakest-link methods to a data set from a cytometry study.

7.1 PENNSYLVANIA LUNG CANCER DATA SET

The study consists of three panels measuring the DNA content and biomarker levels in cells from lung tumor tissue samples. Dr. Stanley Shackney at the Allegheny General Hospital provided the data set from a multiparameter laser-scanning cytometry study of lung cancer patients, collected under a grant from the Pennsylvania Department of Health, Tobacco Formula Funds Grant #ME-01-334.

Biomarker levels within the cell were assessed through the fluorescent emission of three channels with different colors. In Panel 1, the green channel measured her-2/neu, the orange channel p53, and the long red channel Ras. In Panel 2, the same colors were used to identify her-2/neu, EGFR, and VEGF, respectively, while in Panel 3 they were p16, Rb3 and Cyclin E, respectively. The DNA content in each cell was also assessed in each panel to assess aneuploidy. Similar three-color fluorescence measurements were performed in jurkat cells to assess the effects of the day of the measurements.

To account for systematic day effect, the median intensity within jurkat cells was assessed during each day for each color. For each of the m cells, the biomarker intensities were then adjusted for day effect. In all 3-colors used to measure biomarkers, we divided the intensity by the median intensity from the jurkat cells corresponding to the same color from the same

day; for DNA content, we divided the intensity by the peak of a smoothed DNA density plot from the corresponding day.

As the values were skewed to the right, we used the transformation $\log(x+1)$ of these adjusted intensities for all calculations.

7.1.1 Results using quantile stitching weakest-link

Median recurrence-free survival (RFS) was assessed in a group of 57 patients in the Pennsylvania lung cancer data set using stitched weakest-link methods in two separate panels of 3 biomarkers, and DNA content (Tables 18 and 19). There were no noticeably significant results for markers in Panel 3, and as such, the results are not printed here. The superscript notes the set of directions D of the best fitting weakest-link model. We used Cox proportional hazards to model recurrence-free survival due to right-censoring of some of the observations, with p-values from the likelihood ratio test versus the null model. The label $a * b$ is used for models with two variables a and b that include both main effects and the interaction term. For weakest-link models, the Bonferroni adjustment consisted of multiplying the p-value by the number of directionality comparisons. For example, the p-value was multiplied by $16 = 2^4$ for weakest-link models between 4 covariates.

In Panel 2, the weakest-link between combinations of two covariates achieved the best results in terms of the cross-validated partial log-likelihood, *cvl*. The model with the lowest *cvl* was the within-cell weakest-link between two biomarkers from Panel 2: her-2/neu and DNA. The weakest-link models had slightly better fits when the weakest-link was taken at the intracellular level, rather than at the patient level. Conventional linear models were not able to successfully take into account interaction effects, with no interaction terms being statistically significant. The linear model that included all 4 main effects, and all interaction effects, had overfitting problems, noted by the high *cvl*. The weakest-link models for higher-order interactions between all 4 covariates had a relatively low *cvl* and did not have the overfitting that was apparent in the linear models.

Logic regression obtained a model consisting of dichotomous predictors. We used the package *LogicReg* in *R*. The results of the best combination of dichotomous predictors according to logic regression, with the cross-validated log-likelihood values, are also summarized

Table 18: Recurrence-free survival: results from Panel 1

Recurrence-free survival					
Model	**# covars**	**-LL**	**cvl**	**P(unadj)**	**P(Bonf)**
Linear model					
Null	0	56.78			
$\underset{j\in 1:m_i}{\text{mean}}\left[\hat{F}(\text{her-2/neu})\right]$	1	55.57	73.24	0.499	
$\underset{j\in 1:m_i}{\text{mean}}\left[\hat{F}(\text{p53})\right]$	1	54.72	72.00	0.142	
$\underset{j\in 1:m_i}{\text{mean}}\left[\hat{F}(\text{Ras})\right]$	1	53.72	70.93	0.041	
$\underset{j\in 1:m_i}{\text{mean}}\left[\hat{F}(\text{DNA})\right]$	1	55.73	73.47	0.720	
Main effects	4	53.17	74.51	0.262	
All effects	15	47.13	121.78	0.299	
Weakest-link (t_i) between four markers on Panel 1					
$\underset{k\in Panel1}{\text{wl}^{(+--+)}}\left\{\underset{j\in 1:m_i}{\text{mean}}\left[\hat{F}_k(x_{j(i)k})\right]\right\}$	1	53.51	69.69	0.011	0.165
$\underset{j\in 1:m_i}{\text{mean}}\left[\underset{k\in Panel1}{\text{wl}^{(+--+)}}\left\{\hat{F}_k(x_{j(i)k})\right\}\right]$	1	52.15	69.82	0.007	0.111
Best fitting model using logic-regression					
$I\big(<\text{median}\left\{\underset{j\in 1:m_i}{\text{mean}}\left[\hat{F}(\text{her-2/neu})\right]\right\}\big)$	1	52.99	70.88	0.018	
and $I\big(\geq\text{median}\left\{\underset{j\in 1:m_i}{\text{mean}}\left[\hat{F}(\text{DNA})\right]\right\}\big)$					
or $I\big(\geq\text{median}\left\{\underset{j\in 1:m_i}{\text{mean}}\left[\hat{F}(\text{p53})\right]\right\}\big)$					

Linear and weakest-link methods for modeling recurrence-free survival, Panel 1 (n=55)
For pairwise linear models, $*$ symbol denotes model with main effects and interaction
$\text{wl}^{(D)}$: weakest-link with directional vector (D)
-LL: -partial log-likelihood
CVL: cross-validated partial log-likelihood
P(unadj): Unadjusted p-values from Cox proportional hazards regression
P(Bonf): For weakest-link models, Bonferroni-adjusted p-values for number of directions

Table 19: Recurrence-free survival: results from Panel 2

Recurrence-free survival					
Model	**# covars**	**-LL**	**cvl**	**P(unadj)**	**P(Bonf)**
Linear model					
Null	0	56.78			
$\underset{j\in 1:m_i}{\text{mean}}\left[\hat{F}(\text{her-2/neu})\right]$	1	49.42	66.47	0.027	
$\underset{j\in 1:m_i}{\text{mean}}\left[\hat{F}(\text{EGFR})\right]$	1	49.71	67.01	0.038	
$\underset{j\in 1:m_i}{\text{mean}}\left[\hat{F}(\text{VEGF})\right]$	1	51.74	68.20	0.626	
$\underset{j\in 1:m_i}{\text{mean}}\left[\hat{F}(\text{DNA})\right]$	1	46.33	64.09	0.001	
$\underset{j\in 1:m_i}{\text{mean}}\left[\hat{F}(\text{her-2/neu})\right] + \underset{j\in 1:m_i}{\text{mean}}\left[\hat{F}(\text{DNA})\right]$	3	45.99	91.72	0.003	
$\underset{j\in 1:m_i}{\text{mean}}\left[\hat{F}(\text{her-2/neu})\right] * \underset{j\in 1:m_i}{\text{mean}}\left[\hat{F}(\text{DNA})\right]$	3	44.74	113.53	0.003	
$\underset{j\in 1:m_i}{\text{mean}}\left[\hat{F}(\text{EGFR})\right] * \underset{j\in 1:m_i}{\text{mean}}\left[\hat{F}(\text{DNA})\right]$	3	44.65	128.20	0.002	
$\underset{j\in 1:m_i}{\text{mean}}\left[\hat{F}(\text{VEGF})\right] * \underset{j\in 1:m_i}{\text{mean}}\left[\hat{F}(\text{DNA})\right]$	3	45.26	75.92	0.004	
Main effects	4	44.21	98.75	0.004	
All effects	15	30.81	93.06	<0.001	
Weakest-link between two markers on Panel 2, with adj. p-value < 0.01					
$\underset{her-2/neu,DNA}{\text{wl}^{(++)}}\left\{\underset{j\in 1:m_i}{\text{mean}}\left[\hat{F}(\text{her-2/neu}), \hat{F}(\text{DNA})\right]\right\}$	1	46.60	62.48	0.001	0.005
$\underset{j\in 1:m_i}{\text{mean}}\left[\underset{her-2/neu,DNA}{\text{wl}^{(++)}}\left\{\hat{F}(\text{her-2/neu}), \hat{F}(\text{DNA})\right\}\right]$	1	46.28	62.09	0.001	0.003
$\underset{j\in 1:m_i}{\text{mean}}\left[\underset{EGFR,DNA}{\text{wl}^{(++)}}\left\{\hat{F}(\text{EGFR}), \hat{F}(\text{DNA})\right\}\right]$	1	46.60	62.50	0.001	0.005
$\underset{j\in 1:m_i}{\text{mean}}\left[\underset{VEGF,DNA}{\text{wl}^{(+-)}}\left\{\hat{F}(\text{VEGF}), \hat{F}(\text{DNA})\right\}\right]$	1	46.74	63.84	0.001	0.005
Weakest-link (t_i)**between four markers on Panel 2:**					
$\underset{k\in Panel2}{\text{wl}^{(++++)}}\left\{\underset{j\in 1:m_i}{\text{mean}}\left[\hat{F}_k(x_{j(i)k})\right]\right\}$	1	47.76	63.64	0.004	0.067
$\underset{j\in 1:m_i}{\text{mean}}\left[\underset{k\in Panel2}{\text{wl}^{(++++)}}\left\{\hat{F}_k(x_{j(i)k})\right\}\right]$	1	46.85	62.59	0.002	0.025
Best fitting model using logic-regression					
$I(\geq \text{median}\left\{\underset{j\in 1:m_i}{\text{mean}}\left[\hat{F}(\text{her-2/neu})\right]\right\})$ and $I(\geq \text{median}\left\{\underset{j\in 1:m_i}{\text{mean}}\left[\hat{F}(\text{DNA})\right]\right\})$	1	45.99	62.52	0.001	

Linear and weakest-link methods for modeling recurrence-free survival, Panel 2 (n=53)
For pairwise linear models, $*$ symbol denotes model with main effects and interaction
$\text{wl}^{(D)}$: weakest-link with directional vector (D)
-LL: -partial log-likelihood; CVL: cross-validated partial log-likelihood
P(unadj): Unadjusted p-values from Cox proportional hazards regression
P(Bonf): For weakest-link models, Bonferroni-adjusted p-values for number of directions

in Tables 18 and 19. While the model with binary covariates from logic regression provided as good a fit, the cross-validated *cvl* was slightly higher than the best-fitting weakest-link models. This may have been due to redefining the median of the observations for each different set of training data.

The usefulness of the weakest-link model in characterizing joint association between the most highly associated pair of her-2/neu and DNA with recurrence is seen in the scatterplot in Figure 13. The censored observations located at the lower right of the scatterplot cause apparent overfitting from the additive linear model. As a result, contours from the fitted additive linear interaction model do not behave as expected: when DNA content is high, recurrence risk increases with her-2/neu, however when DNA content is low, recurrence decreases with her-2/neu.

7.1.2 Weakest-link for identifying high-risk groups of patients

As seen in Tables 18 and 19, the continuous values obtained from the weakest-link algorithms were predictive of recurrence-free survival for markers in Panel 2, but not those in Panels 1 and 3. However, it is possible to identify low-risk and high-risk patients from some of the weakest-link models by dichotomizing derived covariates that were not significant, from Tables 18 and 19.

To this end, we dichotomized selected covariates t_i from Tables 18 and 19 through unsupervised k-means clustering, with 2 clusters for dichotomization. We also dichotomized the within-cell weakest-link between her-2/neu and DNA on Panel 2, which had the lowest cross-validated log-likelihood in Table 19. The other dichotomized covariates were the within-cell weakest-link between all 4 covariates in the same panel, for each panel 1 to 3.

We summarize these four dichotomized covariates as follows:

- $t_{(4,Panel1)} = \underset{j\in 1:m_i}{\text{mean}} \left[\underset{k\in Panel1}{\text{wl}^{(+--+)}} \left\{ \hat{F}_k(x_{j(i)k}) \right\} \right]$
- $t_{(2,Panel2)} = \underset{j\in 1:m_i}{\text{mean}} \left[\underset{her-2/neu,DNA}{\text{wl}^{(++)}} \left\{ \hat{F}(\text{her-2/neu}), \hat{F}(\text{DNA}) \right\} \right]$
- $t_{(4,Panel2)} = \underset{j\in 1:m_i}{\text{mean}} \left[\underset{k\in Panel2}{\text{wl}^{(++++)}} \left\{ \hat{F}_k(x_{j(i)k}) \right\} \right]$
- $t_{(4,Panel3)} = \underset{j\in 1:m_i}{\text{mean}} \left[\underset{k\in Panel3}{\text{wl}^{(+-++)}} \left\{ \hat{F}_k(x_{j(i)k}) \right\} \right]$

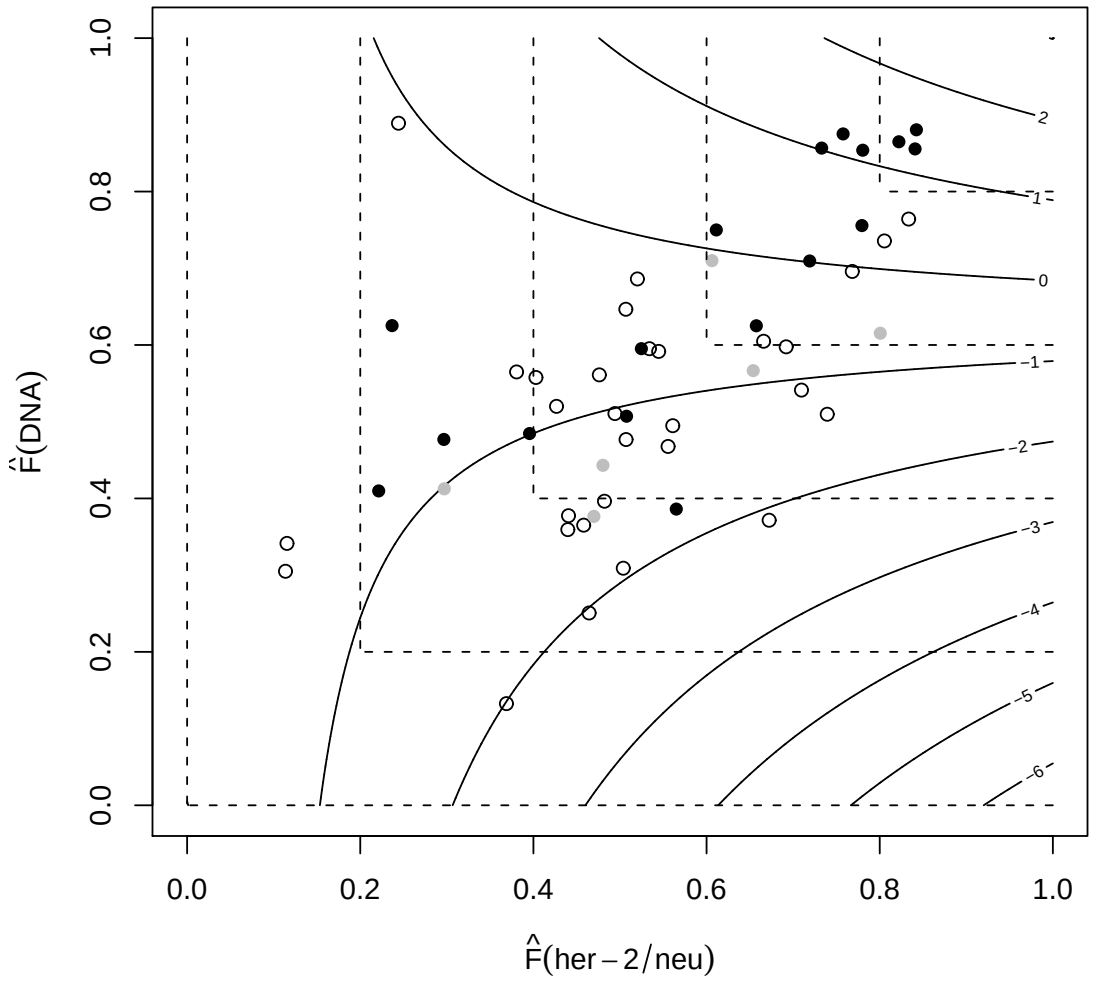

Figure 13: Scatterplot between her-2/neu and DNA

Scatterplot between empirical distribution function $\hat{F}$ of her-2/neu and DNA, from Panel 2. Patients who recurred or died are in black, those who were censored at more than 10 days in white and censored at less than 10 days in gray. Equivalent values of the quantile-stitching weakest-link covariate $\underset{her-2/neu,DNA}{\mathrm{wl}^{(++)}}\left\{\underset{j\in 1:m_i}{\mathrm{mean}}\left[\hat{F}(\text{her-2/neu})\right], \underset{j\in 1:m_i}{\mathrm{mean}}\left[\hat{F}(\text{DNA})\right]\right\}$ are on the same dotted line. The points on the contour have equal predicted values according to the main and interaction effects Cox regression model her-2/neu+DNA+her-2/neu*DNA.

To cluster observations from one covariate (t_i) into 2 clusters, k-means finds a cutoff that minimizes the within-cluster sum of squares, and categorizes the observations according to this cutoff. As the clustering was unsupervised, the cutpoints for defining the groups were not driven by the outcomes. Kaplan-Meier survival was estimated in each of these dichotomized groups, with the log-rank test for intergroup comparisons (Table 20), with both unadjusted and Bonferroni-adjusted p-values.

Recurrence-free survival (RFS) significantly differed in the 2 groups dichotomized from the weakest-link between all 4 biomarkers in both Panel 1 and Panel 2. In particular, the 7 highest observations of derived covariates for Panel 1 appeared to form a clear subgroup that was highly associated with lower recurrence-free survival.

However, RFS was not different between groups dichotomized from the weakest-link between her-2/neu and DNA from Panel 2, $\underset{j\in 1:m_i}{\text{mean}}\left[\underset{her-2/neu,DNA}{\text{wl}^{(++)}}\left\{\hat{F}(\text{her-2/neu}), \hat{F}(\text{DNA})\right\}\right]$. This occurred even though this continuous covariate was highly associated with recurrence-free survival (p=0.003 from Table 19). Weakest-link did not detect any relation between recurrence-free survival and the markers on Panel 3.

To see if recurrence-free survival comparisons would have been different for other group cutpoints, we plotted the p-values from a log-rank test against all the possible cutpoints (Figure 14) of these derived covariates. The patients were dichotomized as follows. For each possible cutpoint t_i, a patient i^* is in the high-risk group if $t_{i^*} > t_i$. This procedure is repeated for all possible cutpoints $t_1, \ldots, t_{56}$. The derived covariates t_i of each patient $i = 1, \ldots, n$ are plotted along the x-axis. Plotted above these points, on the y-axis, are the corresponding adjusted p-values when each of these t_i is the cutpoint for a log-rank test.

For both covariates obtained from Panel 2, $t_{(2,Panel2)}$ and $t_{(4,Panel2)}$, there were several cutpoints that could have successfully identified low and high-risk groups for recurrence-free survival. However the k-means algorithm did not produce the ideal cutpoints (in terms of lowest p-value), especially for the Panel 2 her-2/neu and DNA weakest-link derived covariate $t_{(2,Panel2)}$ (left). There appeared to be no clear clustering in these derived covariates; k-means tends to produce clusters of similar size and may not have been an ideal dichotomization procedure.

Table 20: Recurrence-free survival in subgroups by unsupervised clustering

Recurrence-free survival (months)					
Derived covariate,t_i (k-means clustering; 2 clusters)	**Risk group**	**Events/ At risk**	**Med RFS (95%CI)**	**p (unadj)**	**p (Bonf)**
$\underset{j\in 1:m_i}{\text{mean}}\left[\underset{k\in Panel1}{\text{wl}^{(+--+)}}\left\{\hat{F}_k(x_{j(i)k})\right\}\right]$	low	12/48	62 (48-NA)	1×10^{-5}	7.2×10^{-4}
	high	6/7	15 (6-NA)		
$\underset{j\in 1:m_i}{\text{mean}}\left[\underset{her2neu,DNA}{\text{wl}^{(++)}}\left\{\hat{F}(\text{her2neu}),\hat{F}(\text{DNA})\right\}\right]$	low	7/29	62 (49-NA)	0.029	0.117
	high	10/24	47 (15-NA)		
$\underset{j\in 1:m_i}{\text{mean}}\left[\underset{k\in Panel2}{\text{wl}^{(++++)}}\left\{\hat{F}_k(x_{j(i)k})\right\}\right]$	low	7/33	62 (49-NA)	3×10^{-5}	4.1×10^{-4}
	high	10/20	15 (12-NA)		
$\underset{j\in 1:m_i}{\text{mean}}\left[\underset{k\in Panel3}{\text{wl}^{(+-++)}}\left\{\hat{F}_k(x_{j(i)k})\right\}\right]$	low	6/22	62 (49-NA)	0.0465	0.744
	high	11/28	18 (15-NA)		

Median Kaplan-Meier recurrence-free survival in subgroups detected by unsupervised clustering of weakest-link derived covariates.
Comparisons by log-rank test.
p-value (log-rank): from log-rank test between groups
p-value (Bonf): with Bonferroni adjustment for weakest-link directionality
High and low-risk groups obtained through unsupervised k-means clustering (2 clusters).
Biomarkers used in weakest-link (x_k)
Panel 1 ($k = 1$ to 4): her-2/neu, p53, Ras, DNA
Panel 2 ($k = 5$ to 8): her-2/neu, EGFR, VEGF, DNA
Panel 3 ($k = 9$ to 12): p16, Rb, Cycline E, DNA

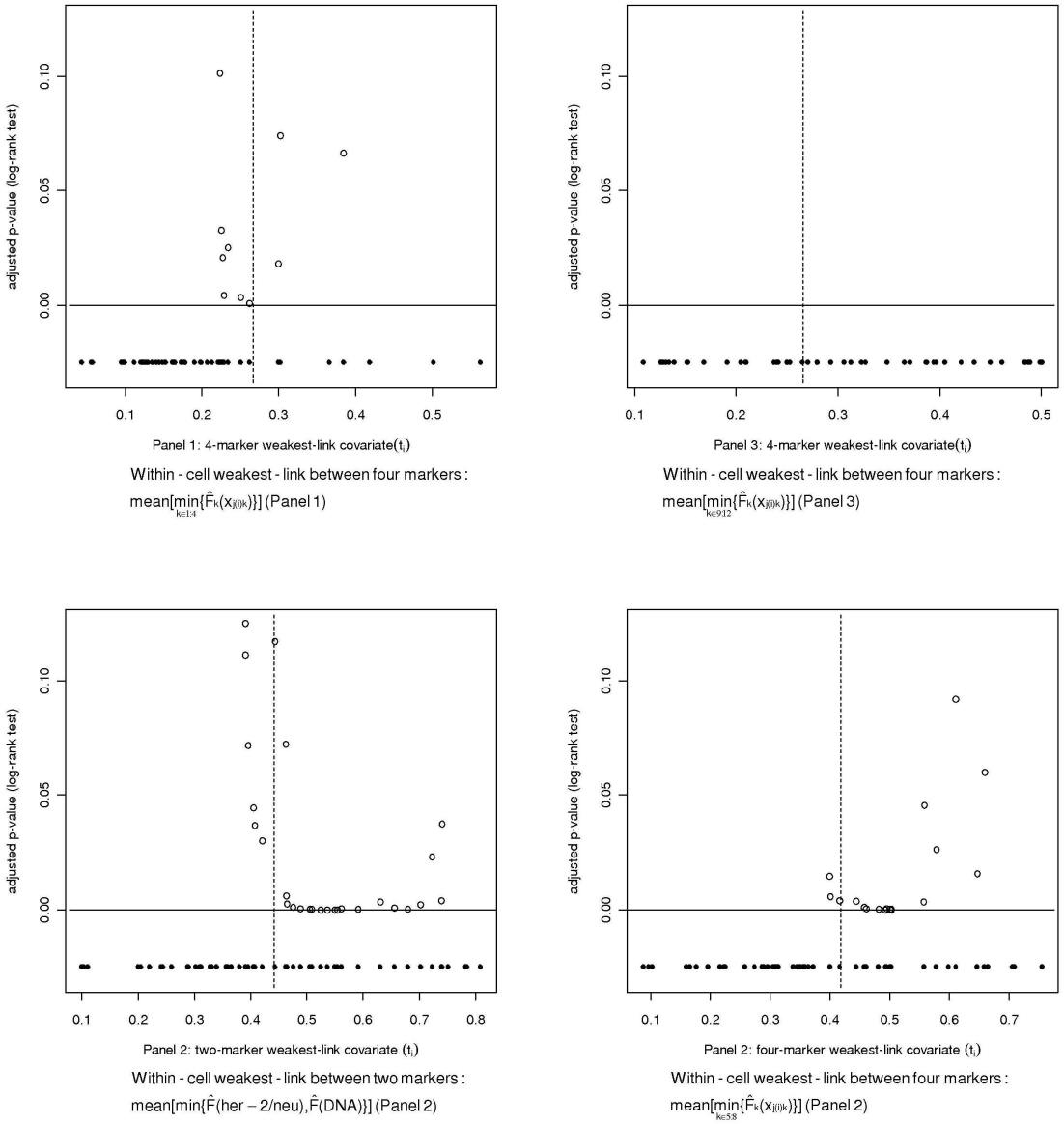

Figure 14: Plots of derived covariates from quantile stitching weakest-link

Plots of the derived covariate t_i from quantile stitching for each patient along the x-axis, for each patient $i = 1, \ldots, n$. Plotted above these points, on the y-axis, are the corresponding Bonferroni-adjusted p-values when each of these t_i is the cutpoint for a log-rank test of recurrence-free survival. Upper left: within-cell weakest-link between four covariates from Panel 1, right: within-cell weakest-link between 4 covariates from Panel 3. Bottom left: within-cell weakest-link between two covariates from Panel 2 (her-2/neu, DNA), right: within-cell weakest-link between 4 covariates from Panel 2. Vertical lines correspond to the cutpoint identified by k-means clustering.

We used the method of maximally selected statistics (Hothorn and Zeileis, 2008), a supervised procedure, to infer an optimal cutpoint for dichotomization relative to the outcomes. The package *maxstat* in *R* was used to identify two groups of observations relative to RFS, according to a log-rank statistic that is maximally selected in terms of an optimal cutpoint. The procedure also approximates the null distribution of the maximally selected statistic; as a result, any p-values take into account the multiple comparisons made from considering different possible cutpoints.

We assessed the ability of the derived covariates in forming categorical risk subgroups by comparing the unsupervised k-means clusters to those from the maximally selected tests, which are supervised. The maximally selected method estimated Kaplan-Meier recurrence-free survival, in optimally chosen subgroups (Table 21), with corresponding log-rank tests adjusting for multiple cutpoints.

We compared the p-values from the maximally selected test to those from the unsupervised k-means clustering of the same derived covariates (Table 20). The unsupervised k-means clustering of weakest-link between 4 covariates from Panel 1 ($t_{(4,Panel1)}$) found the same optimal cutpoint as that from the supervised maximally selected test. Considering the k-means procedure was unsupervised, the corresponding p-value did not require adjustment and stayed below the $p < 0.05$ threshold. However, the difference in RFS between these groups from $t_{(4,Panel1)}$ was not significant after the maximally selected procedure adjusted for multiple cutpoints. On the other hand, the supervised maximally selected test was required to adequately identify dichotomous subgroups with respect to RFS of the two-marker weakest-link between her-2/neu and DNA from Panel 2, $t_{(2,Panel2)}$.

7.1.3 Effects of level of weakest-link assessment in cell-based data

Derived covariates, whether taken at the patient level or cell level, were highly correlated with each other. This pattern is illustrated in two different covariates obtained from the pairwise weakest-link between her-2/neu and DNA, with the weakest-link assessed:

- at the patient level $t_{i,wl(patient)} = \underset{her-2/neu,DNA}{\text{wl}^{(++)}} \left\{ \underset{j\in 1:m_i}{\text{mean}} \left[\hat{F}(\text{her-2/neu}), \hat{F}(\text{DNA}) \right] \right\}$.
- at the cell level $t_{i,wl(cell)} = \underset{j\in 1:m_i}{\text{mean}} \left[\underset{her-2/neu,DNA}{\text{wl}^{(++)}} \left\{ \hat{F}(\text{her-2/neu}), \hat{F}(\text{DNA}) \right\} \right]$.

Table 21: Kaplan-Meier median survival in subgroups detected by maximally selected weakest-link derived covariates

Recurrence-free survival (months)					
Cutpoint selected from Derived covariate, t_i	**Risk group**	**Events/ At risk**	**Med RFS (95%CI)**	**p (unadj)**	**p (max)**
$\underset{j\in 1:m_i}{\text{mean}}\left[\underset{k\in Panel1}{\text{wl}^{(+--+)}}\left\{\hat{F}_k(x_{j(i)k})\right\}\right]$	low	12/48	62 (48-NA)	7.2×10^{-4}	0.174
	high	6/7	15 (6-NA)		
$\underset{j\in 1:m_i}{\text{mean}}\left[\underset{her2neu,DNA}{\text{wl}^{(++)}}\left\{\hat{F}(\text{her2neu}),\hat{F}(\text{DNA})\right\}\right]$	low	7/38	62 (49-NA)	4.0×10^{-5}	0.003
	high	10/15	15 (11-NA)		
$\underset{j\in 1:m_i}{\text{mean}}\left[\underset{k\in Panel2}{\text{wl}^{(++++)}}\left\{\hat{F}_k(x_{j(i)k})\right\}\right]$	low	8/40	62 (49-NA)	9.0×10^{-5}	0.021
	high	9/13	15 (7-NA)		
$\underset{j\in 1:m_i}{\text{mean}}\left[\underset{k\in Panel3}{\text{wl}^{(+-++)}}\left\{\hat{F}_k(x_{j(i)k})\right\}\right]$	low	5/20	62 (49-Inf)	0.082	0.822
	high	12/30	47 (15-Inf)		

Survival was estimated by Kaplan-Meier method, comparisons by log-rank test.
p-value (log-rank): from log-rank test between groups, with Bonferroni adjustment for weakest-link directionality
Maximally selected statistics denote optimal cutpoints for dividing low-risk from high-risk patients using weakest-link derived covariates, with the weakest-link taken at the cell level.
p-value (max log-rank): from maximally selected log-rank test between groups, accounting for multiple cutpoints, with Bonferroni adjustment for weakest-link directionality
Panel 1: her-2/neu, p53, Ras, DNA
Panel 2: her-2/neu, EGFR, VEGF, DNA
Panel 3: p16, Rb, Cycline E, DNA

A scatterplot of these two covariates is in Figure 15 (top).

A *smaller* within-patient intracellular correlation between her-2/neu and DNA was significantly associated ($\rho = -0.494, p < 0.001$) with a larger difference between covariates obtained from weakest-link taken at the patient level compared to weakest-link was taken at the cell level ($t_{i,wl(cell)} - t_{i,wl(patient)}$). A scatterplot illustrates this inverse relationship in Figure 15 (bottom).

The two covariates $t_{i,wl(patient)}$ and $t_{i,wl(cell)}$ were highly correlated ($\rho = 0.990$). Both covariates were also highly associated with recurrence-free survival (Table 19), with Bonferroni adjusted p-values of 0.005 (within-patient) and 0.003 (within-cell).

7.2 DISCUSSION OF ANALYSIS OF CYTOMETRY DATA

The described analysis of the cytometry data set dealt with several issues from fitting a set with hierarchical structure using weakest-link models.

Most cancer phenotypes are driven by individual intact cells due to their heterogeneous nature. However, high-throughput technologies typically require destroying the identity of individual cells, which can obscure the interactions among the covariates. The sheer quantity of covariates cannot make up for the loss of information in the sample processing step. We analyzed a small set of previously selected markers, measured on intact cells by fluorescence cytometry. We compared within-cell weakest link models with weakest link models that amalgamate data across cells before examining interactions. As expected, the weakest-link models performed slightly better in the former, which preserves the joint information of several markers within individual cells. The somewhat improved performance of this model may reflect the importance of intact individual cells in understanding the behavior of the cancer.

The comparison with linear models was clear. While additive models provided fits that were closer to the data, the weakest link models were far more optimal as judged by the cross-validated likelihood criterion. Consider the contour plot for the expected outcome as a function of the two covariates (Figure 13). As one covariate changes, the relationship

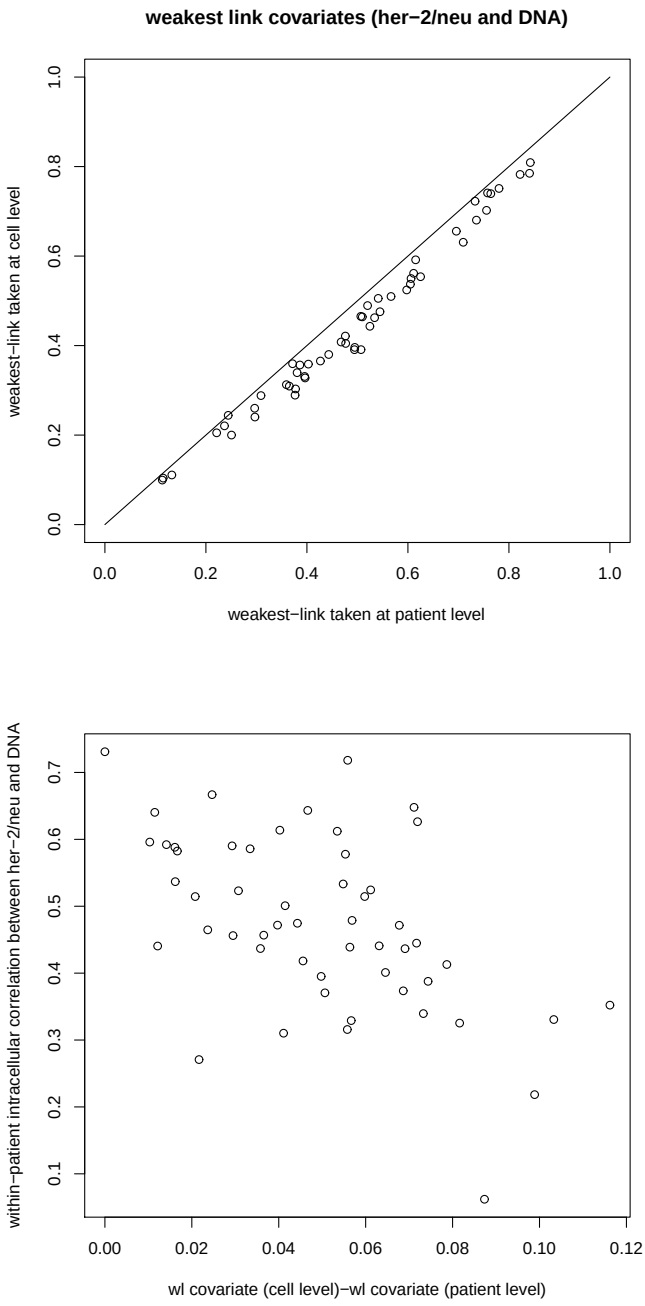

Figure 15: Weakest-link covariates when taken at the patient level and cell level

Top: Comparison of patient covariates when weakest-link is taken at patient level vs cell level. Bottom: Differences between corresponding derived covariates (taken within cell - taken within patient) vs within-patient intracellular correlation between Her-2/neu and DNA. Each patient is marked by a different point.

between the other covariate and the outcome reverses at a point. The interpretation of the interaction term is incorrect if this point is within the range of data. There is some evident overfitting resulting from a group of 10 or so censored points in the bottom right-hand corner of the plot. In this setting, the significant problem of overfitting with linear models is greatly mitigated by the use of weakest link models.

Another issue in most procedures designed for combining several attributes is that they tend to require the features to be categorical. In this limited comparison of weakest link models with logic regression, weakest link models performed equivalently or somewhat better. As well, dichotomizing the data did not improve the classification value of the models unless there is a clear underlying clustering of data.

These results encourage the use of weakest link models as a valuable tool which should be considered when the joint effects of multiple biomarkers are of interest.

8.0 R PACKAGE WEAKESTLINK

An R package *weakestLink* is in development that includes most of the functions used within this dissertation. Some of the features of this package are:

The package includes functions for searching combinations using the described stitched weakest-link models:

- Quantile-stitched
- Probit-stitched

Similar searches also search for combinations using additive linear models of the following types:

- Main effects only
- Main and interaction effects

Search types include:

- Exhaustive searches for each possible combination of covariates
- Filtered searches within a subset of covariates
- Greedy search for locally optimal pairs
- Simulated annealing searches

Model types implemented are:

- Linear regression for continuous outcomes
- Logistic regression for binary outcomes
- Cox proportional hazards regression for survival outcomes

If desired, raw data can be standardized to the following types:

- Standard normal distribution

- Uniform distribution on the unit interval by the empirical cumulative distribution function

Other features include:

- Searches for combinations of more than 2 covariates
- Cross-validation procedures as described within this dissertation, for each of the model types

A description of the functions, arguments, and output is available on the online help files, along with simple examples demonstrating the important functions.

Below is an example of a typical running of the function *wlinkSearch* to find covariate pairs associated with lymph node status, through an exhaustive search on a filtered subset of covariates (Table 4).

```
# covariate indices with p<0.001 marginal association with LN+ by likelihood ratio test
subset.ind <- numeric()

# univariable likelihood ratio tests on all covariates
for (k in 1:7129){
model.temp <- glm(LN.pos~duke.norm.all[,k],family="binomial")
lr.stat <- model.temp$null.deviance-model.temp$deviance
p.unadj <- 1-pchisq(lr.stat,1)
if (p.unadj < 0.001) subset.ind <- append(subset.ind,k)
}

# use exhaustive search on filtered subset of covariates
wlinkSearch(duke.norm.all, subset.ind, n.in.comb=2, test.type=2, outcome=LN.pos,
   matching.type=2)
$sig.results
            best.wl best.score p.value.unadj p.value.bonf      comb wl.dir    p.value
6088   V118V441..00   28.41651  3.294892e-10 1.317957e-09  118.4413     00 0.03348633
6791   V132V417..00   28.77631  3.961591e-10 1.584636e-09  132.4172     00 0.04026204
6803   V132V441..00   21.92835  1.194866e-11 4.779466e-11  132.4413     00 0.00121435
26328 V126.V295..00   26.80089  1.441102e-10 5.764407e-10 1264.2955     00 0.01464606
26393 V126.V441..00   28.78582  3.980921e-10 1.592368e-09 1264.4413     00 0.04045850

$best.results
          best.wl best.score p.value.unadj p.value.bonf     comb wl.dir    p.value
6803 V132V441..00   21.92835  1.194866e-11 4.779466e-11  132.4413     00 0.00121435
```

BIBLIOGRAPHY

Aarts, E., Korst, J., and Michiels, W. (2005). *Simulated Annealing*, pages 187–210. New York: Springer.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716–23.

Belisle, C. (1992). Convergence theorems for a class of simulated annealing algorithms on Rd. *J. of Applied Probability* **29**, 885–92.

Boedigheimer, M. J. and Ferbas, J. (2008). Mixture modeling approach to flow cytometry data. *Cytometry A* **73**, 421–9.

Breiman, L. (2001). Random forests. In *Machine Learning*, pages 5–32.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Monterey, CA: Wadsworth and Brooks.

Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrika* **30**, 89–99.

Culverhouse, R., Klein, T., and Shannon, W. (2004). Detecting epistatic interactions contributing to quantitative traits. *Genet Epidemiol* **27**, 141–52.

Culverhouse, R., Suarez, B. K., Lin, J., and Reich, T. (2002). A perspective on epistasis: limits of models displaying no main effect. *Am J Hum Genet* **70**, 461–71.

Darzynkiewicz, Z., Bedner, E., Li, X., Gorczyca, W., and Melamed, M. R. (1999). Laser-scanning cytometry: A new instrumentation with many applications. *Exp Cell Res* **249**, 1–12.

Dettling, M. and Bühlmann, P. (2003). Boosting for tumor classification with gene expression data. *Bioinformatics* **19**, 1061–069.

Donnenberg, A. D. and Donnenberg, V. S. (2007). Rare-event analysis in flow cytometry. *Clin Lab Med* **27**, 627–52, viii.

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association* **72**, 557–65.

Emlet, D. R., Brown, K. A., Kociban, D. L., Pollice, A. A., Smith, C. A., Ong, B. B. L., and Shackney, S. E. (2007). Response to trastuzumab, erlotinib, and bevacizumab, alone and in combination, is correlated with the level of human epidermal growth factor receptor-2 expression in human breast cancer cell lines. *Mol Cancer Ther* **6**, 2664–74.

Fukuyama, Y., Mitsudomi, T., Sugio, K., Ishida, T., Akazawa, K., and Sugimachi, K. (1997). K-ras and p53 mutations are an independent unfavourable prognostic indicator in patients with non-small-cell lung cancer. *Br J Cancer* **75**, 1125–30.

Gaffney, D. K., Haslam, D., Tsodikov, A., Hammond, E., Seaman, J., Holden, J., Lee, R. J., Zempolich, K., and Dodson, M. (2003). Epidermal growth factor receptor (EGFR) and vascular endothelial growth factor (VEGF) negatively affect overall survival in carcinoma of the cervix treated with radiotherapy. *Int J Radiat Oncol Biol Phys* **56**, 922–8.

Gentle, J. E. (2009). *Computational Statistics*. New York: Springer.

Hahn, L., Ritchie, M. D., and Moore, J. H. (2003). Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics* **19**, 376–82.

Hothorn, T. and Zeileis, A. (2008). Generalized maximally selected statistics. *Biometrics* **64**, 1263–9.

Janes, H., Pepe, M., Kooperberg, C., and Newcomb, P. (2005). Identifying target populations for screening or not screening using logic regression. *Statistics in Medicine* **24**, 1321–38.

Kern, W., Voskova, D., Schoch, C., Schnittger, S., Hiddemann, W., and Haferlach, T. (2004). Prognostic impact of early response to induction therapy as assessed by multiparameter flow cytometry in acute myeloid leukemia. *Haematologica* **89**, 528–40.

Klein, J. and Moeschberger, M. (2005). *Survival Analysis: Techniques for Censored and Truncated Data*. New York, NY: Springer–Verlag.

Konishi, T., Huang, C. L., Adachi, M., Taki, T., Inufusa, H., Kodama, K., Kohno, N., and Miyake, M. (2000). The K-ras gene regulates vascular endothelial growth factor gene expression in non-small cell lung cancers. *Int J Oncol* **16**, 501–11.

Kooperberg, C., Ruczinski, I., LeBlanc, M. L., and Hsu, L. (2001). Sequence analysis using logic regression. *Genet Epidemiol* **21 Suppl 1**, S626–31.

Lo, K., Brinkman, R. R., and Gottardo, R. (2008). Automated gating of flow cytometry data via robust model-based clustering. *Cytometry A* **73**, 321–32.

Marchini, J., Donnelly, P., and Cardon, L. R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics* **37**, 413–417.

Muchardt, C., Reyes, J., Bourachot, B., Leguoy, E., and Yaniv, M. (1996). The hbrm and brg-1 proteins, components of the human snf/swi complex, are phosphorylated and excluded from the condensed chromosomes during mitosis. *EMBO J* **15**, 3394–3402.

Muller-Tidow, C., Metzger, R., Kugler, K., Diederichs, S., Idos, G., Thomas, M., Dockhorn-Dworniczak, B., Schneider, P. M., Koeffler, H. P., Berdel, W. E., and Serve, H. (2001). Cyclin E is the only cyclin-dependent kinase 2-associated cyclin that predicts metastasis and survival in early stage non-small cell lung cancer. *Cancer Res* **61**, 647–53.

Niklinski, J., Niklinska, W., Laudanski, J., Chyczewska, E., and Chyczewski, L. (2001). Prognostic molecular markers in non-small cell lung cancer. *Lung Cancer* **34 Suppl 2**, S53–8.

Nowakowski, G. S., Witzig, T. E., Dingli, D., Tracz, M. J., Gertza, M. A., Lacy, M. Q., Lust, J. A., Dispenzieri, A., Greipp, P. R., Kyle, R. A., and Rajkumar, S. V. (2005). Circulating plasma cells detected by flow cytometry as a predictor of survival in 302 patients with newly diagnosed multiple myeloma. *Blood* **106**, 2276–79.

Pardo, A., Gibson, K., Cisneros, J., Richards, T. J., Yang, Y., Becerril, C., Yousem, S., Herrera, I., Ruiz, V., Selman, M., and Kaminski, N. (2005). Up-regulation and profibrotic role of osteopontin in human idiopathic pulmonary fibrosis. *PLoS Med* **2**, e251.

Park, M. Y. and Hastie, T. (2008). Penalized logistic regression for detecting gene interactions. *Biostatistics* **9**, 30–50.

Perez-Persona, E., Vidriales, M.-B., Mateo, G., Garcia-Sanz, R., Mateos, M.-V., de Coca, A. G., Galende, J., Martin-Nunez, G., Alonso, J. M., de Las Heras, N., Hernandez, J. M., Martin, A., Lopez-Berges, C., Orfao, A., and Miguel, J. F. S. (2007). New criteria to identify risk of progression in monoclonal gammopathy of uncertain significance and smoldering multiple myeloma based on multiparameter flow cytometry analysis of bone marrow plasma cells. *Blood* **110**, 2586–92.

Ratei, R., Karawajew, L., Lacombe, F., Jagoda, K., Poeta, G. D., Kraan, J., Santiago, M. D., Kappelmayer, J., Bjorklund, E., Ludwig, W.-D., Gratama, J. W., and Orfao, A. (2007). Discriminant function analysis as decision support system for the diagnosis of acute leukemia with a minimal four color screening panel and multiparameter flow cytometry immunophenotyping. *Leukemia* **21**, 1204–11.

Richards, T. (2002). *Weakest Link Models*. Ph.D. thesis, University of Pittsburgh.

Ritchie, M. D., Hahn, L. W., Roodi, N., Bailey, L. R., Dupont, W. D., Parl, F. F., and Moore, J. H. (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet* **69**, 138–47.

Ruczinski, I. (2000). *Logic regression and statistical issues related to the protein folding problem*. Ph.D. thesis, University of Washington.

Ruczinski, I., Kooperberg, C., and Leblanc, M. (2003). Logic regression. *Journal of Computational and Graphical Statistics* **12**, 475–511.

Schwender, H. and Ickstadt, K. (2007). Identification of SNP interactions using logic regression. *Biostatistics* **9**, 187–98.

Shackney, S. E., Smith, C. A., Pollice, A., Brown, K., Day, R., Julian, T., and Silverman, J. F. (2004). Intracellular patterns of Her-2/neu, ras, and ploidy abnormalities in primary human breast cancers predict postoperative clinical disease-free survival. *Clin Cancer Res* **10**, 3042–52.

Shannon, W. (1995). *Conjunctive Split Models*. Ph.D. thesis, University of Pittsburgh.

Shek, L. L. and Godolphin, W. (1988). Model for breast cancer survival: relative prognostic roles of axillary nodal status, TNM stage, estrogen receptor concentration, and tumor necrosis. *Cancer Res* **58**, 5565–9.

Toedling, J., Rhein, P., Ratei, R., Karawajew, L., and Spang, R. (2006). Automated in-silico detection of cell populations in flow cytometry readouts and its application to leukemia disease monitoring. *BMC Bioinformatics* **7**, 282.

Verweij, P. J. and Houwelingen, H. C. V. (1993). Cross-validation in survival analysis. *Statistics in Medicine* **12**, 2305–14.

Wang, H. and Huang, S. (2007). Mixture-model classification in DNA content analysis. *Cytometry A* **71A**, 716–23.

West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson, J. A. J., Marks, J. R., and Nevins, J. R. (2001). Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci U S A* **98**, 11462–7. URL http://data.genome.duke.edu/west.php

Yang, C., He, Z., Wan, X., Yang, Q., Xue, H., and Yu, W. (2009). SNPHarvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies. *Bioinformatics* **25**, 504–511.

Zhang, Y. and Liu, J. S. (2007). Bayesian inference of epistatic interactions in case-control studies. *Nat Genet* **39**, 1167–73.

Zheng, T., Wang, H., and Lo, S.-H. (2006). Backward genotype-trait association (BGTA)-based dissection of complex traits in case-control designs. *Hum Hered* **62**, 196–212.

Zhu, J. and Hastie, T. (2004). Classification of gene microarrays by penalized logistic regression. *Biostatistics* **5**, 427–443.