Optimal Decision Rule for Combining Multiple Biomarkers into Tree-based Classifier and its Evaluation

by

Yuxin Zhu

A dissertation submitted to The Johns Hopkins University
in conformity with the requirements for the degree of
Doctor of Philosophy

Baltimore, Maryland
May, 2018

Abstract

In biomedical practices, multiple biomarkers are often combined using a classification rule of the form of some tree structure to make diagnostic decisions. The classification structure and cutoff point at each node of a tree are commonly chosen ad-hoc based on experience of decision makers. There is a lack of analytical approaches that lead to optimal prediction performance, and that guide the choice of optimal cutoff points of a pre-specified classification tree. In this dissertation, we propose to search for and estimate the optimal decision rule through an approach of rank correlation maximization. The proposed method is flexible and computationally feasible using data with reasonably large sample sizes when there are many biomarkers available for classification or prediction. Using this method, for a pre-specified tree-structured classification rule, we are able to guide the choice of optimal cutoff at tree nodes, as well as to estimate optimal prediction performance of multiple biomarkers combined.

In this dissertation, we also propose a semi-marginal and semi-parametric regression model for gap times between successive recurrent events in the presence of time-dependent covariates. Recurrent event data is commonly encountered in longitudinal follow-up studies, when each subject experiences multiple events under observation until loss to follow-up, dropout or end of study occurs. There exists a rich literature of models and methods that focus on time-to-event data in a recurrent event setting, but for applications where time-between-events (also referred to as gap times) is of scientific interest or where there is a strong cyclical pattern, limited techniques were developed, especially for regression with time-dependent covariates. We propose a semi-marginal regression model of a proportional hazard form on gap times such that no event history is included in the conditional statistics of regression except for the time relapse from baseline to last event occurrence. The proposed method is flexible in being semi-parametric, robust to various correlation structures of gap times within subject, and also allows time-dependent covariates to be included in the conditional statistics of regression.

Thesis Committee

Primary Readers

Mei-Cheng Wang (Primary Advisor)
Professor
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Lawrence H. Moulton (Chair)
Professor
Department of International Health
Johns Hopkins Bloomberg School of Public Health

Vadim Zipunnikov
Assistant Professor
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Anja Soldan
Assistant Professor
Neurology
Johns Hopkins School of Medicine

Alternate Readers

Xiaobin Wang
Professor
Department of Population, Family and Reproductive Health
Johns Hopkins Bloomberg School of Public Health

Elizabeth Colantuoni
Associate Scientist
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Acknowledgments

First and foremost, I would like to thank my advisor Dr. Mei-Cheng Wang for her guidance over the past five years. Dr. Wang has been always a great encouragement to me through difficulties, and a great inspiration, both intellectually as a researcher and more generally as a mentor. It has been a great honor to be one of her advisees, and my understanding and appreciation for the elegance of statistical methods, if any, has come from her. I am also grateful to her for organizing the SLAM group, where a student's opinion is always valued and encouraged.

I would like to thank Dr. Lawrence Moulton for being my thesis committee chair, and Dr. Vadim Zipunnikov and Dr. Anja Soldan for taking out precious time to read my thesis and serve the committee duties. I would also like to thank Dr. Xiaobin Wang and Dr. Elizabeth Colantuoni for being my alternates.

I am grateful to Dr. Karen Bandeen-Roche, Dr. Hongkai Ji, and Dr. Mei-Cheng Wang again, for being part of the summer school program at Nanjing University six years ago, during which I was introduced to the field of biostatisitcs, and which laid path for me to pursue studies in this field. I would also like to thank Dr. Fang Han, who has taught be a lot and will alway be an example to follow.

I want to thank all my PhD peers: Jack Fu, Detian Deng, Bing He, Elizabeth Sweeney, Claire Ruberman, Leslie Myint, and Yu Du, and all my instructors: Dr. Brian Caffo, Dr. Vadim Zipunnikov, Dr. Jeff Leek, Dr. Roger Peng, Dr. Mei-Cheng Wang, Dr. Frangaskis Constantine and Dr. Daniel Scharfstein for always being helpful and supportive. I would like to thank my collaborations Dr. Marilyn Albert, Dr. Anja Soldan, and Dr. Corinne Pettigrew in the BIOCARD study, for their support and understanding during our collaborations. I would also like to thank all students, staff and professors in the Department of Biostatistics for making here such a wonderful place to make mistakes, to learn, and to grow.

Last but not least, I want to thank my parents for their unconditional love and support throughout my entire life. Special thanks to my sister, who has taught me how to love and give.

Table of Contents

List of Tables

List of Figures

Chapter 1

Literature Review: Tree-Based Classification Methods, and Related

1.1 Tree Structure and Tree-Based Classifiers

A general tree structure is a graphical representation of the hierachical nature of some structure using nodes and branches, widely used in various fields including computer science and decision making. In a finite tree, each parent node connects to child nodes through branches. The node with no parent node is commonly referred to as the root node, while nodes with no child node connected are referred to as leaf nodes. One special structure commonly used for decision making, or put into statistical context, for classification and regression, is a binary tree in which each node has at most two child nodes, and each leaf node has an estimated outcome value attached, either binary for classification or continuous for regression. See Figure 1.1 for an example of a classification tree that classifies a subject to group 1 when both X1 and X2 are larger than some cutoff values, and to group 0 if otherwise. In

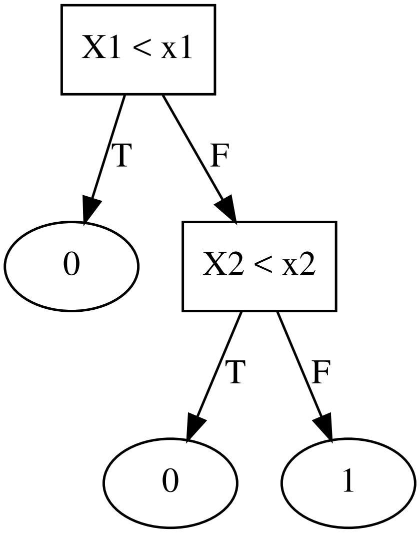

Figure 1.1: Example of a classification tree using two variables X1 and X2 for a binary outcome taking value 0 or 1. Leaf nodes are represented by circle, and non-leaf nodes by box with splitting condition at the node in the box. "T" on a branch indicates that condition is satisfied in the child node that follows, "F" on a branch indicates otherwise. For this particular tree, a subject is classified to have outcome 1 if both X1 and X2 are larger than some cutoff values, x1 and x2 respectively, and outcome 0 if otherwise.

general, tree structures used in statistical analysis are understood as a kind of data structure, where the root node represents the entire sample, and each branching indicates data bifurcating according to a binary outcome of whether some splitting covariate is larger than a certain cutoff value, into child nodes. Covariates used for classification are also referred to as markers.

Tree-based methods gained their popularity in statistical literature since the introduction of classification and regression tree (CART, Breiman et al., 1984). The classic CART procedure grows the tree using an algorithm greedy in the sense that it chooses splitting covariates and corresponding cutoff values by minimizing some loss function at each step, and therefore not necessarily minimizing the overall loss. The algorithm first overgrows the tree until there are only a few observations in each leaf node, and then prunes back by merging neighboring child nodes while minimizing increase in loss at each step. In face of the new challenges brought by big data era, traditional tree-based methods have been coupled with ensemble techniques in machine learning, giving rise to a great variety of powerful prediction methods including boosted trees (Friedman, 2002, Friedman, Hastie, and Tibshirani, 2001), random forest (Breiman, 1996) and rotation forest (Rodriguez, Kuncheva, and Alonso, 2006). The general idea of these ensembled tree methods is to infuse randomness in growing a number of trees by either using random samples or choosing splitting markers and cutoff values randomly, and then averaging. The reason for the popularity of these methods among the statistics community is that tree-based predictors are essentially non-parametric, and are thus flexible and robust to model misspecifications, an advantange especially when the training sample is large.

Not limited to the statistics community, tree-based methods are also well accepted by biomedical researchers and are commonly used in medical practices. Many diagnoses are made if a few tests come out positive; for example, diagnosis of HIV infection is made when both ELISA and Western blot tests detect HIV antibodies. This kind of decision making is similar to human thinking and thus interpretable. Meanwhile, using tree-based methods for decision making can save resources as not necessarily all tests need to be performed. In the example of HIV infection diagnosis, Western blot test is not needed if ELISA test comes out negative.

Despite the wide and successful applications of tree-based methods in statistics and biomedical practices, little progress has been made in understanding the theoretical properties of these methods. There is some work on consistency of random forests (Scornet, Biau, and Vert, 2015, Biau, Devroye, and Lugosi, 2008), but no statistical inferential results have thus far been presented. At the core of these challenges is the lack of understanding for a single tree.

1.2 Evaluating Prediction Performance of Tree-Based Classifiers

Methods to evaluate prediction performance of tree-based classifiers for binary outcomes have been developed from two distinct perspectives, depending on how trees are used. From one perspective, predetermined binary outcome labels are assigned to leaf nodes while cutoff values used for data splitting vary according to further requirements on desired sensitivity or specificity. Although rarely talked about, the simplest tree structure that uses one marker often falls under this category – it is given *a priori* that a subject would be classified as 1 (or 0) if his or her marker value is larger than some constant, while the cutoff value is calculated later according to further conditions. From the other perspective, binary tree classification is considered to be the building block of some regression model, and each subject is assigned an estimated risk that is commonly taken to be the empirical risk of outcome among subjects in the same node (or nodes for bootstrapped and boosted tree methods). The former approach considers a fixed classification structure by assigning outcome to leaf nodes, and has as many degrees of freedom as the number of cutoff values, while the latter approach uses trees to build regression models for risk, determining cutoff values in the process of model fitting (tree growing), and ends up having only one degree of freedom – the estimated risks. Current machine learning methods often take the second perspective.

Essential to evaluation from either perspective are receiver operating characteristics (ROC) curve and its area under curve (AUC), a set of evaluation tools widely used for a single marker (Hanley and McNeil, 1982). Denote by X the marker variable under consideration, and by Y a binary outcome that is correlated with X such that a larger X indicates a higher risk for outcome $Y = 1$. ROC curve is then created by plotting true positive rate (TPR) against false positive rate (FPR), where

$$TPR(x) = \mathbb{P}(X > x | Y = 1), \quad FPR(x) = \mathbb{P}(X > x | Y = 0).$$

More rigorously, we can define inverse functions of TPR and FPR as

$$TPR^{-1}(t) = \inf\left\{x : TPR(x) < t\right\}, \quad FPR^{-1}(t) = \inf\left\{x : FPR(x) < t\right\},$$

and then define ROC curve as the plot of function

$$ROC(t) = TPR\{FPR^{-1}(t)\}.$$

The ROC curve is an intuitive illustration of marker's discriminative power – the more ROC is curved towards the upper left corner, the higher prediction power a marker has, which is quantified by area under the ROC curve.

Formally, we define

$$AUC = \int_0^1 ROC(t)\,dt = \int_0^1 TPR(x)\,d\underset{X|Y=0}{\mathbb{P}}(x).$$

Interestingly, AUC has an interpretation as the concordance probability. Denote two independent and identical copies of (X, Y) by (X_1, Y_1) and (X_2, Y_2), and then note that $AUC = \mathbb{P}(X_1 < X_2 | Y_1 < Y_2) + 0.5\,\mathbb{P}(X_1 = X_2 | Y_1 < Y_2)$. It llustrates the ranked-based feature of ROC and AUC, as compared to linearity association captured by Pearson's correlation coefficient.

Evaluation for tree classifiers based on the second perspective is relatively straightforward. After obtaining a risk estimate for each subject, these estimates are treated as observed marker values allowing ROC and AUC methods to be directly applied. However, in doing so the interpretability of using trees is mostly lost, and therefore this approach is less favored by biomedical researchers. On the contrary, the first approach that focuses on tree structure and allows cutoff values to vary is commonly taken in clinical trials and cohort studies to identify useful markers to collect in future stages or studies, but evaluation of prediction performance brings additional challenges as a price of the flexibility and interpretability of using a tree structure. The major difference between a single-marker tree for which ROC and AUC are developed and a multiple-marker tree is that in the latter case TPR and FPR are no longer one-to-one functions, and ROC and AUC are not well-defined without further adjustments. An algorithmic method was proposed by Baker, 2000 to find optimal cutoff values for a fixed tree, but the approach discretized markers, which is often inapplicable in practice. Wang and Li, 2012 and Wang and

Li, 2013 extended definitions for ROC and AUC by considering an averaged prediction measure for a tree that uses multiple markers. For two markers (X_1, X_2) such that a large value in both implies positive classification $Y = 1$, the ROC function is defined to be

$$ROC(t) = \mathbb{E}\left[TP(X_1, X_2)|FP(X_1, X_2), Y = 0\right],$$

where

$$TP(x_1, x_2) = \mathbb{P}(X_1 > x_1, X_2 > x_2|Y = 1),$$

$$FP(x_1, x_2) = P(X_1 > x_1, X_2 > x_2|Y = 0).$$

To account for the distribution of $Q_0 = FP(X_1, X_2)$ conditional on $Y = 0$, a weighted ROC function is defined as

$$WROC(t) = ROC(t) \cdot h_0(t),$$

where $h_0(t)$ is the derivative of probability measure of Q_0. The area under $WROC(t)$ is shown to be equivalent to the concordance probability of correctly ordering markers under the bivariate scenario, a nice property consistent with the univariate ROC results. These definitions naturally extends to the case of more markers. However, these considerations do not address the more pertinent question of how well a fixed tree can predict outcome when its cutoff values achieve optimality, or how to identify these cutoff values. A solution to these questions is of great practical value as it could be used to guide decision making in biomedical researches that use tree-based classifiers. It is also of theoretical interest, since optimal fixed tree is fundamental to all tree-based methods but its features are not yet well understood from a statistical inferential perspective.

1.3 Optimal Combination of Multiple Markers

As indicated by the Neyman-Pearson Lemma, the uniformly most powerful test for classifying binary outcome Y using marker or marker vector M is based on the risk score $\mathbb{P}(Y = 1|X)$. This result has long been known in the literature of signal detection, but has not been brought to attention to the statistical literature until the paper of McIntosh and Pepe, 2002. Various risk score models were studied to find optimal linear combination of markers by McIntosh and Pepe, 2002 and Pepe and Thompson, 2000 among others. Focusing directly on the evaluation measure instead, Pepe, Cai, and Longton, 2006 investigated the linear coefficient that optimized the area under ROC curve. This approach is closely related to the general linear model studied by Han, 1987 that assumes $\mathbb{P}(Y = 1|X) = g(X^\mathsf{T}\beta)$, for some monotone transformation $g(\cdot)$. A maximum rank correlation estimator was proposed, which is exactly what would be obtained by the direct optimization of AUC.

However, a composite marker formed by linear combination lacks flexibility and may not be relevant to the context when markers are combined from different domains. In contrast, a non-linear combination using tree-based methods could be more flexible and interpretable.

References

Breiman, Leo, Jerome Friedman, Charles J Stone, and Richard A Olshen (1984). *Classification and regression trees*. CRC press.

Friedman, Jerome H (2002). “Stochastic gradient boosting”. In: *Computational Statistics & Data Analysis* 38.4, pp. 367–378.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2001). *The elements of statistical learning*. Vol. 1. Springer series in statistics New York.

Breiman, Leo (1996). “Bagging predictors”. In: *Machine learning* 24.2, pp. 123–140.

Rodriguez, Juan José, Ludmila I Kuncheva, and Carlos J Alonso (2006). “Rotation forest: A new classifier ensemble method”. In: *IEEE transactions on pattern analysis and machine intelligence* 28.10, pp. 1619–1630.

Scornet, Erwan, Gérard Biau, Jean-Philippe Vert, et al. (2015). “Consistency of random forests”. In: *The Annals of Statistics* 43.4, pp. 1716–1741.

Biau, GÃŠrard, Luc Devroye, and GÃĄbor Lugosi (2008). “Consistency of random forests and other averaging classifiers”. In: *Journal of Machine Learning Research* 9.Sep, pp. 2015–2033.

Hanley, James A and Barbara J McNeil (1982). “The meaning and use of the area under a receiver operating characteristic (ROC) curve.” In: *Radiology* 143.1, pp. 29–36.

Baker, Stuart G (2000). “Identifying combinations of cancer markers for further study as triggers of early intervention”. In: *Biometrics* 56.4, pp. 1082–1087.

Wang, Mei-Cheng and Shanshan Li (2012). “Bivariate marker measurements and ROC analysis”. In: *Biometrics* 68.4, pp. 1207–1218.

Wang, Mei-Cheng and Shanshan Li (2013). “ROC analysis for multiple markers with tree-based classification”. In: *Risk Assessment and Evaluation of Predictions*. Springer, pp. 179–198.

McIntosh, Martin W and Margaret Sullivan Pepe (2002). “Combining several screening tests: optimality of the risk score”. In: *Biometrics* 58.3, pp. 657–664.

Pepe, Margaret Sullivan and Mary Lou Thompson (2000). “Combining diagnostic test results to increase accuracy”. In: *Biostatistics* 1.2, pp. 123–140.

Pepe, Margaret Sullivan, Tianxi Cai, and Gary Longton (2006). “Combining predictors for classification using the area under the receiver operating characteristic curve”. In: *Biometrics* 62.1, pp. 221–229.

Han, Aaron K (1987). “Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator”. In: *Journal of Econometrics* 35.2-3, pp. 303–316.

Chapter 2

Literature Review: Survival Analysis for Recurrent Events

2.1 Recurrent Event Data

Recurrent event data is commonly encountered in longitudinal follow-up studies, when each subject experiences multiple events under observation until loss to follow-up, dropout or end of study occurs. These multiple events could be considered to be of different types, such as the events of HIV infections, AIDS diagnosis and death, or of the same type such as repeated hospitalization of cardiovascular disease patients.

For events of different types, researchers often consider the number of possible event occurrences for a subject to be fixed, although some of the later events could be censored or never occur during lifetime. In this case, statistical methods are focused on multivariate or multistage perspective of the data, and are developed to model either the times between successive events, which is sometimes termed gap times, or time from baseline to events. In contrast, for events of the same type, the number of individual event occurrences is naturally considered a random variable, which could be informative of some underlying individual characteristics. For instance, for a study following a group of cardiovascular disease patients for repeated hospitalization over a fixed period of time, one would expect patients with more severe conditions to experience more frequent hospitalizations, and the number of occurrences could vary over a wide range across individuals.

Common to all categories of recurrent event data is the heterogeneity among and correlation within subjects that need to be taken into account in modeling. However when gap times are studied, additional difficulties arise as dependent censoring is induced on all gap times except the first one. Bypassing this difficulty, many methods focus on a point process perspective with time index defined as the time from baseline to events (Lancaster and Intrator, 1998; Cook and Lawless, 2007), although sometimes the scientific interest is actually on gap times. For the remaining sections in this chapter, we review statistical methods developed for recurrent event data of the same type such that the number of observed events for each subject is a random variable.

2.2 Statistical Methods for Time-to-Events Data

Denote by $N(t)$ the recurrent event point process that counts the number of events experienced at or prior to time t since time origin, where $t \in (0, \tau]$ for some constant τ. The intensity function of continuous point process $N(t)$ fully determines its probability structure, and is defined as

$$\lambda_N\{t|\mathcal{H}(t)\} = \lim_{\Delta\to 0^+} \frac{\mathbb{P}\left\{N(t+\Delta) - N(t) > 0|\mathcal{H}(t)\right\}}{\Delta},$$

where $\mathcal{H}(t)$ represents the process history up till time t. Focusing on the intensity function, conditional regression methods are proposed by Andersen and Gill, 1982 that extended Cox's proportional hazard model (Cox, 1975) under independent censoring assumptions. Let $X(t)$ denote possibly time-dependent covariate history prior to time t, and let $Z(t) = \phi\{\mathcal{H}(t), X(t)\}$ denote some transformation of $\{\mathcal{H}(t), X(t)\}$. The model then assumes for $t \in (0, \tau]$ that

$$\lambda_N\{t|\mathcal{H}(t), X(t)\} = \lambda_0(t)e^{Z(t)^\intercal\beta},$$

for some baseline function $\lambda_0(t) > 0$ and linear coefficient β. For estimation of β, partial likelihood methods were extended from univariate survival to recurrent event data, while baseline function can be estimated using the Nelson-Aalen estimator (Aalen, 1978). Asymptotic properties of these estimators were established using martingale theories.

Anderson and Gill's conditional regression model can be considered as a prediction model due to the inclusion of event history in the conditional statistics, but is less appropriate for identifying population-level effects. Meanwhile, validity of the model assumptions depend highly on the transformation function ϕ. If , for instance, $Z(t)$ is taken to be time-independent, the model then requires recurrent event process to be memoryless, which is a very strong assumption especially in the context of biomedical studies related to any kind of progression. Alternatively, we can study the rate function defined as

$$\lambda_N(t) = \lim_{\Delta \to 0^+} \frac{\mathbb{P}\left\{N(t+\Delta) - N(t) > 0\right\}}{\Delta},$$

where the event history is not included as part of the conditional statistics. Focusing on the rate function, a marginal regression model was proposed by Lin et al., 2000 under independent censoring assumption:

$$\lambda_N\{t|X(t)\} = \lambda_0(t)e^{X(t)^\top \beta}.$$

Parameters β and $\lambda_N(t)$ were estimated using partial likelihood methods and Nelson-Aalen estimator similar to those used by Andersen and Gill, 1982. Large sample properties were established using modern empirical process theories.

The marginal regression model is suitable for estimating treatment effects and identifying population risk factors, but the results are contingent upon validity of the independent censoring assumption which is often violated in the presence of death or informative drop-out. To deal with this issue, Wang, Qin, and Chiang, 2001 proposed a latent variable model assuming

$$\lambda_N(t|W, X) = W \cdot \lambda_0(t)e^{X^\top \beta},$$

which allows informative censoring of the recurrent event process through some possibly unobserved random variable W. Their approach then avoided estimating the latent variable and the non-parametric component $\lambda_0(t)$ using conditional likelihood techniques. Related to this work, Huang and Wang, 2004 proposed a joint model for recurrent events and failure time by using a shared latent variable.

In addition to the failure event observed at the end of a recurrent event process, the statistical literature studying time-to-events data is rich in dealing with many other practical issues encountered in biomedical research. To study longitudinal measures collected at the recurrence of events, Wu and Carroll, 1988, Tsiatis, Degruttola, and Wulfsohn, 1995, Hogan and Laird, 1997 and Xu and Zeger, 2001, among others, considered the longitudinal measures to be a marker process, and proposed joint models for the marker and recurrent event processes. Multivariate recurrent event process was studied by Sun, Zhu, and Sun, 2009 and Ning et al., 2015 among others, and a dependency measure between two processes was proposed by the latter. Overall, statistical methods for recurrent time-to-events data are well established in various contexts, but these methods are only applicable when the scientific interest is placed on occurrence rate of events over time. When the outcome variable of interest is the gap time between successive events, or when there is a strong cyclical pattern of recurrence, it is more appropriate to study time-between-events models instead.

2.3 Statistical Methods for Time-Between-Events Data

For a recurrent event process $N(t)$, let T_j be the jth gap time between jth and $(j+1)$th events for $j = 1, 2, \ldots$, and denote by $X(t)$ some associated covariate history prior to time t. Parametric transitional probability models and parametric frailty models can be studied using maximum likelihood methods, but these models lack flexibility and are less favored in practice compared to non-parametric or semi-parametric approaches. Prentice, Williams, and Peterson, 1981 also proposed a conditional proportional hazard model for time-between-events data assuming

$$\lambda_N\{t|N(t^-) = j-1, \mathcal{H}(t), X(t)\} = \lambda_{0j}(t - t_{j-1})e^{X(t)^\intercal \beta_j},$$

where $\lambda_{0j}(t)$ and β_j are possibly gap-specific baseline function and linear coefficient. As a variation of the time-to-events model proposed in the same paper, this model was also estimated using partial likelihood methods, and asymptotic properties were established using martingale theories. As a conditional model, it is more appropriate to be used for prediction than for identifying any population effects.

Time-between-events data can also be considered as clustered survival data. Taking this perspective, Pena, Strawderman, and Hollander, 2001 proposed a multiplicative frailty model extending the Cox's proportional hazards model. They assumed that

$$\mathbb{E}\{dN(t)|X(t), W\} = W \cdot \lambda_0(t)e^{X(t)^\intercal \beta},$$

where W is some subject-specific frailty following some pre-specified parametric distribution, $\lambda_0(\cdot)$ is the baseline hazard function, and β is the linear coefficient. A common choice of frailty distribution is the gamma distribution for computational convenience. This model deals with induced informative censoring by jointly modeling gap times $T_1, T_2, \dots$ through a parametric latent variable, but bases analyses on unverifiable assumptions, and is less robust to various subject-level correlation structures.

In comparison, marginal models are more robust, and are useful when researchers are interested in population-level effects of covariates, but limited techniques have been developed. Extending the accelerated failure time model, Huang, 2002 proposed a marginal model estimated by an estimating equation exploiting the additive property of gap times on the log-transformed scale under the model assumption. Strawderman, 2005 also proposed an accelerated failure time model on gap times, and the methods were developed under the strong assumption that gap times are independent conditional on some baseline covariates. Following risk-set methods, Wang and Chen, 2000 proposed a class of non-parametric estimators for the marginal survival function of exchangeable recurrent gap times, and the method was extended by Huang and Chen, 2003 to the regression model $\lambda\{t|X(0)\} = \lambda_0(t)\exp\left\{X(0)^{\mathsf{T}}\beta\right\}$, where $X(0)$ is some baseline covariate. This model is a natural extension of the classic proportional hazard model, and solves the problem of induced dependent censoring by leaving out the last gap time except when only one gap time is observed for a subject. Similar methods were developed assuming different model forms by Sun, Park, and Sun, 2006, Darlington and Dixon, 2013 and Ding and Sun, 2017. However these models only allow the use of baseline covariates and assume exchageability between gap times, and are therefore inapplicable when there exist temporal trends or when longitudinal covariates are collected and are of scientific interest. Models and methods dealing with these issues are yet to be developed.

References

Lancaster, Tony and Orna Intrator (1998). “Panel data with survival: hospitalization of HIV-positive patients”. In: *Journal of the American statistical association* 93.441, pp. 46–53.

Cook, Richard J and Jerald F Lawless (2007). *The statistical analysis of recurrent events*. Springer Science & Business Media.

Andersen, Per Kragh and Richard D Gill (1982). “Cox’s regression model for counting processes: a large sample study”. In: *The annals of statistics*, pp. 1100–1120.

Cox, David R (1975). “Partial likelihood”. In: *Biometrika* 62.2, pp. 269–276.

Aalen, Odd (1978). “Nonparametric inference for a family of counting processes”. In: *The Annals of Statistics*, pp. 701–726.

Lin, DY, LJ Wei, I Yang, and Z Ying (2000). “Semiparametric regression for the mean and rate functions of recurrent events”. In: *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, pp. 711–730.

Wang, Mei-Cheng, Jing Qin, and Chin-Tsang Chiang (2001). “Analyzing recurrent event data with informative censoring”. In: *Journal of the American Statistical Association* 96.455, pp. 1057–1065.

Huang, Chiung-Yu and Mei-Cheng Wang (2004). “Joint modeling and estimation for recurrent event processes and failure time data”. In: *Journal of the American Statistical Association* 99.468, pp. 1153–1165.

Wu, Margaret C and Raymond J Carroll (1988). “Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process”. In: *Biometrics*, pp. 175–188.

Tsiatis, AA, Victor Degruttola, and MS Wulfsohn (1995). “Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS”. In: *Journal of the American Statistical Association* 90.429, pp. 27–37.

Hogan, Joseph W and Nan M Laird (1997). “Mixture models for the joint distribution of repeated measures and event times”. In: *Statistics in medicine* 16.3, pp. 239–257.

Xu, Jane and Scott L Zeger (2001). “Joint analysis of longitudinal data comprising repeated measures and times to events”. In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 50.3, pp. 375–387.

Sun, Liuquan, Liang Zhu, and Jianguo Sun (2009). “Regression analysis of multivariate recurrent event data with time-varying covariate effects”. In: *Journal of Multivariate Analysis* 100.10, pp. 2214–2223.

Ning, Jing, Yong Chen, Chunyan Cai, Xuelin Huang, and Mei-Cheng Wang (2015). “On the dependence structure of bivariate recurrent event processes: inference and estimation”. In: *Biometrika* 102.2, pp. 345–358.

Prentice, Ross L, Benjamin J Williams, and Arthur V Peterson (1981). “On the regression analysis of multivariate failure time data”. In: *Biometrika* 68.2, pp. 373–379.

Pena, Edsel A, Robert L Strawderman, and Myles Hollander (2001). “Nonparametric estimation with recurrent event data”. In: *Journal of the American Statistical Association* 96.456, pp. 1299–1315.

Huang, Yijian (2002). “Censored regression with the multistate accelerated sojourn times model”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.1, pp. 17–29.

Strawderman, Robert L (2005). “The accelerated gap times model”. In: *Biometrika* 92.3, pp. 647–666.

Wang, Mei-Cheng and Ying-Qing Chen (2000). “Nonparametric and semiparametric trend analysis for stratified recurrence times”. In: *Biometrics* 56.3, pp. 789–794.

Huang, Yijian and Ying Qing Chen (2003). “Marginal regression of gaps between recurrent events”. In: *Lifetime data analysis* 9.3, pp. 293–303.

Sun, Liuquan, Do-Hwan Park, and Jianguo Sun (2006). “The additive hazards model for recurrent gap times”. In: *Statistica Sinica*, pp. 919–932.

Darlington, GA and SN Dixon (2013). “Event-weighted proportional hazards modelling for recurrent gap time data”. In: *Statistics in medicine* 32.1, pp. 124–130.

Ding, Jieli and Liuquan Sun (2017). “Additive mixed effect model for recurrent gap time data”. In: *Lifetime data analysis* 23.2, pp. 223–253.

Chapter 3

Optimal Decision Rule for Combining Multiple Biomarkers into Tree-based Classifier and its Evaluation

3.1 Introduction

Biomarkers, or biological markers, refer to measurements of a specific feature as depiction of a biological state, used for diagnosis concerning biological or pathogenic processes, or of pharmacologic responses to a treatment intervention. Biomarkers used for disease diagnosis are also referred to as prognostic markers. Tools that investigate the performance of a single prognostic biomarker, such as receiver operating characteristic curve (ROC) and area under curve (AUC), have been well studied. In real applications, multiple markers are commonly collected, but it remains a question how to optimally combine multiple markers for predicting disease outcome. ROC for single marker is well-defined, because both $TP(\cdot)$ and $FP^{-1}(\cdot)$ are well-defined functions, a nice property that is not naturally inherited when we have multiple markers. A common practice that deals with this problem is to combine multiple markers linearly, so that multiple markers are reduced to one “combined" marker. Methods to optimally combine markers in a linear fashion have been studied under various model assumptions for predicting binary disease outcomes. For examples, the Neyman-Pearson lemma can be connected to the optimality of risk score and the result was brought to the attention of statistical literature by McIntosh and Pepe, 2002; Pepe and Thompson, 2000 and Pepe, Cai, and Longton, 2006 studied linear discriminant analysis, logistic regression, and direct optimization of area under receiver operating characteristic curve.

However, a composite marker formed by linear combination lacks flexibility and may not be relevant when markers are combined from different domains. In contrast, a non-linear combination using tree-based methods could be more flexible and interpretable, which, specially, is already being commonly used in biomedical applications. For example, when several tests are performed on a patient, one possible practice is to diagnose him or her as diseased if all the test results are positive. Besides, tree-based structures handle correlation between markers with a more nonparametric and flexible manner than linear structures.

While evaluating the performance of a single biomarker using ROC or AUC is straightforward, doing so for a general tree involves additional difficulty, because the true positive and false positive rate functions are not well-defined. To study upper boundary of this band, Baker, 2000 considered discretized positivity region to evaluate marker performance based on a utility function. For continuous markers, to estimate the upper boundary curve of the ROC band based on two markers, Jin and Lu, 2009 proposed a bivariate kernel estimator to estimate the upper boundary curve of the ROC band but indicated the unstable performance of their estimator. In general, when multiple markers are used with a tree-based classifier, the quantile function of false positive rate is not one-to-one and the area under the upper boundary curve does not possess the interpretation as AUC in the single marker case. Wang and Li, 2012; Wang and Li, 2013 proposed a population-averaged ROC curve together with a weighted AUC as tools to evaluate the performance of multiple markers using tree classifiers. Of note, Wang and Li's work focused on the population-averaged performance of ROC and AUC, which is substantially different from the aim of this work, which is to search for and estimate the optimal prediction performance of a fixed classification tree structure.

3.2 Fixed Tree Classifier and its Representation

We consider some fixed tree structure denoted by T that uses multiple markers and allows cutoff values to vary at nodes for classification of some binary outcome $Y = 0$ or 1, and we refer to it as a tree classifier, or simply as a tree when there is no confusion. Denote by $X = (X_1, \ldots, X_K)^{\mathsf{T}} \in S_X$ the markers used as splitting covariates, by $\widetilde{x} = (\widetilde{x}_1, \ldots, \widetilde{x}_K)^{\mathsf{T}} \in S_X$ a generic realization of corresponding markers, and by $c = (c_1, \ldots, c_K)^{\mathsf{T}} \in S_c$ the corresponding varying cutoff values. Tree classifiers of this type are commonly used in biomedical researches due to their flexibility and interpretability, but there is a lack of methods for finding optimal cutoff values at each node and evaluating a tree's prediction performance. Part of the difficulties in developing these methods lie in the lack of algebraic representations of trees that can be used in an extendable statistical framework.

To overcome this, we first observe that, despite various graph structures, a tree's classifying behavior is solely determined by the marker space attributed to $Y = 1$ (the positive group) when cutoff values are given. Formally, define positivity region $R(c; X, T)$ to be the set of $\widetilde{x}$ classified positive ($Y = 1$) by tree T given cutoff values c of marker X. Two trees, T_1 using markers X_1 and T_2 using X_2, with different graph representation are considered identical in terms of classification if for every c_1 there exists c_2 such that $R(c_1; X_1, T_1) = R(c_2; X_2, T_2)$ and vice versa. Therefore it is sufficient to study the positivity region of a tree. The second observation is that if we consider the collection of bifurcated marker spaces created at each node, the final positivity region can be obtained by performing intersection and union operations over a sub-collection of these sets. Intuitively, any tree classification rule can be represented as individual classification rule of the form "$X_k > c_k$" or "$X_k < c_k$" linked by "and" and "or" logical operators.

To find a standard representation, we first consider leaf nodes assigned $Y = 1$ and index these nodes by $j = 1, \ldots, J$. Denote by $R_j(c; X, T)$ the marker region attributed to $Y = 1$ by the jth leaf node, and we have $R(c; X, T) = \cup_{j=1}^{J} R_j(c; X, T)$. Then consider the nodes "traveled" from root node to the jth leaf node, and suppose markers X_k for $k \in \kappa_j$ are used for data splitting at

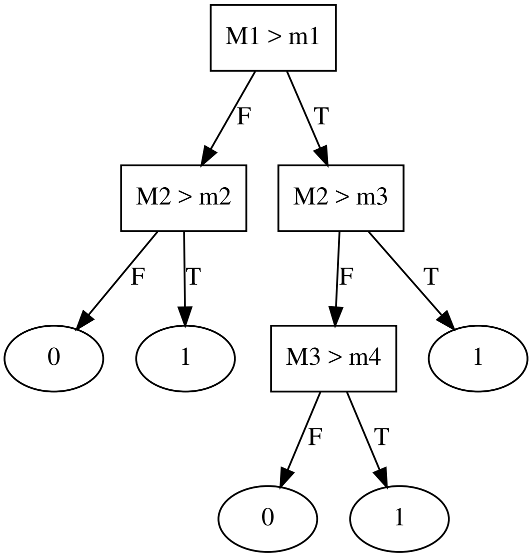

Figure 3.1: Example of a tree classifier that uses three markers M_1, M_2 and M_3, but has four splitting nodes thus four cutoff values to optimize.

these traveled nodes. We can assume without loss of generality that at each node we obtain the marker space satifying $X_k > c_k$, as we can always reverse the sign of a marker. This implies that $R_j(c; X, T) = \cap_{k \in \kappa_j} \{\widetilde{x} \in S_X : \widetilde{x}_k > c_k\}$, which yields what we call the standard representation of positivity region in the following form:

$$R(c; X, T) = \cup_{j=1}^{J} \big[\cap_{k \in \kappa_j} \{\widetilde{x} \in S_X : \widetilde{x}_k > c_k\} \big].$$

For further simplification, we assume that κ_j's are disjoint and $\cup_{j=1}^{J} \kappa_j = \{1, \ldots, K\}$, because if there exists any repeatedly used marker, we can add one or more additional copies of it to the initial marker vector X under consideration along with appropriate modification to S_X and S_c. For simplicity, we sometimes write set $\{\widetilde{x} \in S_X : \widetilde{x}_k > c_k\}$ as $\{X_k > c_k\}$.

Now we illustrate the derivation of a fixed tree's standard representation using the tree classifier as shown in Figure 3.1. Suppose supports of (M_1, M_2, M_3) and (m_1, m_2, m_3) are both Euclidean space $\mathbb{R}^3$. Three leaf nodes are classified as group 1, which we index by $1, 2, 3$ going from left to right. Marker spaces attributed to 1 by three leaf nodes are

$$\big\{M_1 < m_1\big\} \cap \big\{M_2 > m_2\big\},$$

$$\big\{M_1 > M_1\big\} \cap \big\{M_2 < m_3\big\} \cap \big\{M_3 > m_4\big\} \text{ and}$$

$$\big\{M_1 > m_1\big\} \cap \big\{M_2 > m_3\big\}$$

respectively. After changing signs and adding additonal copies of marker when the marker is used repeatedly, we obtain

$$R_1(c; X, T) = \big\{X_1 > c_1\big\} \cap \big\{X_2 > c_2\big\},$$

$$R_2(c; X, T) = \big\{X_3 > c_3\big\} \cap \big\{X_4 > c_4\big\} \cap \big\{X_5 > c_5\big\},$$

$$R_3(c; X, T) = \big\{X_6 > c_6\big\} \cap \big\{X_7 > c_7\big\},$$

where $X = (X_1, \ldots, X_7)^{\mathsf{T}} = (-M_1, M_2, M_1, -M_2, M_3, M_1, M_2)^{\mathsf{T}}$ and $c = (c_1, \ldots, c_7)^{\mathsf{T}} = (-m_1, m_2, m_1, -m_3, m_4, m_1, m_3)$. Correspondingly, we have $S_X = \big\{x \in \mathbb{R}^7 : x_1 = -x_3 = -x_6, x_2 = -x_4 = x_7\big\}$ and $S_c = \big\{c \in \mathbb{R}^7 : c_1 = -c_3 = -c_6, c_4 = -c_7\big\}$. With these specification we obtain the standard representation for tree in Figure 3.1 as

$$R(c; X, T) = \cup_{j=1}^{3} \big[\cap_{k \in \kappa_j} \{\widetilde{x} \in S_X : \widetilde{x}_k > c_k\}\big]$$

for $c \in S_c$, where $\kappa_1 = \{1,2\}$, $\kappa_2 = \{3,4,5\}$, and $\kappa_3 = \{6,7\}$.

3.3 Receiver Operating Characteristic Band and Optimal Receiver Operating Characteristic Curve

Having identified a standard representation of a tree classifier, we are now ready to generalize the definitions of true positive rate and false positive rate in the single marker scenario. We consider continuous marker vector $X \in S_X \subset \mathbb{R}^K$ for the simplicity of discussions, but all results can be extended to include discrete ordinal markers with some minor technical modifications. For positivity region $R(c; X, T)$ we define

$$TPR(c) = \mathbb{P}\left\{X \in R(c; X, T) | Y = 1\right\},$$

$$FPR(c) = \mathbb{P}\left\{X \in R(c; X, T) | Y = 0\right\}.$$

We also generalize the inverse of TPR and FPR to set-valued functions

$$TPR^{-1}(t) = \left\{c \in S_c : TPR(c) = t\right\},$$

$$FPR^{-1}(t) = \left\{c \in S_c : FPR(c) = t\right\},$$

for $t \in [0,1]$, which further implies the generalization of ROC curve to what we call the ROC band (ROCB) as the graph of set-valued function

$$ROCB(t) = TPR\left\{FPR^{-1}(t)\right\} = \left\{TPR(c) : FPR(c) = t\right\}.$$

It was referred to as a “band" since with the generalization using set-valued functions, for each false positive rate there exists multiple true positive rates,

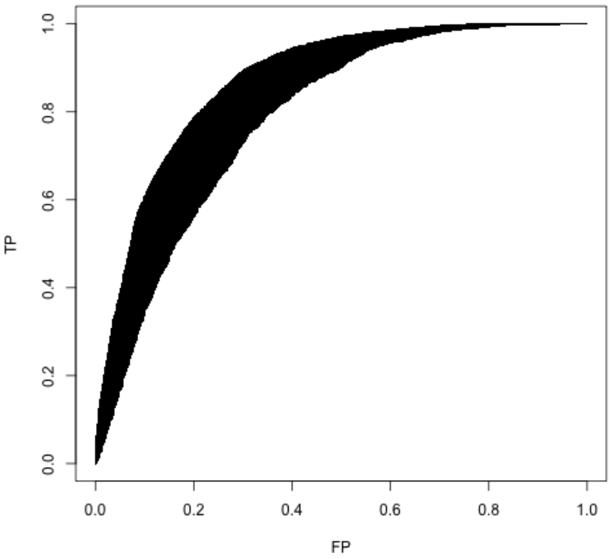

Figure 3.2: ROC band generated for marker $X \in \mathbb{R}^2$ in prediction of binary outcome $Y = 0, 1$, where X follows bivariate standard normal conditional on $Y = 0$, and bivariate normal with mean vector $(1, 1)^\top$, marginal variances 0.5 and correlation 0 conditional on $Y = 1$. A subject is classified to have outcome $Y = 1$ if both marker values exceed some threshold.

and overall the $ROCB$ function spans a band over $[0, 1]$.

What an ROC band captures is the range of prediction performance for a fixed tree classifier – given a cutoff value, the TPR and FPR pair then falls on the ROC band. If practitioners randomly choose the cutoff values, the average prediction performance in the population can be depicted by measures like those proposed in Wang and Li, 2013. However, it is of more practical interest to study the upper boundary of the ROC band, which captures the “best" performance possible using given tree classifier.

See Figure 3.2 for the ROC band generated by simulation for marker $X \in \mathbb{R}^2$ combined using the “and" operator, where X follows bivariate standard normal contional on $Y = 0$, and bivariate normal distribution with mean vector $(1,1)^{\mathsf{T}}$, marginal variances 0.5 and correlation 0 conditional on $Y = 1$. This setup is intended to mimic possible distributions of two independently informative biomarkers for some disease – in the non-disease population biomarkers are lower on average, but there is great heterogeneity, while in the diseased population biomarkers are higher on average but have less heterogeneity. A subject is considered diseased if most marker values exceed some threhold. When FPR is at 0.2, the corresponding TPR ranges from approximately 0.55 to 0.8. If a practitioner wants to have a FPR no greater than 0.2 but then chooses cutoff values without being further informed, he or she could end up with a TPR anywhere between 0.55 and 0.8, risking to lose a lot of efficiency and resources. It is therefore desirable to find, or to approximate “optimality" – the cutoff values that give us the highest TPR for some given FPR.

Due to the optimality implication of the ROC band upper boundary, we refer to it as the optimality ROC curve, which is formally defined as the graph of function

$$OROC(t) = \sup ROCB(t) = \sup \left\{ TP(c) : FP(c) = t \right\}$$

for $t \in [0,1]$. Intuitively, the area under optimality ROC curve (AUOROC) can then be used to evaluate the overall optimal prediction performance of a tree with varying TPR and FPR, and we define it to be

$$AUOROC = \int_0^1 OROC(t)\, dt. \tag{3.1}$$

The ROC band and the optimality ROC curve have some interesting properties that can be linked to the ROC curve. It is easy to establish the equivalent definition

$$OROC(t) = \sup \left\{ TP(c) : FP(c) \leq t \right\}, \tag{3.2}$$

and to show that $OROC(t)$ is monotonically increasing in t. Similar to ROC curve, we have

$$ROCB(0) = 0, ROCB(1) = 1, OROC(0) = 0, \text{ and } OROC(1) = 1,$$

and that both the ROC band and the optimality ROC curve degenerates to the ROC curve in the single marker scenario. It can also be shown that $OROC(t)$ is continuous and monotonically increasing in t.

3.4 Empirical Estimation of Optimal Receiver Operating Characteristic Curve

Suppose we observe data consisting of (x_i, y_i) for $i = 1, \ldots, n$ that are n i.i.d. copies of (X, Y). For positivity region $R(c; X, T)$ of given tree classifier T, we can estimate TPR and FPR empirically as

$$\widehat{TPR}(c) = \frac{\sum_{i=1}^{n} \mathbb{1}\left\{ x_i \in R(c; X, T), y_i = 1 \right\}}{\sum_{i=1}^{n} \mathbb{1}(y_i = 1)},$$

$$\widehat{FPR}(c) = \frac{\sum_{i=1}^{n} \mathbb{1}\left\{ x_i \in R(c; X, T), y_i = 0 \right\}}{\sum_{i=1}^{n} \mathbb{1}(y_i = 0)}.$$

These two estimators can then be plugged into (3.2) and (3.1) to obtain empirical estimator of $OROC$ and $AUOROC$ as

$$\widehat{OROC}(t) = \sup\left\{\widehat{TPR}(c) : \widehat{FPR}(c) \leq t\right\},$$

$$\widehat{AUOROC} = \int_0^1 \widehat{OROC}(t)\,dt.$$

We prove the following result for estimators $\widehat{OROC}(t)$ and $\widehat{AUOROC}$ in Section 3.9.

Theorem 1. $\widehat{OROC}$(t) is uniformly strongly consistent for $OROC$(t), that is,

$$\sup_t \left|\widehat{OROC}(t) - OROC(t)\right| \to 0$$

almost surely. As a result, $\widehat{AUOROC}$ is strongly consistent for $AUOROC$.

However, not only is statistical inference difficult to obtain, the estimators also have non-negligible positive biases, issues both arising from the use of supremum in the definition of $OROC(t)$. An intuitive explanation comes from the asymptotic behavior of $\widehat{OROC}(t)$ when cutoff values, and thus t, have discrete support. Suppose c_j's for $j = 1, \ldots, J$ are some cutoff values such that $\widehat{TPR}\{c_{(j)}\}$ for $j = 1, \ldots, J$ are all possible true positive rates when $\widehat{FPR}\{c_{(j)}\}$ is equal to some constant, and $TPR\{c_{(j)}\}$ forms an increasing sequence. Each $\widehat{TPR}\{c_{(j)}\}$ is asymptotically normal, and the Jensen's inequality then implies that the expectation of $\widehat{OROC}(t)$ is greater than that of $\widehat{TPR}\{c_{(J)}\}$, which converges to $TPR\{c_{(J)}\}$. For these biases to be small, an unrealistically large sample size is needed. But even if one could collect a sample of sufficient size, one faces computational difficulties because computational complexity of the estimators grows at the rate of n^K. This is a rate exponential in K, implying the infeasibility in applying the empirical estimators to even only slightly complicated tree classifiers like the example in Figure 3.1 with $K = 7$.

3.5 Semi-Parametric and Rank-Based Estimation

Taking an optimization perspective, the problem we are interested in is to find those $c \in S_c$ such that

$$TPR(c) = OROC(t), \text{ and } FPR(c) = t \tag{3.3}$$

for any given $t \in [0, 1]$, and to estimate $OROC(t)$ with identified optimal cutoff values. Equation (3.3) implies

$$TPR(c) - OROC\{FPR(c)\} = 0,$$

which inspires us to define function

$$m(c) = TPR(c) - OROC\{FPR(c)\},$$

a continuous mapping from S_c to $\mathbb{R}$ whose solution graph forms some hypersurface $\mathcal{S} \subset S_c$. Our goal then translates into identifying those $c \in \mathcal{S}$ such that $FPR(c) = t$ and evaluating function value $TPR(c)$. We refer to $\mathcal{S}$ as the optimality hypersurface due to its connection to the optimality ROC curve.

This optimality hypersurface could have highly complicated structure and it is unrealistic and often impossible to derive closed forms for various distributions and tree classifiers. Even for the simple example as shown in Figure 3.2, we are not able to obtain closed form formula for $OROC(t)$

and thus not for $m(c)$. Instead of deriving closed form of $m(c)$ under some distributional assumption on (X, Y), we consider a class $\mathcal{H}$ of curves that are likely to lie in the optimality hypersurface. We would want this class to be large enough so that it likely contains a curve from $\mathcal{S}$ and that when it does not, it contains a curve that is close enough to $\mathcal{S}$ under some distance measure; we would also want this class to be structural enough to give us some theoretical properties and insights.

With these goals, we propose to assume that there exists in the optimality hypersurface a curve from the class of curves $\mathcal{H}$ with the parametric representation (in the context of calculus terminology rather than statistics terminology)

$$h_k(c_k; \theta) = c_0, \text{ for } k = 1, \ldots, K, \tag{3.4}$$

for continuous and monotonically increasing functions $h_k(\cdot; \theta) : \mathbb{R} \to \mathbb{R}$ indexed by p-dimensional parameter $\theta \in \Theta \subset \mathbb{R}^p$. For identifiability and without loss of generality we can take $h_1(\cdot; \theta)$ to be the identity mapping. For choices of $\mathcal{H}$, we can take $h_k(\cdot; \theta)$ to be continuous and monotonically increasing piece-wise linear functions with knots at percentiles, polynomial function, or smooth spline functions, all commonly used to approximate general continuous functions. In practice, we can even introduce some tuning parameters so that the parametrization is adapted to specific data structure. For instance, we can use as tuning parameters the number of knots for piece-wise linear functions, degree of polynomial for polynomials, and smoothness penalty parameter for smooth splines. Parametrizations can be highly flexible and tailored to context with appropriate specification. Assume that θ_0 indexes a curve belonging to the optimality hypersurface $\mathcal{S}$.

We obtain some interesting insights under this assumption. For $c \in S_c$ satisfying (3.4) and $FPR(c) = t$, some algebra gives us

$$\begin{aligned}
TPR(c) &= \mathbb{P}\left\{X \in R(c; X, T) | Y = 1\right\} \\
&= \mathbb{P}\left(\cup_{j=1}^{J} \left[\cap k \in \kappa_j \{X_k > c_k\}\right] \middle| Y = 1\right) \\
&= \mathbb{P}\left(\cup_{j=1} \left[\cap_{k\in\kappa_j} \left\{h_k(X_k; \theta_0) > c_1\right\}\right] \middle| Y = 1\right).
\end{aligned}$$

Writing $\vee$ for taking maximum over a set, and $\wedge$ for minimum over a set, we further derive from the last display that

$$TPR(c) = \mathbb{P}\left[\vee_{j=1}^{J} \left\{ \wedge_{k\in\kappa_j} h_k(X_k; \theta_0)\right\} > c_1 \middle| Y = 1\right].$$

Therefore when $\mathcal{H}$ indeed contains a curve indexed by θ_0 that belongs to the optimality hypersurface $\mathcal{S}$, the optimality ROC curve corresponding to positivity region $R(c; X, T)$ is exactly the ROC curve of random variable

$$H(X; \theta_0) = \vee_{j=1}^{J}\left\{ \wedge_{k\in\kappa_j} h_k(X_k; \theta_0)\right\}.$$

Derivations above can also be used to show that θ indexes a class of random variables $H(X; \theta)$ whose ROC curves fall on the ROC band. The definition of $OROC(t)$ further implies that θ_0 maximizes AUC of $H(X; \theta)$, which is equivalent to the concordance probability of correctly ranking two observations

(Hanley and McNeil, 1982). That is,

$$\theta_0 \in \operatorname*{argmax}_{\theta \in \Theta} S(\theta),$$

where

$$S(\theta) = \mathbb{P}\left\{H(X;\theta) > H(X';\theta) | Y = 1, Y' = 0\right\},$$

(X', Y') being an independent and identical copy of (X, Y).

These properties inspire us to consider estimator

$$\widehat{\theta} \in \operatorname*{argmax}_{\theta \in \Theta} S_n(\theta),$$

where

$$\begin{aligned} S_n(\theta) =& \binom{n}{2}^{-1} \sum_{i=1}^{n-1} \sum_{i<i'=2}^{n} \big[\mathbb{1}\left\{H(x_i;\theta) > H(x_{i'};\theta), y_i > y_{i'}\right\} + \\ & \qquad \mathbb{1}\left\{H(x_i;\theta) < H(x_{i'};\theta), y_i < y_{i'}\right\}\big] \\ =& \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{i'=n_0+1}^{n_0+n_1} \mathbb{1}\left\{H(x_i;\theta) < H(x_{i'};\theta)\right\} \times \frac{2n_0 n_1}{n(n-1)}, \end{aligned}$$

which is proportional to the empirical counterpart of concordance probability $S(\theta)$ based on observed data $(x_i, y_i), i = 1, \ldots, n$, where $y_i = 0$ for $i = 1, \ldots, n_0$ and $y_i = 1$ for $i = n_0 + 1, \ldots, n_0 + n_1 = n$. Asymptotic properties of $\widehat{\theta}$ are discussed in Section 3.6.

When $\mathcal{H}$ does contain a curve from the optimality surface, we identify the best classification rule with the given tree structure. Further, if the tree structure under investigation actually yields the globally optimal rule when cutoff values are chosen appropriately, $H(X;\theta_0)$ is then the overall optimal decision rule for classification, and we have

$$\mathbb{P}(Y = 1|X) = g\{H(X;\theta_0)\}$$

for some monotonically increasing function g, or equivalently

$$Y = \mathbb{1}\left\{H(X;\theta_0) + U_i\right\},$$

where U_i's are some i.i.d. errors. This model is a variation of the general linear model proposed by Han, 1987.

When $\mathcal{H}$ does not contain any curve that comes from the optimality surface $\mathcal{S}$, we have a misspecified model but θ_0 still indexes a random variable of the form $H(X;\theta)$ that has an AUC closest to $AUOROC$. With appropriate model tuning, we expect the difference to be small and that the random variable $H(X;\widehat{\theta})$ has an ROC curve that is close to the optimality ROC curve.

3.6 Asymptotic Properties and Statistical Inference

We study and present asymptotic properties of $\widehat{\theta}$ in this section. Write $Z = (X^\mathsf{T}, Y)^\mathsf{T}$ and the support of Z as S_Z. For a generic vector $z = (x^\mathsf{T}, y)^\mathsf{T} \in S_Z$ and $\theta \in \Theta$, we define

$$\tau(z;\theta) = \mathbb{E}\big[H(x;\theta) > H(X;\theta), y > Y\big] + \mathbb{E}\big[H(x;\theta) < H(X;\theta), y < Y\big].$$

For some function $f(\theta)$, denote

$$\nabla_m f(\theta) = \frac{\partial^m f(\theta)}{\partial \theta^m},$$

$$\big|\nabla_m f(\theta)\big| = \sum \Big|\frac{\partial^m f(\theta)}{\partial \theta^m}\Big|,$$

for $m = 1, 2$. Weak convergence of $\widehat{\theta}$ is then established in the following Theorem.

Theorem 2. Under regularity conditions given in Section 3.9, we have

$$n^{1/2}(\widehat{\theta} - \theta_0) \to N(0, V^{-1}\Delta V)$$

in distribution, where

$$2V = \mathbb{E}_Z\big\{\nabla_2 \tau(Z); \theta_0\big\},$$

$$\Delta = \mathbb{E}_Z\big\{\nabla_1 \tau(Z; \theta_0) \cdot \nabla_1 \tau(Z; \theta_0)^\mathsf{T}\big\}.$$

Consistent estimators of V and Δ can be constructed by numerical derivatives as discussed in Sherman, 1993. Specifically, let $\{\epsilon_n\}_{n=1}^{\infty}$ denote a sequence of real numbers going to zero as $n \to \infty$, and denote by $u_j \in \mathbb{R}^p$ a vector that has one as its jth component and zeros elsewhere. Define

$$\tau_n(z; \theta) = \sum_{i=1}^{n} \{g(z, z_i; \theta) + g(z_i, z; \theta)\},$$

where

$$g(z_1, z_2; \theta) = \mathbb{1}\{H(x_1; \theta) > H(x_2; \theta), y_1 > y_2\},$$

and $z_1 = (x_1^\mathsf{T}, y_1)^\mathsf{T}, z_2 = (x_2^\mathsf{T}, y_2)^\mathsf{T} \in S_Z$. We can then estimate Δ by $\widehat{\Delta} =$

$\left(\widehat{\delta}_{jl}\right)_{p \times p}$, where for $j, l = 1, \ldots, p$,

$$\widehat{\delta}_{jl} = \sum_{i=1}^{n} \widehat{q}_j(z_i; \widehat{\theta}) \cdot \widehat{q}_l(z_i; \widehat{\theta}),$$

$$\widehat{q}_j(z; \theta) = \epsilon_n^{-1} \cdot \left[\tau_n(z; \theta + \epsilon_n u_j) - \tau_n(z; \theta)\right].$$

We can similarly estimate V by $\widehat{V} = \left(\widehat{v}_{jl}\right)_{p \times p}$, where for $j, l = 1, \ldots, p$,

$$2\widehat{v}_{ij} = \sum_{i=1}^{n} \widehat{r}_{jl}(z_i; \widehat{\theta}),$$

$$\begin{aligned} \widehat{r}_{jl}(z; \theta) =& \epsilon_n^{-2} \cdot \left[\tau_n\left\{z; \theta + \epsilon_n(u_j + u_l)\right\} - \tau_n(z; \theta + \epsilon_n u_j) \right. \\ & \left. - \tau_n(z; \theta + \epsilon_n u_l) + \tau_n(z; \theta)\right]. \end{aligned}$$

Again by arguments in Sherman, 1993, $\widehat{\Delta}$ is consistent when $n^{1/2}\epsilon_n \to \infty$ and $\widehat{V}$ is consistent when $n^{1/4}\epsilon_n \to \infty$, implying the reasonable choices of ϵ_n. However, it can be tricky to choose a proper bandwidth ϵ_n in practice, because the empirical objective function could be not sufficiently small, especially when sample size is small. Choosing too small or too large a bandwidth would result in numerical instability or estimation bias. An alternative is to use bootstrap techiniques (Efron and Tibshirani, 1994) to approximate the asymptotic distribution of $\widehat{\theta}$ for statistical inferences.

After obtaining an estimate $\widehat{\theta}$ of θ_0 using the training sample $\left\{(x_i, y_i) : i = 1, \ldots, n\right\}$, we can estimate $AUOROC$ using an independent testing sample

$\{(x_i, y_i) : i = n+1, \ldots, n+n'\}$ as

$$\widehat{AUOROC} = \sum_{i=n+1}^{n+n'-1} \sum_{i<i'=2}^{n+n'} \Big[\mathbb{1}\left\{H(x_i;\widehat{\theta}) > H(x_i;\widehat{\theta}), y_i > y_{i'}\right\} + \\ \mathbb{1}\left\{H(x_i;\widehat{\theta}) < H(x_i;\widehat{\theta}), y_i < y_{i'}\right\}\Big] \times \\ \Big\{\sum_{i=n+1}^{n+n'} \mathbb{1}(y_i = 0) \times \sum_{i=n+1}^{n+n'} \mathbb{1}(y_i = 1)\Big\}^{-1}.$$

To construct confidence intervals for estimated prediction performance of obtained classification rule, we use bootstrap techniques again. Generate testing samples indexed by $b = 1, \ldots, B$ and obtain $\widehat{AUOROC}_b$. $(1-\alpha)\%$ confidence intervals can then be constructed using $(\alpha/2)\%$ and $(1-\alpha/2)\%$ percentiles of $\{\widehat{AUOROC}_b\}_{b=1}^{B}$. Finally, we can derive optimal cutoff values as

$$c = \left\{c_1, h_2^{-1}(c_1;\widehat{\theta}), \ldots, h_K^{-1}(c_1;\widehat{\theta})\right\}^{\mathsf{T}},$$

and c_1 is further determined by requirements on false positive rate, true positive rate, or some other measure of loss.

3.7 Simulation Studies

3.7.1 Simulation with Correctly Specified Model

We conduct simulation studies to evaluate finite sample performance of proposed estimator when the model is correctly specified. Specifically we generate i.i.d. (X_i, Y_i) for $i = 1, \ldots, n$ following the relationship

$$Y_i = \mathbb{1}\left\{H(X_i;\theta_0) + U_i > 0\right\},$$

where X_i is a three-dimensional marker vector following normal distribution with mean $(0,0,0)^\mathsf{T}$, marginal variances 10 and covariances 10ρ. We take $H(X;\theta_0) = \min(\theta_{01}X_1 + \theta_{02}, \theta_{03}X_2 + \theta_{04}, X_3)$, where $\theta_0 = (\theta_{01}, \theta_{02}, \theta_{03}, \theta_{04})^\mathsf{T} = (1, -1, 2, 0.5)^\mathsf{T}$, and $U_i \sim N(2, \delta^2)$. Under this data generating scheme, the optimality hypersurface contains a curve that has parameter representation

$$(c_3 + 1, c_3/2 - 1/4, c_3),$$

for $c_3 \in \mathbb{R}$. Also, probability of $Y_i = 1$ is monotone in $H(X_i;\theta_0)$, implying that the ROC curve of random variable $H(X_i;\theta_0)$ corresponds to the optimality ROC curve of tree classifier with positivity region

$$R(c; X, T) = \cap_{k=1}^{3}\{X_k > c_k\},$$

where $c = (c_1, c_2, c_3)^\mathsf{T} \in \mathbb{R}^3$. Various scenarios are considered varying $\delta^2 = 1, 3$, $\rho = 0.2, 0.5, 0.8$ and $n = 50, 100, 200$. We report empirical bias, empirical standard error, empirical mean of standard error estimates, and empirical 95% confidence interval coverage probability of estimator $\widehat{\theta} = (\widehat{\theta}_1, \widehat{\theta}_2, \widehat{\theta}_3, \widehat{\theta}_4)^\mathsf{T}$, calculated over 1,000 replications. All variances were calculated through bootstrap over 10,000 samples.

Simulation results are summarized in Tables 3.1, 3.2 and 3.3. We can see that the estimators are slightly biased; the bootstrapped standard error estimates is close to empirical standard error, and the difference becomes smaller

Table 3.1: Simulation summary statistics for $\widehat{\theta}$ when $\rho = 0.2$

		$\delta^2 = 1$				$\delta^2 = 3$			
		$\widehat{\theta}_1$	$\widehat{\theta}_2$	$\widehat{\theta}_3$	$\widehat{\theta}_4$	$\widehat{\theta}_1$	$\widehat{\theta}_2$	$\widehat{\theta}_3$	$\widehat{\theta}_4$
$n = 50$	Bias	0.061	0.024	0.041	-0.067	0.065	-0.003	0.084	-0.089
	ESE	0.405	0.498	0.435	0.578	0.643	0.728	0.628	0.784
	MSE	0.380	0.448	0.411	0.486	0.534	0.637	0.587	0.671
	CP	0.926	0.903	0.915	0.905	0.910	0.905	0.910	0.908
$n = 100$	Bias	0.075	0.022	0.033	-0.081	0.079	0.006	0.073	-0.119
	ESE	0.373	0.443	0.399	0.485	0.554	0.694	0.593	0.749
	MSE	0.450	0.505	0.447	0.586	0.539	0.693	0.607	0.766
	CP	0.944	0.945	0.940	0.916	0.935	0.938	0.938	0.928
$n = 200$	Bias	0.032	0.022	0.017	-0.020	0.053	-0.048	0.084	-0.045
	ESE	0.289	0.377	0.310	0.396	0.368	0.639	0.468	0.579
	MSE	0.407	0.459	0.394	0.525	0.482	0.664	0.549	0.718
	CP	0.966	0.942	0.963	0.951	0.968	0.931	0.957	0.943

Note: Bias is the empirical bias; ESE is the empirical standard error; MSE is the empirical mean of standard error estimates; CP is the empirical coverage probability of 95% confidence intervals.

as sample size increases; the 95% confidence interval coverage probability converges to 0.95, and is generally close enough to 0.95 when sample size is as large as 200. We would expect the performance of estimators to further improve with even larger sample sizes.

3.7.2 Simulation with Misspecified Model

We conduct another set of simulation studies to evaluate the finite sample bias of estimated AUOROC using proposed methods when the model is misspecified. Specifically we generate bivariate marker $M_i = (M_{i1}, M_{i2})^\mathsf{T}$ associated with binary outcome D_i, i.i.d. for $i = 1, \ldots, n$, where M_i's follow bivariate

Table 3.2: Simulation summary statistics for $\widehat{\theta}$ when $\rho = 0.5$

		$\delta^2 = 1$				$\delta^2 = 3$			
		$\widehat{\theta}_1$	$\widehat{\theta}_2$	$\widehat{\theta}_3$	$\widehat{\theta}_4$	$\widehat{\theta}_1$	$\widehat{\theta}_2$	$\widehat{\theta}_3$	$\widehat{\theta}_4$
$n = 50$	Bias	0.094	0.010	0.016	-0.068	0.060	-0.023	0.051	-0.034
	ESE	0.495	0.527	0.487	0.555	0.622	0.699	0.640	0.737
	MSE	0.372	0.433	0.406	0.465	0.517	0.618	0.579	0.649
	CP	0.899	0.896	0.903	0.901	0.898	0.901	0.913	0.906
$n = 100$	Bias	0.096	0.008	0.034	-0.087	0.074	-0.078	0.141	-0.098
	ESE	0.409	0.491	0.445	0.491	0.626	0.886	0.703	0.763
	MSE	0.485	0.526	0.478	0.612	0.565	0.721	0.641	0.774
	CP	0.940	0.922	0.930	0.939	0.947	0.918	0.936	0.925
$n = 200$	Bias	0.059	0.014	0.031	-0.058	0.089	-0.054	0.121	-0.119
	ESE	0.326	0.393	0.331	0.416	0.541	0.731	0.574	0.746
	MSE	0.461	0.480	0.434	0.573	0.529	0.701	0.601	0.750
	CP	0.962	0.953	0.959	0.951	0.949	0.931	0.955	0.939

Note: Bias is the empirical bias; ESE is the empirical standard error; MSE is the empirical mean of standard error estimates; CP is the empirical coverage probability of 95% confidence intervals.

normal distribution with mean 0, variance 1 and covariance ρ in the subgroup of $D_i = 0$, and follow bivaraite normal distribution with mean μ, variance 0.5 and covariance 0.5ρ when $D_i = 1$. The prevalence of $D_i = 1$ is 0.5. We consider two tree classifiers to combine the markers – the “and" tree that classifies a subject as positive if both marker values exceed some thresholds, and the “or" tree that classifies a subject as positive if either marker value exceeds some threshold. Since there is no closed form solution for the true AUOROC under these scenarios, we calculate the empirical AUOROC, denoted by $\widehat{AUOROC}^{emp}$, using method as described in Section 3.4 in a simulated large dataset with

Table 3.3: Simulation summary statistics for $\widehat{\theta}$ when $\rho = 0.8$

		$\delta^2 = 1$				$\delta^2 = 3$			
		$\widehat{\theta}_1$	$\widehat{\theta}_2$	$\widehat{\theta}_3$	$\widehat{\theta}_4$	$\widehat{\theta}_1$	$\widehat{\theta}_2$	$\widehat{\theta}_3$	$\widehat{\theta}_4$
$n = 50$	Bias	0.075	-0.019	0.029	-0.039	0.051	-0.040	0.038	0.005
	ESE	0.491	0.505	0.510	0.652	0.559	0.688	0.634	0.740
	MSE	0.348	0.401	0.388	0.416	0.497	0.574	0.562	0.604
	CP	0.900	0.905	0.893	0.893	0.868	0.874	0.883	0.890
$n = 100$	Bias	0.109	-0.035	0.067	-0.089	0.016	-0.018	0.097	-0.056
	ESE	0.512	0.529	0.501	0.625	0.617	0.834	0.706	0.833
	MSE	0.521	0.551	0.529	0.636	0.581	0.716	0.664	0.770
	CP	0.927	0.927	0.931	0.916	0.914	0.913	0.924	0.933
$n = 200$	Bias	0.098	-0.052	0.079	-0.073	0.090	-0.116	0.224	-0.169
	ESE	0.396	0.609	0.537	0.489	0.737	1.039	0.887	1.041
	MSE	0.641	0.608	0.585	0.763	0.608	0.795	0.709	0.819
	CP	0.969	0.956	0.962	0.953	0.935	0.930	0.940	0.937

Note: Bias is the empirical bias; ESE is the empirical standard error; MSE is the empirical mean of standard error estimates; CP is the empirical coverage probability of 95% confidence intervals.

sample size 5,000. Due to the consistency of the empirical AUOROC estimator, simulated AUOROC is numerically close enough to the true value, but computational difficulties in calculating $\widehat{AUOROC}^{emp}$ as discussed in Section 3.4 limited our simulation studies to only two markers. We then use proposed rank-based methods to search for the optimal splitting criteria using piece-wise linear function as h_k's. Knots of piece-wise linear functions are chosen to be evenly distributed over the range of marker, and the numbers of knots are selected from 0 to 5 according to 5-fold cross validations. We choose the class of piece-wise linear functions for illustration because it is one of the most generic and simple approximations for continuous functions.

Table 3.4: Summary statistics for simulation when model is misspecified

	$\mu = 0.5$			$\mu = 1$		
	$\rho = 0$	$\rho = 0.3$	$\rho = 0.7$	$\rho = 0$	$\rho = 0.3$	$\rho = 0.7$
				"and" tree		
$\widehat{AUOROC}^{emp}$	0.744	0.723	0.691	0.876	0.855	0.824
$n = 50$	0.667	0.635	0.605	0.816	0.779	0.741
$n = 100$	0.664	0.635	0.611	0.821	0.780	0.737
$n = 200$	0.660	0.625	0.596	0.824	0.781	0.747
				"or" tree		
$\widehat{AUOROC}^{emp}$	0.663	0.652	0.650	0.803	0.789	0.787
$n = 50$	0.645	0.641	0.631	0.792	0.787	0.776
$n = 100$	0.648	0.647	0.643	0.793	0.790	0.782
$n = 200$	0.649	0.649	0.650	0.794	0.790	0.788

The prediction performance of the estimated rule is then evaluated using an independently generated large testing set of sample size 5,000. Again, this sample size is large enough so that we approximately obtain the true prediction power of the estimated rule. We consider different scenarios varying $\mu = 0.5, 1$, $\rho = 0, 0.3, 0.7$ and training sample size $n = 50, 100, 200$, and report the empirical average of approximated AUOROC estimated using methods in Section 3.5 over 1,000 replication. Since it is more appropriate to consider the expectation of a logistic-transformed AUOROC that ranges over the entire real line, we calculate the empirical average as

$$\text{expit}\Big\{\frac{1}{1000}\sum_{k=1}^{1000}\text{logit}(\widehat{AUOROC}_k)\Big\},$$

where $\widehat{AUOROC}_k$ is the approximated AUOROC for the kth simulation replication, $\text{expit}(x) = e^x/(1+e^x)$ and $\text{logit}(x) = \log\{x/(1-x)\}$.

Simulation results are summarized in Table 3.4, showing that the bias of estimated AUOROC is reasonably small in all cases investigated, especially for the “or" tree structure. For the “and" tree structure, we expect the biases to be further improved by using another function class $\mathcal{H}$ that better fits the data.

3.8 Data Analysis: Biomarker Prediction Performance for 5-Year Progression using BIOCARD Study Data

The BIOCARD (BIOCARD: Biomarkers of Cognitive Decline Among Normal Individuals) study is an active longitudinal study that follows a cohort of 349 initially cognitively normal individuals for progression of mild cognitive impairment (MCI) and dementia related to Alzheimer’s Disease, and collects annual cognitive and biannual MRI and CSF scans and blood specimens since study initiation in 1995 at National Institute of Health until 2005 and after reinitiation at Johns Hopkins in 2009. The overarching goal of the study is to identify predictors of cognitive decline among normal individuals.

Using BIOCARD study data, we illustrate proposed methods by evaluating predictive performance of several tree classifiers combining markers collected at baseline for the binary outcome of whether an individual progressed to MCI within 5 years. Markers from four domains are considered: baseline age and ApoE-4 status from the demographic domain, Digital Symbol Substitution Test and Wechsler Adult Scale from the cognitive test domain, right hippocampus volume and right entorhinal cortex thickness from the MRI domain, Abeta and P-Tau from the CSF domain. These markers are selected because they are indicated in previous analyses to be predictive of progression from normal to MCI or dementia (Moghekar et al., 2013, Albert et al., 2014, Soldan et al., 2015) or are marginally associated with our binary outcome of interest. Analysis results of this type can potentially be used in clinical trials studying treatment of Alzheimer's Disease to recruit subjects that are at higher risk of MCI in the near future, and thus improving efficacy of statistical analysis and reducing cost of study by targeting people at greater risk of progression.

We investigate three sets of tree structures combining available markers. Due to the constraints on model complexity imposed by the small sample size and especially the small number of cognitive impairment cases, we study tree classifiers combining no more than three markers or markers coming from no more than two regions. For the first set of tree classifiers, we follow the general philosophy of diagnosing cognitive impairment in practice that uses the "or" operation to combine markers within the same domain, and the "and" operation to combine domains. For the second set, we consider the alternative logic that combines markers using "and" within domain and combines domains using "or". For the third set of tree classifiers, we consider combinations of top correlated markers with the outcome, under both the general philosophy and the alternative in treating domains. See Tables 3.5, 3.6 and 3.7 for a detailed description of tree structures studied.

Out of 224 subjects with available marker information, we include in the analyses 218 subjects that are observed to have either progressed within 5 years from baseline, which is referred to as positive, or remained cognitively

Table 3.5: Descriptions of tree classifiers in the first set

Tree classifier name	Description of positivity criteria
Demo	Subject passes some cutoff age or is an ApoE-4 carrier.
Cog	Subject's Digital Symbol Substitution Test or Wechsler Adult Scale are below some thresholds.
MRI	Subject's right hippocampus volume or right entorhinal cortex thickess are below some thresholds.
CSF	Subject's Abeta falls below some threshold or P-Tau passes some threshold.
Demo and Cog	Subject is classified as positive by both the Demo and the Cog tree.
Demo and MRI	Subject is classified as positive by both the Demo and the MRI tree.
Demo and CSF	Subject is classified as positive by both the Demo and the CSF tree.
Cog and MRI	Subject is classified as positive by both the Cog and the MRI tree.
Cog and CSF	Subject is classified as positive by both the Cog and the CSF tree.
MRI and CSF	Subject is classified as positive by both the MRI and the CSF tree.

Table 3.6: Descriptions of tree classifiers in the second set

Tree classifier name	Description of positivity criteria
Demo*	Subject passes some cutoff age and is an ApoE-4 carrier.
Cog*	Subject's Digital Symbol Substitution Test and Wechsler Adult Scale are both below some thresholds.
MRI*	Subject's right hippocampus volume and right entorhinal cortex thickess are both below some thresholds.
CSF*	Subject's Abeta falls below some threshold and P-Tau passes some threshold.
Demo* or Cog*	Subject is classified as positive by the Demo* or the Cog* tree.
Demo* or MRI*	Subject is classified as positive by the Demo* or the MRI* tree.
Demo* or CSF*	Subject is classified as positive by the Demo* or the CSF* tree.
Cog* or MRI*	Subject is classified as positive by the Cog* or the MRI* tree.
Cog* or CSF*	Subject is classified as positive by the Cog* or the CSF* tree.
MRI* or CSF*	Subject is classified as positive by the MRI* or the CSF* tree.

Table 3.7: Descriptions of tree classifiers in the third set

Tree classifier name	Description of positivity criteria
Top1	Subject's Digital Symbol Substitution Test is below some threshold.
Top2	Subject's Digital Symbol Substitution Test is below some threshold, and subject's P-Tau passes some threshold.
Top3	Subject's Digital Symbol Substitution Test is below some threshold, and subject's P-Tau passes some threshold or Abeta falls below some threshold.
Top2*	Subject's Digital Symbol Substitution Test is below some threshold, or subject's P-Tau passes some threshold.
Top3*	Subject's Digital Symbol Substitution Test is below some threshold, or subject's P-Tau passes some threshold and Abeta falls below some threshold.

normal beyond 5 years, which is referred to as negative (the reduced risk set, Kaplan and Meier, 1958). The dataset is equally split into a training set on which the optimal rule is estimated, and a testing set on which the prediction performance of estimated rule is evaluated. For estimating the optimal rule, function h_k's are taken to be piece-wise linear functions with knots evenly spread over the range of corresponding covariate, and the numbers of knots are selected from 0 to 2 by five-fold cross validations. We choose to use piecewise linear functions because it is the one of the most generic approximation of continuous functions and has proven to yield good approximations of AUOROC under many scenarios in our simulation study in Section 3.7.2.

We choose covariate with the strongest marginal Kendall's tau correlation with outcome as the “reference" that corresponds to index $k = 1$. The logic behind this choice is that there always exists some parameter value $\widetilde{\theta}$ such that markers other than the reference marker are effectively not contributing to classification, which means that tree structures including more markers always perform better than the reference marker in the population. After obtaining the estimates $\widehat{\theta}$ and thus the optimal rule for given tree classifier, we evaluate $\widehat{AUOROC}$, that is the estimated AUC of $H(X;\widehat{\theta})$ in association with outcome, as the measure of prediction performance. 95% confidence intervals of $\widehat{AUOROC}$ given estimated optimal rules are obtained by bootstrapping on the testing set over 10,000 samples. We also report sensitivities and specificities maximizing Youden’s Index (Youden, 1950) as an illustration of one posssible way to utilize the analysis results in practice – having chosen a cutoff value, the complete classification rule that yields correponding sensitivity and specificity can be constructed. See Table 3.8 for a summary of these analysis results. ROC curves of the composite marker variables $H(X;\widehat{\theta})$ are given in Figures 3.3, 3.4, and 3.5. P-values comparing the predictive powers of tree classifiers are plotted in the form a heatmap in Figure 3.6, where a darker color indicates a smaller p-value. We can see that CSF and cognitive are the most predictive domains using the “or" combination within domain, and they often improves prediction on top of another domain. The MRI domain on the other side, is the least predictive using either the “and" or “or" combination within domain, and often adds more noise than predictive power on top of another domain. Of note, analysis results indicate serious overfitting issues when more than two markers are used, mostly due to the small number of cases in the BIOCARD dataset (10 in training sample, and 8 in testing sample). While allowing more

Table 3.8: Summary statistics of BIOCARD analysis tree evaluation results

Tree classifier	$\widehat{AUOROC}$	95% CI	Sensitivity	Specificity
Demo	0.710	(0.514, 0.902)	0.625	0.810
Cog	0.759	(0.527, 0.942)	0.875	0.640
MRI	0.646	(0.417, 0.844)	0.750	0.660
CSF	0.813	(0.569, 0.968)	0.875	0.750
Demo and Cog	0.731	(0.523, 0.894)	0.875	0.620
Demo and MRI	0.654	(0.481, 0.814)	0.875	0.380
Demo and CSF	0.652	(0.392, 0.880)	0.625	0.810
Cog and MRI	0.747	(0.526, 0.924)	0.875	0.590
Cog and CSF	0.788	(0.577, 0.949)	0.875	0.740
MRI and CSF	0.653	(0.417, 0.844)	0.375	0.930
Demo*	0.625	(0.507, 0.904)	0.500	0.940
Cog*	0.749	(0.522, 0.924)	0.875	0.640
MRI*	0.645	(0.400, 0.862)	0.625	0.730
CSF*	0.746	(0.581, 0.899)	0.875	0.540
Demo* or Cog*	0.873	(0.753, 0.970)	0.750	0.870
Demo* or MRI*	0.736	(0.504, 0.914)	0.625	0.860
Demo* or CSF*	0.719	(0.661, 0.908)	0.625	0.910
Cog* or MRI*	0.691	(0.554, 0.816)	0.875	0.500
Cog* or CSF*	0.801	(0.626, 0.944)	0.750	0.840
MRI* or CSF*	0.741	(0.571, 0.894)	0.750	0.630
Top1	0.759	(0.538, 0.925)	0.875	0.640
Top2	0.749	(0.531, 0.915)	0.875	0.740
Top3	0.787	(0.570, 0.948)	0.875	0.740
Top2*	0.751	(0.521, 0.938)	0.875	0.620
Top3*	0.734	(0.524, 0.906)	0.875	0.600

flexibility, our proposed rank-based method does require sufficient sample size to be effective.

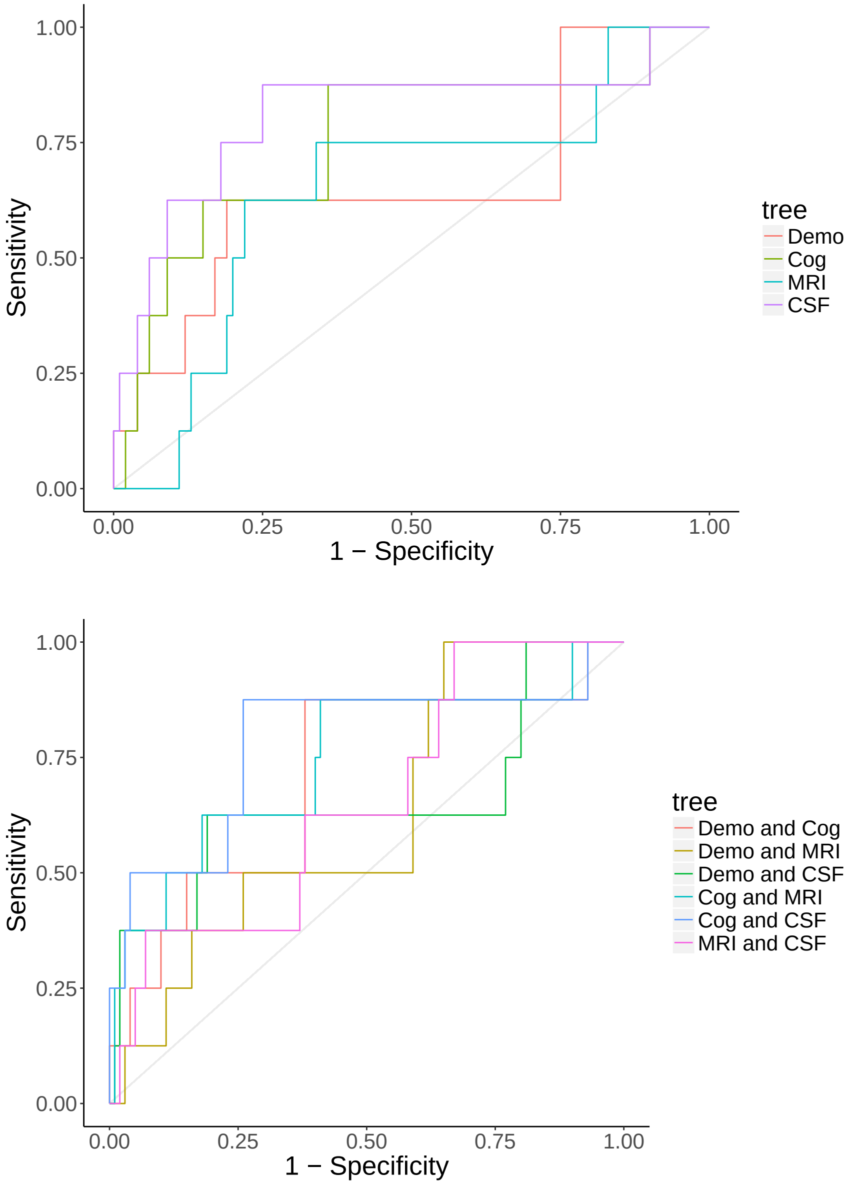

Figure 3.3: ROC curves of composite marker variable corresponding to tree classifiers in the first set.

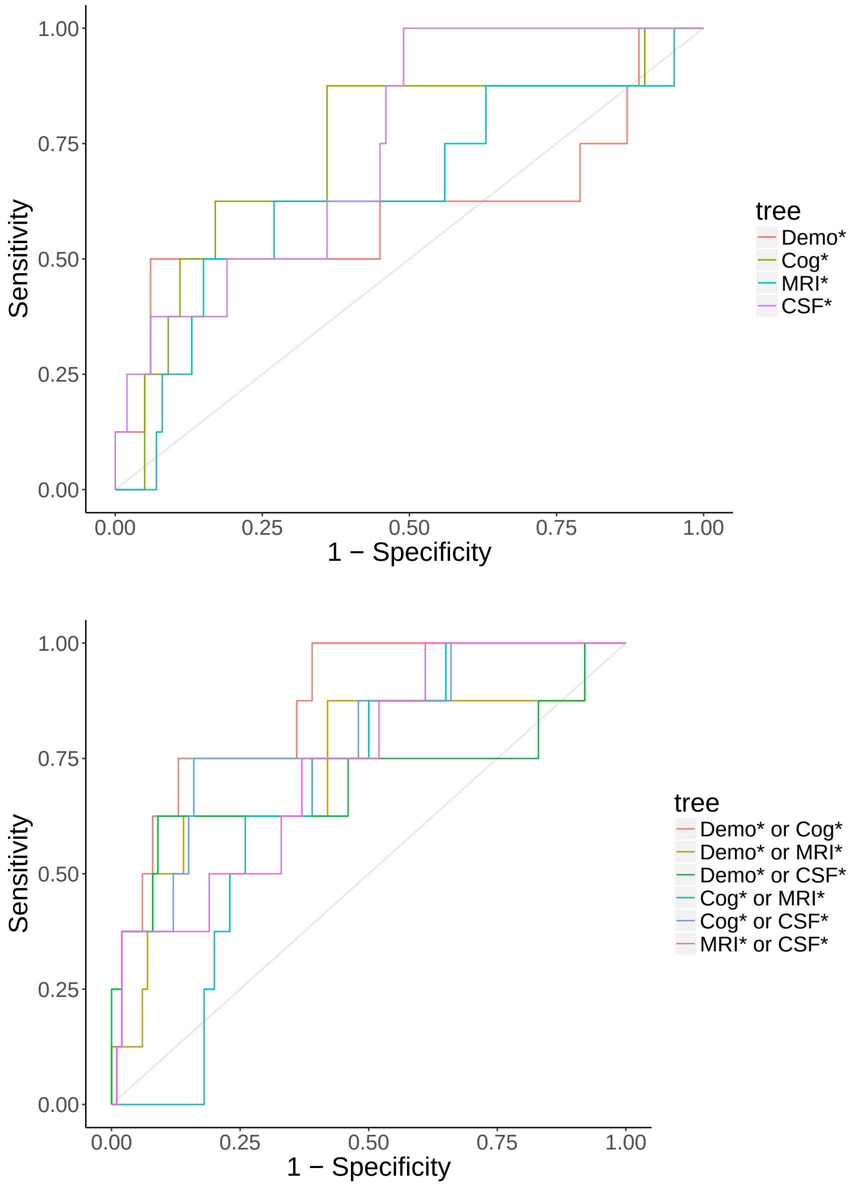

Figure 3.4: ROC curves of composite marker variable corresponding to tree classifiers in the second set.

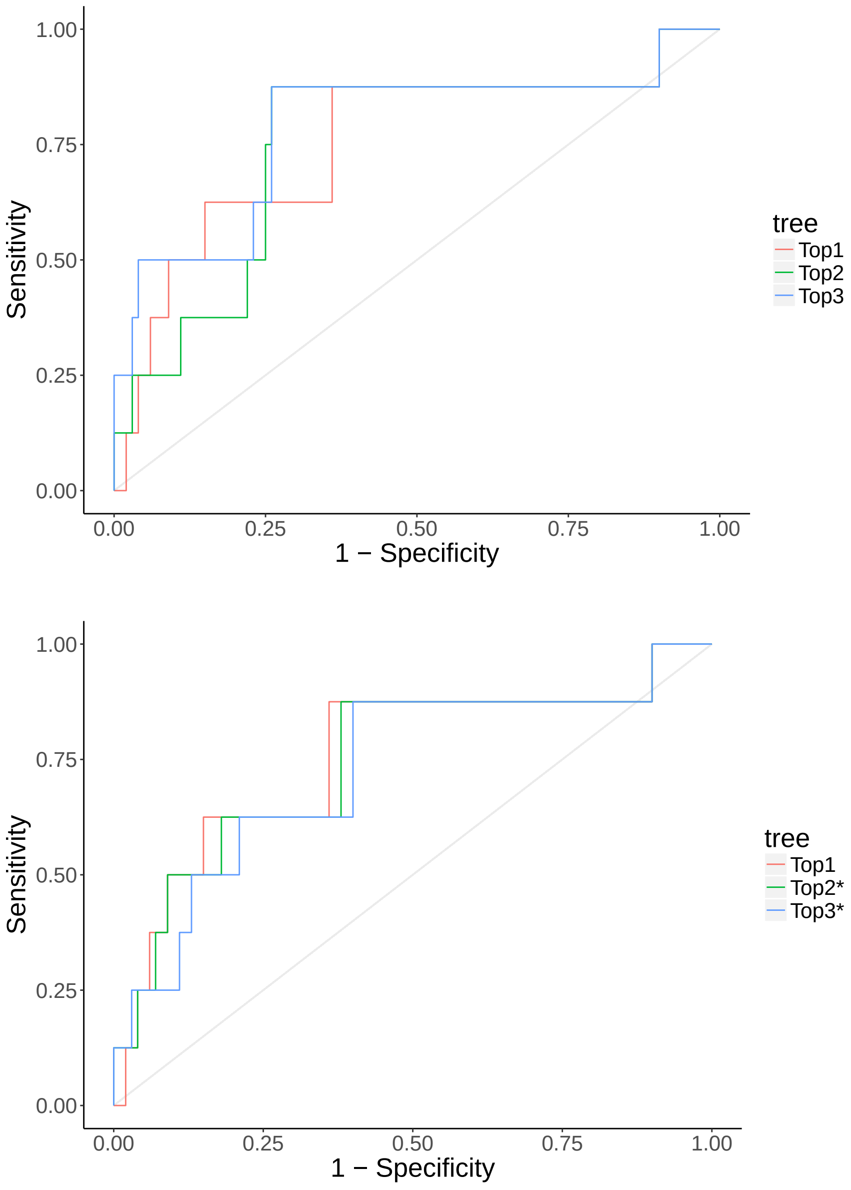

Figure 3.5: ROC curves of composite marker variable corresponding to tree classifiers in the third set.

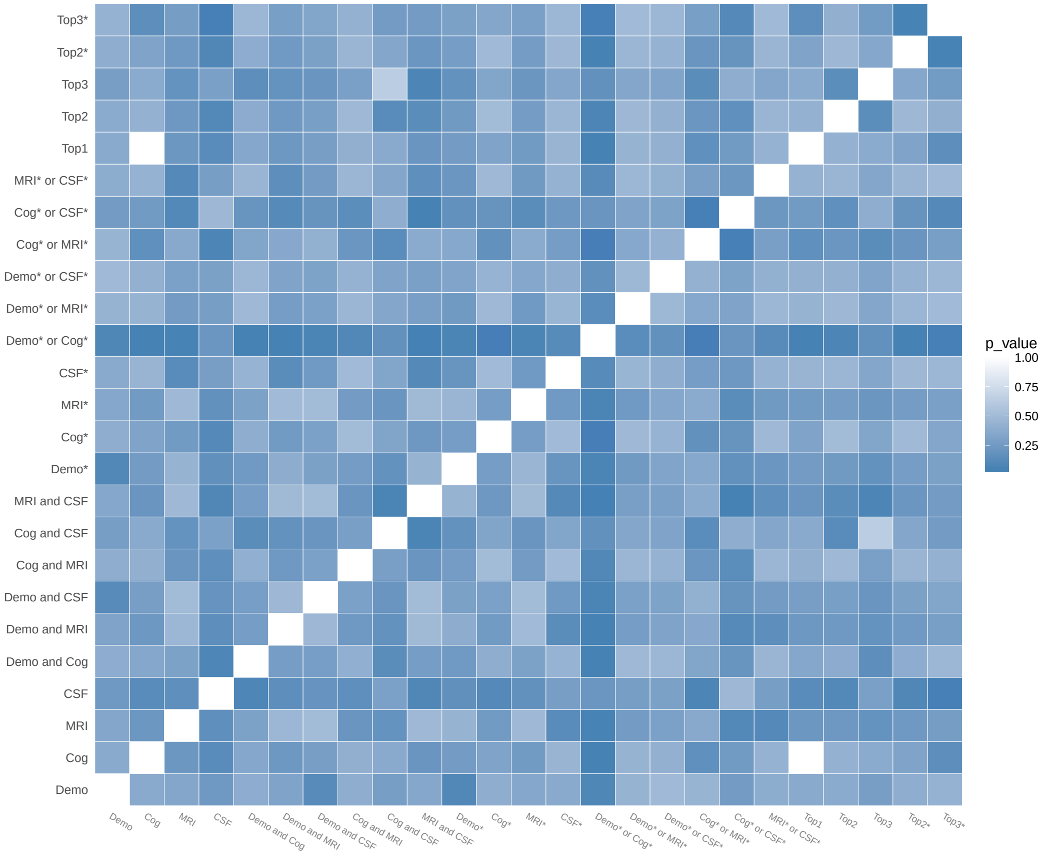

Figure 3.6: Heatmap of p-values from pair-wise comparison of all tree structures under consideration in their predictive power depicted by $\widehat{AUOROC}$. A darker color indicates a smaller p-value.

3.9 Proofs of Asymptotic Results

3.9.1 Proof of Theorem 1

Denote by F_0 the distribution function of (X, Y). Let $A(F)$ be the corresponding $OROC(\cdot)$ of a general distribution function F, where A is some functional. Also let $T(F) = \sup_{t \in [0,1]} \left| A(F)(t) - A(F_0)(t) \right|$. By continuity, we have that if a sequence of distribution functions $\{F_n\}_{n=1}^{\infty}$ converges uniformly to F_0, then $A(F_n)$ converges to $A(F_0)$ point-wise. Now we further show that the convergence is uniform. By monotonicity of $OROC(\cdot)$, for any $\delta > 0$, there exsits a sequence $0 = t_0 < t_1 < \cdots < t_L = 1$ such that

$$\left| A(F_0)(q_l) - A(F_0)(q_{l-1}) \right| < \delta,$$

for $l = 1, \ldots, L$. This implies that for any $t \in [t_{l-1}, t_l]$ we have

$$\begin{aligned}
&\left| A(F_n)(t) - A(F_0)(t) \right| \\
\leq &\max \left\{ \left| A(F_n)(q_l) - A(F_0)(t) \right|, \left| A(F_n)(t_{l-1}) - A(F_0)(t) \right| \right\} \\
\leq &\max \left\{ \left| A(F_n)(q_l) - A(F_0)(q_l) \right|, \left| A(F_n)(q_{l-1}) - A(F_0)(q_{l-1}) \right| \right\} + \delta.
\end{aligned}$$

Therefore we have

$$\sup_{t \in [0,1]} \left| A(F_n)(t) - A(F_0)(t) \right| \leq \max_{l=0,\ldots,L} \left| A(F_n)(q_l) - A(F_0)(q_l) \right| + \delta.$$

Let $n \to \infty$ in the last display, we show that

$$\lim_{n \to \infty} \sup_{t \in [0,1]} \left| A(F_n)(t) - A(F_0)(t) \right| \leq \delta.$$

Further let $\delta \to 0$, and we prove the uniform convergence of $T(F_n) \to T(F_0)$. Finally because empirical estimators of F_0 converges uniformly and almost surely (Van Der Vaart and Wellner, 1996), we prove results of Theorem 1.

3.9.2 Proof of Theorem 2

To establish asymptotic properties of proposed estimator $\widehat{\theta}$, we impose the following regularity assumptions.

Assumption 1. Vector θ_0 is an interior point of compact set $\Theta \in \mathbb{R}^p$ that indexes a series of continuous and monotonically increasing functions $h_k(\cdot;\theta_0)$ for $k = 1, \ldots, K$.

Assumption 2. We assume that S_X, the support of X, is an open set in a possibly lower rank subspace of $\mathbb{R}^K$.

Assumption 3. Let $\mathcal{N}$ denote a neighborhood of θ_0. We then make the following assumptions:

(i) For each $z \in S_Z$, all second order partial derivatives of $\tau(z;\theta)$ exist on $\mathcal{N}$;

(ii) There exists an integrable function $M(z)$ such that for all $z \in S_Z$ and $\theta \in \mathcal{N}$, we have

$$\left\|\nabla_2\tau(z;\theta) - \nabla_2(\tau;\theta_0)\right\| \leq M(z) \cdot \|\theta - \theta_0\|;$$

(iii) $\mathbb{E}_Z\left[\|\nabla_1\tau(Z;\theta_0)\|_1^2\right]$ and $\mathbb{E}_Z\left|\nabla_2\middle|\tau(Z;\theta_0)\right|$ are finitely upper bounded;

(iv) Matrix $\mathbb{E}_Z\left[\nabla_2\tau(Z;\theta_0)\right]$ is negative definite.

Assumption 4. $h_k(\cdot;\theta)$ comes from a finite dimensional vector space of functions for $k = 1, \ldots, K$.

Under Assumption 3, θ_0 uniquely maximizes $S(\theta)$ locally in a neighborhood. Following arguments in Han, 1987 or by uniform strong convergence of U-statistics, there exists a sequence of solutions $\widehat{\theta}$ that that converge almost surely to θ_0. Next, we prove asymptotic normality.

Consider a class of functions $\mathcal{G} = \{g(\cdot, \cdot; \theta) : \theta \in \Theta\}$, where

$$g(z_1, z_2; \theta) = \mathbb{1}\{H(x_1; \theta) > H(x_2; \theta), y_1 > y_2\}$$

for $z_1, z_2 \in S_Z$. Given consistency, it is sufficient to show that the set of subgraphs of functions belonging to $\mathcal{G}$ forms a VC class of sets. For each $\theta \in \Theta$ we have that

$$\begin{aligned}
&\text{subgraph}\{g(\cdot, \cdot; \theta)\} \\
=&\{(z_1, z_2, t) \in S_Z \otimes S_Z \otimes \mathbb{R} : 0 < t < g(z_1, z_2; \theta)\} \\
=&\{y_1 - y_2 > 0\} \cap \{t \geq 1\}^c \cap \{t > 0\} \cap \\
&\cup_{j_1=1}^{J} \cap_{j_2=1}^{J} \cap_{k_1 \in \kappa_{j_1}} \cap_{k_2 \in \kappa_{j_2}} \{h_{k_1}(x_{1,k_1}; \theta) > h_{k_2}(x_{2,k_2}; \theta) > 0\}.
\end{aligned}$$

By Assumption 4 and Lemmas 2.4 and 2.5 in Pakes and Pollard, 1989, $\{\text{subgraph}(g); g \in \mathcal{G}\}$ forms a VC class of sets, thus proving Theorem 2.

References

McIntosh, Martin W and Margaret Sullivan Pepe (2002). "Combining several screening tests: optimality of the risk score". In: *Biometrics* 58.3, pp. 657–664.

Pepe, Margaret Sullivan and Mary Lou Thompson (2000). "Combining diagnostic test results to increase accuracy". In: *Biostatistics* 1.2, pp. 123–140.

Pepe, Margaret Sullivan, Tianxi Cai, and Gary Longton (2006). "Combining predictors for classification using the area under the receiver operating characteristic curve". In: *Biometrics* 62.1, pp. 221–229.

Baker, Stuart G (2000). "Identifying combinations of cancer markers for further study as triggers of early intervention". In: *Biometrics* 56.4, pp. 1082–1087.

Jin, Hua and Ying Lu (2009). "The ROC region of a regression tree". In: *Statistics & Probability Letters* 79.7, pp. 936–942.

Wang, Mei-Cheng and Shanshan Li (2012). "Bivariate marker measurements and ROC analysis". In: *Biometrics* 68.4, pp. 1207–1218.

Wang, Mei-Cheng and Shanshan Li (2013). "ROC analysis for multiple markers with tree-based classification". In: *Lifetime data analysis* 19.2, pp. 257–277.

Hanley, James A and Barbara J McNeil (1982). “The meaning and use of the area under a receiver operating characteristic (ROC) curve.” In: *Radiology* 143.1, pp. 29–36.

Han, Aaron K (1987). “Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator”. In: *Journal of Econometrics* 35.2-3, pp. 303–316.

Sherman, Robert P (1993). “The limiting distribution of the maximum rank correlation estimator”. In: *Econometrica: Journal of the Econometric Society*, pp. 123–137.

Efron, Bradley and Robert J Tibshirani (1994). *An introduction to the bootstrap*. CRC press.

Moghekar, Abhay, Shanshan Li, Yi Lu, Ming Li, Mei-Cheng Wang, Marilyn Albert, Richard O’Brien, BIOCARD Research Team, et al. (2013). “CSF biomarker changes precede symptom onset of mild cognitive impairment”. In: *Neurology* 81.20, pp. 1753–1758.

Albert, Marilyn, Anja Soldan, Rebecca Gottesman, Guy McKhann, Ned Sacktor, Leonie Farrington, Maura Grega, Raymond Turner, Yi Lu, Shanshan Li, et al. (2014). “Cognitive changes preceding clinical symptom onset of mild cognitive impairment and relationship to ApoE genotype”. In: *Current Alzheimer Research* 11.8, pp. 773–784.

Soldan, Anja, Corinne Pettigrew, Yi Lu, Mei-Cheng Wang, Ola Selnes, Marilyn Albert, Timothy Brown, J Tilak Ratnanather, Laurent Younes, Michael I Miller, et al. (2015). “Relationship of medial temporal lobe atrophy, APOE genotype, and cognitive reserve in preclinical Alzheimer’s disease”. In: *Human brain mapping* 36.7, pp. 2826–2841.

Kaplan, Edward L and Paul Meier (1958). “Nonparametric estimation from incomplete observations”. In: *Journal of the American statistical association* 53.282, pp. 457–481.

Youden, William J (1950). “Index for rating diagnostic tests”. In: *Cancer* 3.1, pp. 32–35.

Van Der Vaart, Aad W and Jon A Wellner (1996). “Weak convergence”. In: *Weak convergence and empirical processes*. Springer, pp. 16–28.

Pakes, Ariel and David Pollard (1989). “Simulation and the asymptotics of optimization estimators”. In: *Econometrica: Journal of the Econometric Society*, pp. 1027–1057.

Chapter 4

Semi-Marginal and Semi-Parametric Analysis for Recurrent Gap Time with Time-Dependent Covariates

4.1 Introduction

In this chapter, we study gap times between successive recurrent events of the same type, a data structure that is often of scientific interest in applications such as hospitalizations, HIV opportunistic infections, and menstrual cycle studies. To facilitate our discussion, we introduce the following notations. Let i be the index of a subject, and j be the index of an event. For subject i, let T_{ij} be the gap time from event $j-1$ to event j, $j = 1, 2, \ldots$, and without loss of generality, let $T_{i0} = 0$ represent the time origin. Using the notation of gap times, we may assume the recurrent event process starts from $T_{i0} = 0$, and a study subject experiences recurrent events of interest at times $T_{i0}, T_{i0} + T_{i1}, \ldots$, since time origin, and remains under observation until occurrence of censoring at time C_i. We consider the regression of gap time T_{ij}, and let $X_{ij} \in \mathbb{R}^p$ be covariates corresponding to T_{ij}. We allow X_{ij} to be either baseline, or time-dependent, and are thus able to include time trends into covariate information.

In real data applications, it is often true that there exist heterogeneity among subjects and correlation among gap times within subject. Not only does this correlation structure need to be taken into account when modeling, it also induces informative censoring on gap times except the first one. To see this, let $N_i(t) = \sum_{j=1}^{\infty} \mathbb{1}(T_{ij} \leq t)$, $t > 0$, be the point process associated with recurrent events for subject i. Consider the simple case when X_{ij} is time-invariant, that is, $X_{ij} = X_i$ for all j, and when censoring C_i is conditional independent of process $N_i(\cdot)$, given X_i. Each gap time T_{ij} is then censored by $C_{ij} = \min\left\{C_i - \sum_{k=0}^{j-1} T_{ik}, 0\right\}$, which is in general correlated with T_{ij} for $j \neq 0$ even after conditioning on X_i. To deal with induced informative censoring, a frailty model was proposed by Pena, Strawderman, and Hollander, 2001 to account for correlations within subject, the idea of which can be traced back to Aalen and Husebye, 1991. Focusing on time to event instead, Prentice, Williams, and Peterson, 1981 proposed full conditional model for the recurrent event process. Compared with frailty models that base analyses on unverifiable assumptions and full conditional models, marginal models are more robust to subject-level correlation structure, and are useful when researchers are interested in population-level effect of covariates. Due to these advantages, a semi-marginal model is adopted in this work, in the sense that the model is marginal and leaves the correlation between gap times within the same subject completely unspecified, except for the conditioning on the location of last recurrent event. To be more specific, let $L_{ij} = \sum_{k=0}^{j-1} T_{ik}$ for $j = 1, 2, \ldots,$ represent the elapse from time origin to the $(j-1)$th event, a variable that captures the “location" of a subject in the progression of the recurrent event process. After the inclusion of L_{ij} as part of the covariates in the gap time regression model, independent censoring is obtained given covariates. Compared with other marginal regression models such as that proposed in Huang and Chen, 2003, our method do not require the assumption on exchangeability of gap times within subject, and is therefore more suitable for modeling time trend and the effect of time-varying covariates.

4.2 Semi-Marginal Regression Model

Denote by $D_{ij}^0 = (T_{ij}, L_{ij}, X_{ij})$ the variables measured at jth gap time for subject i, and let $D_i^0 = (D_{i1}^0, D_{i2}^0, \ldots)$ be the multivariate recurrent event process for subject i. Suppose (D_i^0, C_i)'s follow some i.i.d. distribution for $i = 1, \ldots, n$. For simplicity of writing, we denote $Z_{ij} = \left\{X_{ij}^{\mathsf{T}}, \phi_i(L_{ij})\right\}^{\mathsf{T}}$ to be a set of covariates including last event location information, where $\phi_i(\cdot)$ is a pre-specified transformation function allowed to be subject-specific and vector-valued. It is assumed that T_{ij} is correlated with L_{ij} on the transformed scale. Denoting $\lambda_{ij}(t; X_{ij}, L_{ij})$ to be the hazard of T_{ij} given (X_{ij}, L_{ij}), we propose the following semi-marginal model:

$$\text{(M1)} \qquad \lambda_{ij}(t; X_{ij}, L_{ij}) = \lambda_0(t) e^{Z_{ij}^{\mathsf{T}} \theta_0},$$

where θ_0 is regression coefficient, and $\lambda_0(\cdot)$ is a non-negative baseline hazard function shared by gap times within the same subject. Included in the regression covariates are covariates that can be baseline or specific to each recurrent event, and specially, L_{ij}, a variable that indicates the location of an occurrence in the recurrent event process of a subject. Model (M1) is semi-marginal in the sense that model assumptions are made not conditional on full history, or given underlying frailty that captures subject heterogeneity, but only conditional on the location of the last observed event. The location variable is incorporated to resolve the induced dependency between C_{ij} and T_{ij}, and allows the model to capture time trend.

Since the proposed model is semi-marginal, the existence of a sensible full model which guarantees the validity of the semi-parametric model deserves further examination. The existence of a full model for marginal proportional hazard model to hold turns out to be a non-trivial point, as the common practice of using multiplicative random effect on hazard may not work. That is, if we adopt the widely used frailty model

$$\lambda(t; X_{ij}, L_{ij}, W_i) = W_i \lambda_0(t) e^{Z_{ij}^{\mathsf{T}}\theta_0}, \tag{4.1}$$

then there is no non-degenerate W_i that yields desired model (M1). To see this, first observe that (M1) can be expressed as a semi-transformation model

$$g(T_{ij}) = -Z_{ij}^{\mathsf{T}}\theta_0 + E_{ij}, \tag{4.2}$$

where $g(t) = \log \Lambda_0(t) = \log \int_0^t \lambda_0(u)\, du$ is the log transformed cumulative baseline hazard function, and E_{ij} has extreme density $f(x) = e^{x-e^x}$, conditional on Z_{ij}. We require that E_{ij}'s be independent across different i's, but possibly correlated within the same i. For (4.1) to yield (M1), it is then required that $E_{ij} - \log W_i$ have extreme density $f(x)$. This condition is satisfied only when W_i is degenerated to 1, that is, $W_i = 1$, which does not allow correlation between gap times within the same subject.

To generate data from a non-degenerate full model, the error terms need to be generated sequentially. Specifically, we start from generating independent error terms E_{i1} and covariate $Z_{i1} = \{X_{i1}^{\mathsf{T}}, \phi_i(0)\}^{\mathsf{T}}$ for $i = 1, \ldots, n$, and get $T_{i1} = g^{-1}(-Z_{i1}^{\mathsf{T}}\theta_0 + E_{i1})$. Given generated data $(Z_{i1}, E_{i1}, T_{i1}), \ldots, (Z_{ij}, E_{ij}, T_{ij})$, we obtain $L_{i,j+1} = L_{ij} + T_{ij}$, and generate independent $X_{i,j+1}$, and then $E_{i,j+1}$ such that it is independent of $L_{i,j+1}$ and $X_{i,j+1}$, but, for instance, correlated with E_{ij}. This is in a sense an autocorrelation 1 structure of error terms and can be replaced by other structures.

4.3 Estimation Based on Weighted Pairwise Comparison

We first describe the observed data under consideration. Write $\Delta_{ij} = \mathbb{1}(T_{ij} \leq C_{ij})$ to be the censoring indicator of gap time T_{ij}, and denote by Y_{ij} the minimum of gap time T_{ij} and censoring time C_{ij}. Define $J_i = \max\left\{j : \sum_{k=1}^{j-1} T_{ij} < C_i\right\}$, such that T_{i,J_i-1} is the last uncensored gap time for subject i. We note that for $j > J_i$ all gap times are censored by 0 yielding $Y_{ij} = C_{ij} = 0$, and Z_{ij} are actually unobserved and can be denoted to have value zero without loss of generality. Denote the observed data for subject i at jth occurrence of recurrent event by $D_{ij} = (Y_{ij}, L_{ij}, X_{ij}, \Delta_{ij})$, and the collection of events for subject i by $D_i = (D_{i1}, D_{i2}, \dots)$. In this paper we consider continuous T_{ij} and assume the existence of upper limit $\tau > 0$ such that $\mathbb{P}(C_{ij} \geq \tau) > 0$ for $j = 1, \dots, J_i$. We consider the conditional indepdent right censoring mechanism under which C_i is indepdendent of T_{ij} conditional on X_{ij}.

We formulate the problem from a point process perspective, and consider the observed gap time process $N_{ij}(t) = \mathbb{1}(Y_{ij} \leq t)\, \Delta_{ij}$. Also denote by $R_{ij}(t) = \mathbb{1}(Y_{ij} \geq t)$ the at-risk process. Define σ-filed for the jth event of subject i at time t as

$$\mathcal{F}_{ij,t} = \sigma\big\{ \mathbb{1}(Y_{ij} \leq u, \Delta_{ij} = 1), \mathbb{1}(Y_{ij} \leq t, \Delta_{ij} = 0), X_{ij}, L_{ij} : 0 \leq u \leq t \big\}.$$

Based on model and censoring assumptions, we have for $t > 0$,

$$\mathbb{E}\big\{ \mathrm{d}N_{ij}(t) \big| C_{ij} > 0, \mathcal{F}_{ij,t-} \big\} = R_{ij}(t) e^{Z_{ij}^{\mathsf{T}} \theta_0}\, \mathrm{d}\Lambda_0(t).$$

Denote $M_{ij}(t) = N_{ij}(t) - R_{ij}(t)\Lambda_0(t) e^{Z_{ij}^{\mathsf{T}}\theta_0}$, and observe that $\mathbb{E}\big\{dM_{ij}(t) | j \leq J_i, \mathcal{F}_{ij,t-}\big\} = 0$ and that $M_{ij}(t) \equiv 0$ for $j > J_i$. Now consider pairwise comparison function

$$\begin{aligned} h(D_{ij}, D_{i'j'}; \theta) =& Q(Z_{ij}, Z_{i'j'}) \int_0^\tau R_{ij}(t) Z_{ij}\, dN_{ij}(t) + R_{i'j'}(t) Z_{i'j'}\, dN_{i'j'}(t) - \\ & \frac{R_{ij}(t) Z_{ij} e^{Z_{ij}^{\mathsf{T}}\theta} + R_{i'j'}(t) Z_{i'j'} e^{Z_{i'j'}^{\mathsf{T}}\theta}}{R_{ij}(t) e^{Z_{ij}^{\mathsf{T}}\theta} + R_{i'j'}(t) e^{Z_{i'j'}^{\mathsf{T}}\theta}} \big\{ dN_{ij}(t) + dN_{i'j'}(t) \big\}, \end{aligned}$$

for some bounded function Q. With some algebra we obtain that for $i, i' =$

$1, \ldots, n$ and $j, j' = 1, \ldots, J$

$$h(D_{ij}, D_{i'j'}; \theta_0) = Q(Z_{ij}, Z_{i'j'}) \int_0^\tau R_{ij}(t) Z_{ij}\, dM_{ij}(t) + R_{i'j'}(t) Z_{i'j'}\, dM_{i'j'}(t) -$$

$$\frac{R_{ij}(t) Z_{ij} e^{Z_{ij}^\mathsf{T}\theta_0} + R_{i'j'}(t) Z_{i'j'} e^{Z_{i'j'}^\mathsf{T}\theta_0}}{R_{ij}(t) e^{Z_{ij}^\mathsf{T}\theta_0} + R_{i'j'}(t) e^{Z_{i'j'}^\mathsf{T}\theta_0}} \big\{ dM_{ij}(t) + dM_{i'j'}(t) \big\},$$

which yields that

$$\mathbb{E}\big\{ h(D_{ij}, D_{i'j'}; \theta_0) \big\} = 0.$$

Under some mild regularity conditions, $\mathbb{E}[|h(D_{ij}, D_{i'j'}; \theta_0)|]$ is finitely upper bounded. Then Fubini's lemma further implies that

$$\mathbb{E}\big[\sum_{j=1}^{\infty} \sum_{j'=1}^{\infty} h(D_{ij}, D_{i'j'}; \theta_0) \big] = \sum_{j=1}^{\infty} \sum_{j'=1}^{\infty} \mathbb{E}\big[h(D_{ij}, D_{i'j'}; \theta_0) \big] = 0$$

whose empirical counterpart is an unbiased estimating equation

$$U_n(\theta) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{i' \neq i} \sum_{j=1}^{J_i} \sum_{j'=1}^{J_{i'}} h(D_{ij}, D_{i'j'}; \theta) = 0.$$

We obtain its solution as estimator $\widehat{\theta}$.

For estimating cumulative baseline hazard $\Lambda_0(t)$ we adopt the Nelson-Aalen estimator modified as follows for $t \in [0, \max_{i,j} C_{ij}]$:

$$\widehat{\Lambda}_0(t) = \int_0^t \frac{\sum_{i=1}^{n} \sum_{j=1}^{J_i} dN_{1ij}(u)}{\sum_{i=1}^{n} \sum_{j=1}^{J_i} R_{ij}(u) e^{Z_{ij}^\mathsf{T}\widehat{\theta}}},$$

which was inspired by the equality

$$\mathbb{E}\Big[\sum_{j=1}^{\infty} \big\{ dN_{ij}(t) - R_{ij}(t) e^{Z_{ij}^\mathsf{T}\theta_0}\, d\Lambda_0(t) \big\} \Big] = \mathbb{E}\Big[\sum_{j=1}^{J_i} \big\{ dN_{ij}(t) - R_{ij}(t) e^{Z_{ij}^\mathsf{T}\theta_0}\, d\Lambda_0(t) \big\} \Big] = 0$$

for $t > 0$, as implied by Fubini's lemma when $\mathbb{E}\big[|dN_{ij}(t) - R_{ij}(t)e^{Z_{ij}^{\mathsf{T}}\theta_0}\, d\Lambda_0(t)|\big]$ is finitely upper bounded under some mild conditions.

4.4 Asymptotic Properties

In this section, we discuss asymptotic properties of proposed estimator $(\widehat{\theta}, \widehat{\Lambda}_0)$. Strong consistency of $\widehat{\theta}$ is shown in Appendix by utilizing monotonicity and strong consistency (to zero) of $U_n(\theta)$. Define for $i = 1, \ldots, n$,

$$\varphi(D_i; \theta) = 2 \cdot \mathbb{E}\big[\sum_{j=1}^{J_i} \sum_{j'=1}^{J_{i'}} h(D_{ij}, D_{i'j'}; \theta)\big| D_i\big]$$

where i, i' are distinct indices. By Taylor's expansion and central limit theorem for U-statistics, we show in Section 4.7 that $n^{1/2}U_n(\theta)$ is asymptotically normal and can be written as

$$n^{1/2}U_n(\theta) = n^{-1/2} \sum_{i=1}^{n} \varphi(D_i; \theta) + o_P(1),$$

under some regularity conditions. Further, define matrices

$$\Gamma = \mathbb{E}\Big[\frac{\partial\, \varphi(D_i; \theta_0)}{\partial\, \theta}\Big],$$

$$\Sigma = \mathbb{E}\big[\varphi(D_i; \theta_0)\, \varphi(D_i; \theta_0)^{\mathsf{T}}\big],$$

and it can then be shown that Γ is negative definite and therefore invertible. Applying Taylor's expansion to $U_n(\widehat{\theta})$ at θ_0, and using the asymptotic normality of $n^{1/2}U_n(\theta_0)$, we show that $n^{1/2}(\widehat{\theta} - \theta_0)$ is asymptotically normal with mean zero and variance $\Gamma^{-1}\,\Sigma\,\Gamma^{-\mathsf{T}}$, and the variance can be consistently estimated by $\widehat{\Gamma}^{-1}\,\widehat{\Sigma}\,\widehat{\Gamma}^{-\top}$, where

$$\widehat{\Gamma} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial\,\widehat{\varphi}(D_i;\widehat{\theta})}{\partial\,\theta},$$

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\varphi}(D_i;\widehat{\theta})\,\widehat{\varphi}(D_i;\widehat{\theta})^{\top}.$$

Here $\widehat{\varphi}(D_i;\theta) = \frac{2}{n-1}\sum_{i'\neq i}\sum_{j=1}^{J_i}\sum_{j'=1}^{J_{i'}} h(D_{ij}, D_{i'j'};\theta)$ is the empirical estimator of $\varphi(D_i;\theta)$.

To study the weak convergence of $V(t) = n^{1/2}\big\{\widehat{\Lambda}_0(t) - \Lambda_0(t)\big\}$ ($t \in [0,\tau]$), we show in Section 4.7 that the process $V(t)$ is asymptotically equivalent to $n^{-1/2}\sum_{i=1}^{n}\psi_i(t)$, where

$$\psi_i(t) = \int_0^t \frac{\sum_{j=1}^{J_i} dM_{ij}(t)}{s_0(u,\theta_0)} - \int_0^t \frac{s_1(u,\theta_0)}{s_0(u,\theta_0)}\,d\Lambda_0(u)\cdot\Gamma^{-1}\varphi(D_i;\theta_0),$$

and

$$s_0(t,\theta) = \mathbb{E}\big[\sum_{j=1}^{J_i} R_{ij}(t)e^{Z_{ij}^{\top}\theta}\big], \quad s_1(t,\theta) = \mathbb{E}\big[\sum_{j=1}^{J_i} R_{ij}(t)Z_{ij}e^{Z_{ij}^{\top}\theta}\big].$$

By multivariate central limit theorem and a proof of tightness similar to Lin et al., 2000, we establish in Section 4.7 that $n^{1/2}\big\{\widehat{\Lambda}_0(t) - \Lambda_0(t)\big\}$ for $t \in [0,\tau]$ converges weakly to mean-zero Gaussian process with covariance function $W(s,t) = \mathbb{E}\big[\psi_i(s)\,\psi_i(t)^{\top}\big]$. We also show in Section 4.7 that $W(s,t)$ can be consistently estimated by $\widehat{W}(s,t) = n^{-1}\sum_{i=1}^{n}\widehat{\psi}_i(s)\,\widehat{\psi}_i(t)^{\top}$, where

$$\widehat{\psi}_i(t) = \int_0^t \frac{\sum_{j=1}^{J_i} d\widehat{M}_{ij}(u)}{\widehat{S}_0(u,\widehat{\theta})} - \int_0^t \frac{\widehat{S}_1(u,\widehat{\theta})}{n\widehat{S}_0(u,\widehat{\theta})^2}\,d\bar{N}(u)\cdot\widehat{\Gamma}^{-1}\,\widehat{\varphi}(D_i;\widehat{\theta}),$$

and

$$\widehat{S}_0(t,\theta) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J_i} R_{ij}(t)e^{Z_{ij}^{\mathsf{T}}\theta},$$

$$\widehat{S}_1(t,\theta) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J_i} R_{ij}(t)Z_{ij}e^{Z_{ij}^{\mathsf{T}}\theta},$$

$$\bar{N}(t) = \sum_{i=1}^{n}\sum_{j=1}^{J_i} N_{ij}(t),$$

$$\widehat{M}_{ij}(t) = N_{ij}(t) - R_{ij}(t)\widehat{\Lambda}_0(t)e^{Z_{ij}^{\mathsf{T}}\widehat{\theta}}.$$

Confidence intervals and confidence bands of estimators can then be constructed following routine proccedures. Specifically, let $z_{\alpha/2}$ be the upper $100\alpha/2$ percentile of standard normal distribution. Then for any linear transformation of $\widehat{\theta}$, $\ell\,\widehat{\theta}$, an approximate $1-\alpha$ confidence interval can be constructed as $\ell\,\widehat{\theta} \pm n^{-1/2}z_{\alpha/2}\ell\,\widehat{\Gamma}^{-1}\,\widehat{\Sigma}\,\widehat{\Gamma}^{-\mathsf{T}}\,\ell^{\mathsf{T}}$. Based on asymptotic distribution on the log transformed scale, an approximate $1-\alpha$ point-wise confidence interval for $\Lambda_0(t)$ can be constructed as

$$\widehat{\Lambda}_0(t)\exp\big\{\pm n^{-1/2}z_{\alpha/2}\widehat{W}(t,t)^{1/2}/\widehat{\Lambda}_0(t)\big\},$$

and confidence bands over interval $[t_1,t_2]\,(0 < t_1 < t_2 \leq \tau)$ can be constructed as discussed in Lin et al., 2000, by generating standard normal random variables $G_1,\ldots,G_n$ and using $n^{-1/2}\sum_{i=1}^{n}\widehat{\psi}_i(t)G_i$ as an approximation for $n^{-1/2}\sum_{i=1}^{n}\psi_i(t)$.

4.5 Simulation Studies

We conduct simulation studies to investigate the finite sample performance of proposed estimators and to validate our theoretical results. We simulate gap times Tij for subjects indexed by $i = 1, \dots, n$, such that T_{ij} has semi-marginal hazard function

$$\lambda(t; X_i, L_{ij}) = \lambda_0(t) e^{\theta_1 X_i + \theta_2 L_{ij}},$$

where $\lambda_0(t) = 0.3 \cdot \mathbb{1}(0 \leq t \leq 1) + 0.2 \cdot \mathbb{1}(t \geq 1)$, and $\theta_1 = \theta_2 = 1$. Baseline covariate X_i is generated from uniform distribution over $[0, 5]$, and $L_{ij} = \sum_{k=1}^{j-1} L_{ik} + T_{ij}$ where L_{i0} is generated from uniform distribution over $[0, 1]$. We also induce correlation between gap times for the same individual by considering the alternative form of simulated model

$$g(T_{ij}) = -\theta_1 X_i - \theta_2 L_{ij} + E_{ij},$$

where $g(t) = \log \Lambda_0(t)$, and $E_{ij} = \log \big[-\log\{1 - \Phi(\epsilon_{ij})\} \big]$ where ϵ_{ij} follows standard normal and $\Phi(\cdot)$ is its distribution function. Error ϵ_{ij}'s are generated sequentially such that ϵ_{ij} is independent of (X_i, L_{ij}). Correlation between ϵ_{ij} and $\epsilon_{ij'}$ is set to be some constant s if $|j - j'| = 1$, and 0 if otherwise. We generate independent censoring C_i from uniform distribution over $[1, 1 + A]$ for some constant A to control the number of recurrent events experienced by subjects. Various scenarios are considered varying $s = 0.3$ or 0.5, $A = 0.5$ or 1, and $n = 50, 100$ or 200. We report empirical bias, empirical standard error, empirical mean standard error estimates, and empirical coverage probability 95% confidence intervals under each scenario over 1,000 replications of simulations in Table 4.1.

Simulations show that in all scenarios as sample size grows, bias converges to zero, empirical mean of standard error gets closer to empirical standard error, and coverge probability goes to 95%, which corroborates theoretical results in Section 4.4. We also observe that convergence rate is faster when correlation between errors is smaller, and biases in parameter and standard error estimates are both generally smaller when subjects are observed to experience more events on average. This makes sense as with small s and larger A, one would expect to obtain more information from observed data. In most cases when sample size is as large as 100, bias becomes ignorable and mean standard error estimate well approximates the empirical standard error.

4.6 Data Analysis: CPCRA ddI/ddC Trial

We illustrate proposed methods and estimators by analyzing data from a randomized clinical trial conducted by Terry Beirn Community Programs for Clinical Research on AIDS, a federally funded national network of community-based research groups. The study compared didanosine (ddI) and zalcitabine (ddC) as treatments for HIV-infected patients who were intolerant or had failed treatment with zidovudine. The outcome of interest is the gap time between opportunistic events[1] and we include in the analysis patients that have experienced at least one opportunistic event after randomization.

[1] Opportunistic events considered are: candidiasis, CMV disease, cryptococcosis, cryptosporidiosis, histoplasmosis, Herpes Simplex virus infection, hist of Herpes zoster, mycobacterium avium complex (MAC), other mycobacterial infection, pneumocystis pneumonia (PCP), Progressive multifocal leukoencephalopathy (PML), tuberculosis, toxoplasmosis, lymphoma, Kaposi's Sarcoma, AIDS dementia complex (ADC), and wasting syndrome.

Table 4.1: Simulation summary statistics for $\widehat{\theta}_1$, $\widehat{\theta}_2$ and $\log\widehat{\Lambda}_0(0.1)$

		$s = 0.3$			$s = 0.5$		
		$\widehat{\theta}_1$	$\widehat{\theta}_2$	$\log\widehat{\Lambda}_0(0.1)$	$\widehat{\theta}_1$	$\widehat{\theta}_2$	$\log\widehat{\Lambda}_0(0.1)$
				$A = 0.5, \bar{J} = 4.4$			
$n = 50$	Bias	0.019	0.100	-0.130	0.019	0.161	-0.166
	ESE	0.107	0.382	0.497	0.116	0.419	0.535
	MSE	0.097	0.341	0.531	0.102	0.375	0.560
	CP	0.922	0.893	0.964	0.902	0.878	0.957
$n = 100$	Bias	0.007	0.075	-0.073	0.003	0.146	-0.103
	ESE	0.075	0.254	0.346	0.079	0.283	0.370
	MSE	0.070	0.247	0.346	0.075	0.273	0.367
	CP	0.932	0.928	0.951	0.934	0.885	0.945
$n = 200$	Bias	0.006	0.065	-0.040	-0.002	0.138	-0.075
	ESE	0.051	0.175	0.227	0.056	0.195	0.246
	MSE	0.051	0.176	0.234	0.054	0.194	0.251
	CP	0.950	0.931	0.944	0.940	0.902	0.950
				$A = 1, \bar{J} = 5.8$			
$n = 50$	Bias	0.013	0.036	-0.069	0.015	0.065	-0.088
	ESE	0.091	0.246	0.420	0.099	0.278	0.460
	MSE	0.084	0.226	0.526	0.089	0.253	0.542
	CP	0.916	0.925	0.978	0.902	0.913	0.976
$n = 100$	Bias	0.006	0.026	-0.035	0.004	0.057	-0.044
	ESE	0.064	0.174	0.293	0.070	0.194	0.320
	MSE	0.061	0.164	0.322	0.065	0.185	0.341
	CP	0.931	0.928	0.968	0.921	0.922	0.964
$n = 200$	Bias	-0.003	0.024	-0.011	-0.001	0.053	-0.023
	ESE	0.044	0.117	0.195	0.047	0.134	0.214
	MSE	0.044	0.117	0.210	0.047	0.132	0.226
	CP	0.948	0.946	0.962	0.946	0.927	0.963

Note: Bias is the empirical bias; ESE is the empirical standard error; MSE is the empirical mean of standard error estimates; CP is the empirical coverage probability of 95% confidence intervals. $\bar{J}$ is the empirical averaged number of events experienced by one subject.

Out of 467 subjects in this study, 363 are included in our analysis, among which 172 received ddI treatment and 191 received ddC treatment. In addition to the treatment variable, AIDS diagnosis indicator is available at baseline – 276 subjects were diagnosed with AIDS and 87 subjects were only HIV infected. Meanwhile, CD4 counts and Karnofsky performance scores (Mor et al., 1984) were collected every two months starting from randomization. Assuming linear change over every two-month interval, we calculate CD4 counts and karnofsky scores at events. We also obtain a quality-of-life score at each event occurrence calculated based on Table II in Neaton et al., 1994 to capture the severity of different types of opportunistic events. A lower value in CD4 count and Karnofsky score indicates deterioration in health status, while a higher quality-of-life score indicates higher severity of opportunistic event experienced. Each subject contributed 1.5 gap times to the analysis on average, and the number of gap times observed range from 1 to 5.

To model the gap time between events, we include in the regression model as covariates the CD4 count, the Karnofsky score and the quality-of-life score at last event occurrence. We also consider the effect of treatment, previous diagnosis of AIDS, and their interaction. Used as the location variable L_{ij} is the time from randomization to last observed event occurrence. To take into account the possible effect of randomization procedure on patients' health status, we introduce a "burn-in" period of three days, after which the patient was expected to have recovered from any disturbance of the study and have settled down to receive treatments. The effect of location variable is modeled separately for within and after the "burn-in" period. Effects of all continuous

Table 4.2: Analaysis results of CPCRA data under the main model

Variable	Estimate	95% CI
Location within “burn-in"	-0.12	(-0.33, 0.10)
Location after “burn-in"	2.85	(2.57, 3.12)*
CD4 count	-0.09	(-0.29, 0.11)
Karnofsky score	-0.04	(-0.27, 0.18)
Quality-of-life score	0.12	(-0.08, 0.32)
AIDS and ddI	-0.58	(-1.21, 0.05) †
no AIDS and ddC	-0.64	(-1.46, 0.18)
AIDS and ddC	1.17	(0.26, 2.08)*

Note: The reference level is set to be previous diagnosis of no AIDS and ddI treatment. Statistical significance is marker by $*$, and marginal statistical significance is marked by †.

variables are modeled on a properly log-transformed scale, and all continuous covariates are standardized to have mean zero and standard deviation one before model fitting for numerical stability. We assume that censoring is independent of gap time given covariates. Although death is part of the censoring, this assumption is made more valid by comprehensively including covariates that capture dynamic health status of patients. We also fit another model excluding CD4 count, Karnofsky score, and quality-of-life score as covariates as a way of checking robustness of results and influence of possible violation of indenpendent censoring assumption. We consider an estimate to be statistically significant if it has a p-value no greater than 0.05, and consider an estimate to be marginally significant if it has a p-value greater than 0.05 but no greater than 0.1.

Analysis results for linear coefficients under the main model are summarized in Table 4.2. We observe that higher CD4 count and Karnofsky score at

Table 4.3: Analaysis results of CPCRA data under the alternative model

Variable	Estimate	95% CI
Location within “burn-in"	-0.12	(-0.33, 0.09)
Location after “burn-in"	2.30	(2.02, 2.58)*
AIDS	-0.51	(-1.15, 0.13)
ddC	-0.57	(-1.39, 0.25)
AIDS and ddC	1.09	(0.18, 2.00)*

Note: The reference level is set to be previous diagnosis of no AIDS and ddI treatment. Statistical significance is marker by $*$.

last event occurrence are associated with longer gap time until the next event, and higher severity of opportunistic event experienced is associated with shorter gap time. However these associations are not statistically significant. There is a strong association, in both magnitude and statistial significance, between time progression and gap time – the gap times between events became shorter as time progressed after patients stabilized. This association has proven to be robust to the inclusion and exclusion of other covariates, see Table 4.3.

By testing for the differences between four groups defined by previous diagnosis and treatment under the main model, we observe interesting effect of the interaction between the two binary variables. See Table 4.4 for a summary of the testing results. In the HIV-infected group, ddC assigment is associated with a longer gap time compared with ddI, but in the AIDS group, ddC assignment is associated with a shorter gap time. Both associations are marginally statistically significant.

We also estimate the cumulative baseline function under the main model,

Table 4.4: Summary of hypothesis testing results comparing groups defined by previous diagnosis and treatment received under main model.

	HIV with ddI	HIV with ddC	AIDS with ddI	AIDS with ddC
HIV with ddI	/	0.58 (0.07)†	0.64 (0.13)	0.05 (0.88)
HIV with ddC	-0.58 (0.07)†	/	0.06 (0.87)	-0.53 (0.01)*
AIDS with ddI	-0.64 (0.13)	-0.06 (0.87)	/	-0.59 (0.10)†
AIDS with ddC	-0.05(0.88)	0.53 (0.01)*	0.59 (0.10)†	/

Note: Test statistic is coefficient for the row group minus that for the column group. P-values are in brackets. Statistical significance is marker by $*$, and marginal statistical significance is marked by †.

which is plotted in Figure 4.1 along with point-wise 95% confidence intervals. We can see that the hazard is almost constant over time, with a slight increase around 60 days and a minor decrease near 120 days after randomization.

4.7 Proofs of Asymptotic Results

4.7.1 Regularity Assumptions

To study asymptotic properties of proposed weighted pairwise comparison estimator, we impose the following regularity assumptions:

(1) We have $\theta_0 \in \Theta$, where Θ is a compact subset of Euclidean space; and $\lambda_0(t)$ is non-negative and upper bounded.

(2) Design Z_{ij}'s are bounded for all i and j.

(3) Weight function $Q(z, z')$ is a bounded positive function symmetric in z and z'.

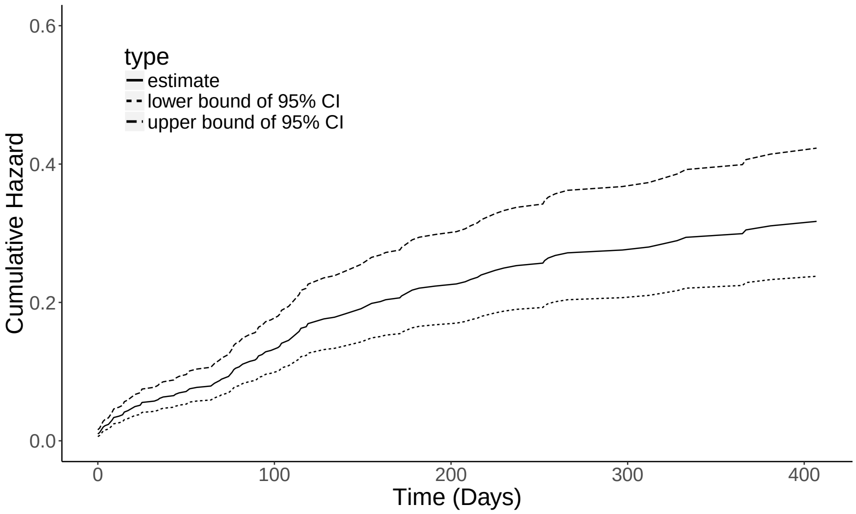

Figure 4.1: Estimate of baseline cumulative hazard function for gap times between opportunistic events using CPCRA data, along with point-wise 95% confidence intervals, under main model.

(4) Matrix Σ is positive definite, and matrix $\mathbb{E}[\partial h(D_{ij}, D_{i'j'}; \theta_0)/\partial\theta]$ is negative definite.

Under these regularity conditions, we have the following main result.

Theorem 3. Under regularity conditions (1)-(4), $\widehat{\theta}$ converges in probability to θ_0, and we further have $n^{1/2}(\widehat{\theta} - \theta_0) \xrightarrow{D} N(0, \Gamma^{-1}\,\Sigma\,\Gamma^{-\mathsf{T}})$.

Theorem 4. Under regularity conditions (1)-(4), $n^{1/2}\{\widehat{\Lambda}_0(t) - \Lambda_0(t)\}$ for $t \in [0, \tau]$ converges weakly to mean-zero Gaussian process with covariance function $W(s,t) = \mathbb{E}\big[\psi_i(s)\,\psi_i(t)^{\mathsf{T}}\big]$.

4.7.2 Consistency of $\widehat{\theta}$

Define $U(\theta) = \mathbb{E}[h(D_{ij}, D_{i'j'}; \theta)]$. The consistency of $\widehat{\theta}$ can be obtained by showing that $U(\theta) = 0$ has a unique solution at θ_0, and that $U_n(\theta)$ converges to $U(\theta)$ uniformly in θ. By arguments in Section 4.3, $U(\theta_0) = 0$. We now show the uniqueness of solution θ_0, and assume for now the negative semi-definiteness of matrix $\Gamma(\theta) = \partial U(\theta)/\partial\theta$. Then by regularity assumption (5), $\Gamma(\theta_0)$ is negative definite, implying that θ_0 is the unique solution of $U(\theta) = 0$.

Now we show that matrix $\Gamma(\theta)$ is negative definite. Denote $a^{\otimes 2} = a\, a^\top$ for some vector a. We have

$$\begin{aligned}\frac{\partial h(D_{ij}, D_{i'j'};\theta)}{\partial\theta} =& Q(Z_{ij}, Z_{i'j'}) \int_0^\tau \Big\{ \Big(\frac{R_{ij}(t) Z_{ij} e^{Z_{ij}^\top\theta} + R_{i'j'}(t) Z_{i'j'} e^{Z_{i'j'}^\top\theta}}{R_{ij}(t) e^{Z_{ij}^\top\theta} + R_{i'j'}(t) e^{Z_{i'j'}^\top\theta}} \Big)^{\otimes 2} \\ & - \frac{R_{ij}(t) Z_{ij}^{\otimes 2} e^{Z_{ij}^\top\theta} + R_{i'j'}(t) Z_{i'j'}^{\otimes 2} e^{Z_{i'j'}^\top\theta}}{R_{ij}(t) e^{Z_{ij}^\top\theta} + R_{i'j'}(t) e^{Z_{i'j'}^\top\theta}} \Big\} \cdot \big\{ dN_{ij}(t) + dN_{i'j'}(t) \big\},\end{aligned} \tag{4.3}$$

which is negative semi-definite. Therefore

$$\Gamma(\theta) = \mathbb{E}\Big[\frac{\partial h(D_{ij}, D_{i'j'};\theta)}{\partial\theta} \Big]$$

is negative semi-defnite. This completes the proof showing that $U(\theta)$ has a unique solution at θ_0. For future proofs, note that using an almost identical argument replacing probability measure with empirical measure and true coefficient value with estimator, we can show that matrix $\partial U_n(\theta)/\partial\theta$ is also negative semi-definite, thus implying that $U_n(\theta)$ is monotone in θ.

Now combined with the boundedness of θ_0 and Z_{ij} as imposed by regularity assumptions (1) and (2), we obtain the result that

$$\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{i' \neq i} \sum_{j=1}^{J} \sum_{j'=1}^{J} h(D_{ij}, D_{i'j'};\theta) \to U(\theta),$$

almost surely and point-wise in θ, as implied by the strong law of large numbers for U-statistics. By monotonicity of $U(\theta)$ and $U_n(\theta)$ in θ and continuity of $U(\theta)$, we show that $U_n(\theta)$ converges almost surely to $U(\theta)$ uniformly in θ, thus yielding $\widehat{\theta} \to \theta_0$ almost surely. This completes the proof.

4.7.3 Asymptotic Normality of $\widehat{\theta}$

By Taylor's expansion and some algebra, we have

$$n^{1/2}(\widehat{\theta} - \theta_0) = -\frac{\partial U_n(\theta^*)}{\partial \theta} \cdot n^{1/2} U_n(\theta_0),$$

where θ^* is on the line segment between $\widehat{\theta}$ and θ_0. Theorem 3 can then by proved by showing that $n^{1/2}U_n(\theta_0)$ converges in distribution to $N(0, \Sigma)$, and that $\partial U_n(\theta^*)/\partial\theta$ converges in probability to Γ.

By central limit theorem for U-statistics, under boundedness assumptions and implied by monotonicity and continuity of $U_n(\theta)$ and $\varphi(D_i; \theta)$ in θ, for any $\theta \in \Theta$

$$n^{1/2} U_n(\theta) = n^{-1/2} \sum_{i=1}^{n} \varphi(D_i; \theta) + o_P(1). \tag{4.4}$$

Specially, this implies that $n^{1/2}U_n(\theta_0)$ converges in distribution to $N(0, \Gamma)$, and that

$$\begin{aligned} \frac{\partial U_n(\theta^*)}{\partial \theta} &= \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \varphi(D_i; \theta^*)}{\partial \theta} + o_P(1) \\ &= \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \varphi(D_i; \theta_0)}{\partial \theta} + \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \varphi(D_i; \theta^{**})}{\partial \theta^2} \cdot (\theta^* - \theta_0)^\mathsf{T} + o_P(1), \end{aligned} \tag{4.5}$$

where θ^{**} is on the line segment between θ^* and θ_0. By weak law of large numbers, $n^{-1}\sum_{i=1}^{n}\partial\varphi(D_i;\theta_0)/\partial\theta$ converges in probability to Γ. Since

$$n^{-1}\sum_{i=1}^{n}\partial^2\varphi(D_i;\theta^{**})/\partial\theta^2$$

is upper bounded uniformly, and θ^* converges to θ_0, the second term on the rightmost side of last display converges to zero in probability. Therefore, $\partial U_n(\theta^*)/\partial\theta$ converges to Γ in probability. This completes the proof of Theorem 3.

4.7.4 Consistency of $\widehat{\Gamma}$ and $\widehat{\Sigma}$

By arguments identical to those used for (4.5), the $\|\widehat{\Gamma}-\Gamma\|$ converges to zero in probability. Similarly for $\widehat{\Sigma}$ we have

$$\begin{aligned}\|\widehat{\Sigma}-\Sigma\| \leq & \Big\|\frac{1}{n}\sum_{i=1}^{n}\varphi(D_i;\widehat{\theta})\,\varphi(D_i;\widehat{\theta}) - \frac{1}{n}\sum_{i=1}^{n}\varphi(D_i;\theta_0)\,\varphi(D_i;\theta_0)\Big\| + \\ & \Big\|\sum_{i=1}^{n}\varphi(D_i;\theta_0)\,\varphi(D_i;\theta_0)-\Sigma\Big\| \\ = & \Big\|\frac{1}{n}\sum_{i=1}^{n}2\,\varphi(D_i;\theta^*)\frac{\partial\,\varphi(D_i;\theta^*)}{\partial\,\theta}\cdot(\widehat{\theta}-\theta_0)^{\mathsf{T}}\Big\| + \\ & \Big\|\sum_{i=1}^{n}\varphi(D_i;\theta_0)\,\varphi(D_i;\theta_0)-\Sigma\Big\|,\end{aligned}$$

where θ^* is on the line segment between $\widehat{\theta}$ and θ_0. By uniform boundedness, and combined with the result that $\widehat{\theta}$ converges in probability to θ_0, the first term on the rightmost side of last display converges to zero. The third term also converges to zero by weak law of large numbers for U-statistics. This completes the proof showing that $\widehat{\Sigma}$ converges to Σ in probability.

4.7.5 Weak Convergence of $V(t)$ and Consistency of $\widehat{W}(s,t)$

We make the decomposition

$$\begin{aligned}V(t) =&n^{1/2}\Big\{\int_0^t \frac{d\bar{N}(u)}{n\widehat{S}_0(u,\theta_0)} - \Lambda_0(t)\Big\} + n^{1/2}\Big\{\int_0^t \frac{d\bar{N}(u)}{n\widehat{S}_0(u,\widehat{\theta})} - \int_0^t \frac{d\bar{N}(u)}{n\widehat{S}_0(u,\theta_0)}\Big\}\\
=&n^{-1/2}\sum_{i=1}^n \int_0^t \frac{d\sum_{j=1}^{J_i} M_{ij}(u)}{\widehat{S}_0(u,\theta_0)} - n^{1/2}\int_0^t \frac{d\bar{N}(u)\sum_{i=1}^n\sum_{j=1}^{J_i} R_{ij}(u)Z_{ij}e^{Z_{ij}^\top\theta^*}}{\big\{\sum_{i=1}^n\sum_{j=1}^{J_i} R_{ij}(u)e^{Z_{ij}^\top\theta^*}\big\}^2}\cdot(\widehat{\theta}-\theta_0)\\
=&V_1+V_2,\end{aligned}$$

where θ^* is on the line segment between $\widehat{\theta}$ and θ_0.

By arguments similar to those in Appendix A.2 of Lin et al., 2000, V_1 is tight and equivalent to

$$n^{-1/2}\sum_{i=1}^n \int_0^t \frac{d\sum_{j=1}^{J_i} M_{ij}(u)}{s_0(u,\theta_0)} + o_P(1).$$

By Lemma 1 of Lin et al., 2000 and uniform strong law of large numbers for i.i.d. sums and for U-statistics (Pollard, 1990; Nolan and Pollard, 1987), and combined with the result of 4.7.3, V_2 is tight and equals

$$-n^{1/2}\sum_{i=1}^n \int_0^t \frac{s_1(u,\theta_0)}{s_0(u,\theta_0)}\, d\Lambda_0(u)\Gamma^{-1}\varphi(D_i;\theta_0) + o_P(1).$$

This implies that $V(t)$ is tight and equals $n^{-1/2}\sum_{i=1}^n \psi_i(t) + o_P(1)$, which completes the proof of Theorem 4. By arguments similar to those in A.3 of Lin et al., 2000 and 4.7.4, $\widehat{W}(s,t)$ converges to $W(s,t)$ in probability uniformly in s and t.

References

Pena, Edsel A, Robert L Strawderman, and Myles Hollander (2001). “Nonparametric estimation with recurrent event data”. In: *Journal of the American Statistical Association* 96.456, pp. 1299–1315.

Aalen, Odd O and Einar Husebye (1991). “Statistical analysis of repeated events forming renewal processes”. In: *Statistics in medicine* 10.8, pp. 1227–1240.

Prentice, Ross L, Benjamin J Williams, and Arthur V Peterson (1981). “On the regression analysis of multivariate failure time data”. In: *Biometrika* 68.2, pp. 373–379.

Huang, Yijian and Ying Qing Chen (2003). “Marginal regression of gaps between recurrent events”. In: *Lifetime data analysis* 9.3, pp. 293–303.

Lin, DY, LJ Wei, I Yang, and Z Ying (2000). “Semiparametric regression for the mean and rate functions of recurrent events”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 62.4, pp. 711–730.

Mor, Vincent, Linda Laliberte, John N Morris, and Michael Wiemann (1984). “The Karnofsky performance status scale: an examination of its reliability and validity in a research setting”. In: *Cancer* 53.9, pp. 2002–2007.

Neaton, James D, Deborah N Wentworth, Frank Rhame, Carlton Hogan, Donald I Abrams, and Lawrence Deyton (1994). “Considerations in choice of a clinical endpoint for AIDS clinical trials”. In: *Statistics in medicine* 13.19-20, pp. 2107–2125.

Pollard, David (1990). “Empirical processes: theory and applications”. In: *NSF-CBMS regional conference series in probability and statistics*. JSTOR, pp. i–86.

Nolan, Deborah and David Pollard (1987). “U-processes: rates of convergence”. In: *The Annals of Statistics*, pp. 780–799.

Chapter 5

Discussion

In this dissertation, we first consider a maximum rank correlation approach to seek the optimal classification rule under a fixed tree structure for a binary outcome. We establish a general representation for a classification tree structure, which allows the definition of ROC curves for a univariate marker to be generalized to the ROC band and optimality ROC curve (OROC). The area under OROC is also proposed to measure optimal predictive performance for a tree structure. We then study a consistent estimator for the OROC, the infeasibility of which then inspire us to propose the maximum rank correlation approach through the parametrization of the so-called optimality hypersurface. Similation studies are carried out to evaluate finite sample performances of proposed approach under both correctly specified and misspecified models. Finally, we illustrate the use of the approach using the BIOCARD dataset.

The proposed methods are flexible and allows for the use of tuning parameters so that the estimation can be more tailored to specific data structures, while guanranteeing computational feasibility. Large sample properties were established under regularity conditions and validated through simulations.

Simulation studies also show that the methods yield small biases in estimated prediction performance under various scenarios even when the model is misspecified. Given the wide use of fixed classification tree structures in biomedical practices and research, the methods could provide physicians and researchers with practical tools to obtain optimal decision rules for some existing experience-based classification trees. Analysis results using these methods could also be used in clinical trials to recruit individuals at higher risk of disease, while saving resources and improving statistical efficacy of analyses at a later stage.

For future work, the proposed methods can be extended to take into account demographic variables by using functions $h_k(c_k; \theta_0, W, \beta_0)$ instead of $h_k(c_k; \theta_0)$. Here W is some additional demographic variables that do not contribute directly to the classification, but define subgroups for which the optimal cutoff values may vary. And β_0 is some parameter for demographic variables. It would also be interesting to develop variable selection techniques for markers to be included in the tree structure, which could potentially become an alternative of the greedy growing algorithm adopted by CART. Another possible direction of future work is to extend the results for binary outcomes to continous or even survival outcomes, as time-dependent ROC approaches built upon Cox's proportional hazard model have been developed and adopted in biomedical research over recent years (Heagerty, Lumley, and Pepe, 2000, Albert et al., 2018). Definitions of ROC band and OROC could be extended to the time-dependent case similar to those proposed by Heagerty, Lumley, and Pepe, 2000, and maximum rank correlation estimation for censored data can be obtained using approaches similar to those proposed in Cheng, Wei, and Ying, 1995.

We also propose the use and estimation of a semi-parametric and semi-marginal model for gap time data regression in this dissertation. We model the hazard function of gap times between successive recurrent events conditional on the last event occurrence time and some possibly time-varying covariates. The model takes a proportional hazard form, and is semi-marginal in the sense that no event occurrence history is included except for the last event time. A pair-wise comparison approach is proposed for the estimation of the model, and its large sample properties are established using U-process theories. Simulation studies illustrate the finite sample performance of proposed estimators, and the methods are further illustrated through an analysis of the CPCRA data.

The proposed model is highly flexible, and robsut to both model misspecification and various correlation structure among gap times within the same subject. The model is also innovative in allowing the inclusion of time-dependent covariates as part of the conditonal statistics, and is thus appropriate for studying time trend of gap times in a disease progression context.

For possible future work, the authors are considering extending the model to allow for non-parametric transformation of last event occurrence time L_{ij}. The inclusion of L_{ij} as part of the conditional statistics is necessary for resolving the induced dependent censoring, but in real applications, it is difficult to find appropriate forms of transformation function $\phi_i(\cdot)$, and the effect of last event occurrence is often of less scientific interest. It would also be of interest to consider simultaneous modeling of a marker process observed at the recurrence of event, a data structure commonly encountered in longitudinal studies.

References

Heagerty, Patrick J, Thomas Lumley, and Margaret S Pepe (2000). “Time-dependent ROC curves for censored survival data and a diagnostic marker”. In: *Biometrics* 56.2, pp. 337–344.

Albert, Marilyn, Yuxin Zhu, Abhay Moghekar, Susumu Mori, Michael I Miller, Anja Soldan, Corinne Pettigrew, Ola Selnes, Shanshan Li, and Mei-Cheng Wang (2018). “Predicting progression from normal cognition to mild cognitive impairment for individuals at 5 years”. In: *Brain*.

Cheng, SC, LJ Wei, and Z Ying (1995). “Analysis of transformation models with censored data”. In: *Biometrika* 82.4, pp. 835–845.

Yuxin (Daisy) Zhu

Address:	615 N Wolfe St, E3034, Baltimore, MD, 21205
Phone:	443 895 1132
email:	daisy.zhu.yx@gmail.com

Education

2013 - Now	Doctor of Philosophy in Biostatistics, **Johns Hopkins University** Expected in 2018
2009 - 2013	Bachelor of Science in Mathematics, **Nanjing University** Thesis: "Survival Analysis for Competing Risk Models"

Research Experiences

2014 - Now	Research Assistant in **BIOCARD** study at **Johns Hopkins University** PI: Dr. Marilyn Albert, co-I: Dr. Mei-Cheng Wang Working to identify **biomarkers** that predict cognitive decline among normal individuals
2015 - Now	Disability Monitoring Project using **Wearable Computing** Working with Dr. Vadim Zipunnikov to develop summary score from **activity data** to predict occurrence of adverse events like hospitalization
2016 Summer	Research Assistant in **Down Syndrome** study at **Johns Hopkins University** PI: Dr. Wayne Silverman, co-I: Dr. Mei-Cheng Wang Survival analysis for left and right censored data with competing risks
2015 Sept	**top performer** for **Prostate Cancer Dream Challenge** subchallenge 1b as part of Bmore Biostat Dream Team

Publications

with Han F. and Ren Z. (alphabetical). Adaptive Estimation of High Dimensional Partially Linear Model. Submitted.

Zhu, Y. and Wang M.C.. Optimal Decision Rule for Multiple Biomarkers Combined as Tree-based Classifiers. In preparation.

Zhu, Y. and Wang M.C.. Semi-Marginal and Semi-Parametric Analysis for Recurrent Gap Time. In preparation.

Albert, M., **Zhu, Y.**, Moghekar, A., Mori, S., Miller, M. I., Soldan, A., ... Wang, M. C. (2018). Predicting progression from normal cognition to mild cognitive impairment for individuals at 5 years. Brain.

Pettigrew, C., Soldan, A., **Zhu, Y.**, Wang, M. C., Moghekar, A., Brown, T., ... & BIOCARD Research Team. (2016). Cortical thickness in relation to clinical symptom onset in preclinical AD. NeuroImage: Clinical, 12, 116-122.

Pettigrew, C., Soldan, A., **Zhu, Y.**, Wang, M. C., Brown, T., Miller, M., ... & BIOCARD Research Team. (2016). Cognitive reserve and cortical thickness in preclinical Alzheimer's disease. Brain Imaging and Behavior, 1-11. Chicago

Deng, D., Du, Y., Ji, Z., Rao, K., Wu, Z., **Zhu, Y.**,& Coley, R. Y. (2016). Predicting survival time for metastatic castration resistant prostate cancer: An iterative imputation approach. F1000Research, 5.