Statistical Modeling and Statistical Learning for Disease Prediction and Classification

Tianle Chen

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2014

ABSTRACT

Statistical Modeling and Statistical Learning for Disease Prediction and Classification

Tianle Chen

This dissertation studies prediction and classification models for disease risk through semiparametric modeling and statistical learning. It consists of three parts. In the first part, we propose several survival models to analyze the Cooperative Huntington's Observational Research Trial (COHORT) study data accounting for the missing mutation status in relative participants (Kieburtz and Huntington Study Group, 1996a). Huntington's disease (HD) is a progressive neurodegenerative disorder caused by an expansion of cytosine-adenine-guanine (CAG) repeats at the IT15 gene. A CAG repeat number greater than or equal to 36 is defined as carrying the mutation and carriers will eventually show symptoms if not censored by other events. There is an inverse relationship between the age-at-onset of HD and the CAG repeat length; the greater the CAG expansion, the earlier the age-at-onset. Accurate estimation of age-at-onset based on CAG repeat length is important for genetic counseling and the design of clinical trials for HD. Participants in COHORT (denoted as probands) undergo a genetic test and their CAG repeat number is determined. Family members of the probands do not undergo the genetic test and their HD onset information is provided by probands. Several methods are proposed in the literature to model the age specific cumulative distribution function (CDF) of HD onset as a function of the CAG repeat length. However, none of the existing methods can be directly used to analyze COHORT proband and family data because family members' mu-

tation status is not always known. In this work, we treat the presence or absence of an expanded CAG repeat in first-degree family members as missing data and use the expectation-maximization (EM) algorithm to carry out the maximum likelihood estimation of the COHORT proband and family data jointly. We perform simulation studies to examine finite sample performance of the proposed methods and apply these methods to estimate the CDF of HD age-at-onset from the COHORT proband and family combined data. Our results show a slightly lower estimated cumulative risk of HD with the combined data compared to using proband data alone.

We then extend the approach to predict the cumulative risk of disease accommodating predictors with time-varying effects and outcomes subject to censoring. We model the time-specific effect through a nonparametric varying-coefficient function and handle censoring through self-consistency equations that redistribute the probability mass of censored outcomes to the right. The computational procedure is extremely convenient and can be implemented by standard software. We prove large sample properties of the proposed estimator and evaluate its finite sample performance through simulation studies. We apply the method to estimate the cumulative risk of developing HD from the mutation carriers in COHORT data and illustrate an inverse relationship between the cumulative risk of HD and the length of CAG repeats at the IT15 gene.

In the second part of the dissertation, we develop methods to accurately predict whether pre-symptomatic individuals are at risk of a disease based on their various marker profiles, which offers an opportunity for early intervention well before definitive clinical diagnosis. For many diseases, existing clinical literature may suggest the risk of disease varies with some markers of biological and etiological importance, for example age. To identify effective prediction rules using nonparametric decision functions, standard statistical learning approaches treat markers with clear biological importance (e.g., age) and other markers without prior knowledge on disease etiology interchangeably as input variables. Therefore, these approaches may be inadequate in singling out and preserving the effects from the biologically important variables, especially in the presence of potential noise markers. Using age as an example of a salient marker to receive special care in the analysis, we propose a local smoothing large margin classifier implemented with support vector machine to construct effective age-dependent classification rules. The method adaptively adjusts age effect and separately tunes age and other markers to achieve optimal performance. We derive the asymptotic risk bound of the local smoothing support vector machine, and perform extensive simulation studies to compare with standard approaches. We apply the proposed method to two studies of premanifest HD subjects and controls to construct age-sensitive predictive scores for the risk of HD and risk of receiving HD diagnosis during the study period.

In the third part of the dissertation, we develop a novel statistical learning method for longitudinal data. Predicting disease risk and progression is one of the main goals in many clinical studies. Cohort studies on the natural history and etiology of chronic diseases span years and data are collected at multiple visits. Although kernel-based statistical learning methods are proven to be powerful for a wide range of disease prediction problems, these methods are only well studied for independent data but not for longitudinal data. It is thus important to develop time-sensitive prediction rules that make use of the longitudinal nature of the data. We develop a statistical learning method for longitudinal data by introducing subject-specific long-term and short-term latent effects through designed kernels to account for within-subject correlation of longitudinal measurements. Since the presence of multiple sources of data is increasingly common, we embed our method in a multiple kernel learning framework and propose a regularized multiple kernel statistical learning with random effects to construct effective nonparametric prediction rules. Our method allows easy integration of various heterogeneous data sources and takes advantage of correlation among longitudinal measures to increase prediction power. We use different kernels for each data source taking advantage of distinctive feature of data modality, and

then optimally combine data across modalities. We apply the developed methods to two large epidemiological studies, one on Huntington's disease and the other on Alzhemeier's Disease (Alzhemeier's Disease Neuroimaging Initiative, ADNI) where we explore a unique opportunity to combine imaging and genetic data to predict the conversion from mild cognitive impairment to dementia, and show a substantial gain in performance while accounting for the longitudinal feature of data.

Key words: Huntington's disease; Age-at-onset; Disease prediction; Varying-coefficient model; Self-consistency equation; Statistical learning; Local smoothing; Reproducing kernel Hilbert space; Risk bound; longitudinal data; Integrative analysis; Latent effects

Table of Contents

List of Figures

List of Tables

Acknowledgments

This work would not have been possible without the support of many people. Many thanks to my advisor, Professor Yuanjia Wang, for her invaluable guidance and help while this work being conducted. Special thanks to Professor Donglin Zeng for providing invaluable comments and help. I also thank Professor Douglas Langbehn for providing help and suggestions. Also thanks to my committee members, Professor Todd Odgen, Professor Ken Cheung, Professor Jeff Goldsmith, and Professor Karen Marder, who offered guidance and support. Last but not least, thanks to all my classmates at Columbia University for their support and help.

To my family

Chapter 1

Introduction

1.1 Overview

This dissertation develops several new methods for disease classification and prediction for time-to-event data, cross-sectional and longitudinal binary data. The dissertation consists of three parts. In the first part (Chapter 2), we propose statistical modeling approaches for the analysis of age-at-disease-onset data for chronic diseases. We model the relationship between mutation status and age-at-onset for Huntington's disease through parametric and non-parametric survival models with methods to handle missing information on mutation status and account for right-censoring. In the second and third parts (Chapters 3 and 4), we propose statistical learning approaches for the analysis of cross-sectional and longitudinal binary data. In Chapter 3, we consider a targeted local kernel support vector machine to construct effective age-dependent rules for classifying mutation status in pre-symptomatic subjects and predicting disease onset in mutation carriers. By using the local kernel method we are able to catch the nonlinear age effect on disease risk and other markers associated with a chronic disease and provide age-specific prediction rules for subjects in different age groups. In Chapter 4, we propose a multi-kernel support vector machine to construct prediction rules for disease onset and progression. We use separate kernels for modeling markers from heterogeneous data sources and integrate the information in a non-sparse fashion. We design kernels to model the subject-specific long-term and short-term latent effects to extract information from correlated longitudinal outcome data. Our method provides prediction rules at subject-specific level and improves prediction accuracy, especially in the situation of predicting future outcomes based on existing data collected from the same subject.

1.2 Introduction to the statistical modeling for disease risk prediction

Huntington's disease (HD) is a severe dominantly inherited neurodegenerative disorder that affects motor, cognitive, and psychiatric function and is uniformly fatal. HD is caused by the expansion of CAG repeats in the IT15 gene that codes for the protein Huntingtin (Ross, 1995; Ross and Tabrizi, 2011). Affected individuals typically begin to show motor signs around 30-50 years of age and eventually die 15-20 years after the disease onset (Foroud et al., 1999). Despite identification of the causative gene, there is currently no treatment that delays or stops disease progression.

One large genetic epidemiological study of HD, the Cooperative Huntington's Observational Research Trial (COHORT), including 42 Huntington Study Group research centers in North America and Australia was initiated in 1996, and concluded in 2011 (Kieburtz and Huntington Study Group, 1996a; Dorsey et al., 2008). Participants in COHORT (denoted as probands) underwent a clinical evaluation and DNA from whole blood was genotyped for mutations in the IT15 gene. Since 2005, COHORT probands particiapte in family history interviews and provide information on HD affection status in their family members. While CAG repeat length is ascertained in probands, the high cost of conducting in-person interviews of family members prevents the collection of all family members' blood samples. Family members' morbidity and mortality information such as age-at-onset (AAO) of HD, is obtained through systematic interviews of the probands or the family members themselves. Although a relative's genotype is unavailable, the corresponding distribution of the HD genotypes can be obtained based on the relative's relationship with the proband and the proband's CAG repeat lengths (Wang et al., 2008).

In a genetic counseling setting, CAG repeat length greater than or equal to 36 is defined as carrying the HD mutation (carrier), CAG repeat length between 27 and 35 is defined as intermediate, and CAG repeat length less than or equal to 26 is defined as normal, or non-carrier (Rubinsztein et al., 1996; Ha et al., 2012). It is known that there is an inverse association between the CAG repeat length and AAO of HD, i.e. the longer the repeat length, the earlier the motor onset (Langbehn et al., 2004). Modeling such a relationship as well as the conditional distribution of HD onset given CAG repeat length accurately and precisely is important for genetic counseling and the design of clinical trials for HD. The AAO of HD is subject to right censoring by the constraint of the study period. Several formulae were proposed in the literature to estimate the survival function of HD onset given CAG repeat length (e.g., Stine et al., 1993; Rubinsztein et al., 1996; Langbehn et al., 2004). Langbehn et al. (2004) has shown that the standard semiparametric survival models, such as the Cox proportional hazards model, do not fit the HD data and proposed a new logistic-exponential parametric model. Specifically, the conditional distribution of HD onset given the CAG repeat length is modeled as a logistic function, with a location and a scale parameter depending on CAG through a non-linear relationship. This model allows the mean and variance of HD onset to depend on CAG by exponential functions respectively, offering flexibility to the distribution function. Other parametric models, such as Gamma distribution, were also proposed in the literature (Gutierrez and MacDonald, 2004). Langbehn et al. (2010a) examines several population models in the literature and show the superior performance of Langbehn et al. (2004) in terms of predicting the two-year onset probability in an independent prospective data.

None of the aforementioned existing methods can be directly used to analyze CO-

HORT family data because family members are not always genotyped and their HD mutation status is unknown. When the inclusion of family data contributes additional information, however, the unobserved mutation status complicates the analysis. To see this, note that the affected parent carrying the mutation has a 50% chance of transmitting the mutation to an offspring. An added complexity is that the likelihood of the offspring having a longer CAG repeat length than the parent is higher if the parent is the father. Since the offspring is not genotyped, whether he or she carries expanded CAG repeats is unknown. In this work, we treat the mutation transmission status as missing data and use the EM algorithm to carry out the maximum likelihood estimation of the proband and family data jointly. Conditionally on the transmission status in family members, we use the logistic-exponential model in Langbehn et al. (2004) to model the AAO as a function of CAG repeat length. We perform simulation studies to examine finite sample performances of the proposed methods. Finally, we apply these methods to analyze the COHORT proband and family combined data. Our results show a slightly lower estimated cumulative risk of HD with the combined data compared to using proband data alone.

We then extend the parametric model from Langbehn et al. (2004) to a nonparametric model. Instead of modeling the mean and variance of HD onset as exponential functions on CAG repeats, we consider a nonparametric varying-coefficient model of the cumulative risk using a logistic link

$$\operatorname{logit}\{\operatorname{pr}(T_i \le t | X_i)\} = \beta_0(t) + c_0(X_i) + \beta_1(t)c_1(X_i), \tag{1.1}$$

where $c_0(x)$ and $c_1(x)$ are known functions of covariates. To provide flexibility and protect against misspecification, $\beta_0(t)$ and $\beta_1(t)$ are left as unknown nonparametric functions. Note that when $c_1(x) = 1/s(x;\gamma)$, $c_0(x) = -\mu(x;\alpha)/s(x;\gamma)$, $\beta_0(t) = 0$, and $\beta_1(t) = t$, model (1.1) reduces to that in Langbehn et al. (2004). For the sake of simple illustration, further consider a special case of (1.1), a varying-coefficient proportional odds model, where

$$\operatorname{logit}\{\operatorname{pr}(T_i \leq t|X_i)\} = \beta_0(t) + \beta_1(t)X_i. \tag{1.2}$$

The interpretation of $\beta_1(t)$ is then directly related to the cumulative risk of disease, since $\exp\{\beta_1(t)\}$ is the odds ratio of experiencing disease onset by age t for subjects with one unit difference in X. Since $\operatorname{pr}(T_i \leq t|X_i)$ is a cumulative distribution function, $\beta_0(t)$ and $\beta_0(t) + \beta_1(t)X_i$ are constrained to be non-decreasing functions of t. In applications where X_i's are positive, we require $\beta_1(t)$ to be non-decreasing as well. When $\beta_0(t)$ and $\beta_1(t)$ are constants that do not vary with t, model (1.2) reduces to a standard proportional odds model.

In the literature, Jung (1996) directly modeled survival function using regression model at a fixed time point without considering temporal effect. There are a number of other works on extending proportional odds model to account for temporal covariate effect or time-varying covariates. Peng and Huang (2007) proposed a strict extension of Cox proportional hazards model in the context of proportional odds model to account for temporal effect. The procedure involves solving a series of estimating equation sequentially, which may be computationally heavy. Chen et al. (2012) proposed methods to extend transformation models considered in for example, Zeng and Lin (2007), to account for external time-varying covariates.

Here, we take a completely different approach that does not involve counting process and with straightforward and simple computational algorithm. When there is no censoring, to estimate the cumulative risk function at a time point t_0 given a covariate, e.g., $\operatorname{pr}(T_i \leq t_0|X_i)$, a straightforward analysis is to fit a logistic regression of $I(T_i \leq t_0)$ on the covariates X_i. When the outcome is subject to censoring, $I(T_i \leq t)$ may not be observed for some of the censored subjects. Let C_i denote the censoring time, Efron (1967) proposed a nonparametric estimator of a survival function by redistributing the conditional masses for the censored subjects, $\operatorname{pr}(T_i > C_i|C_i)$, equally to the non-censored observations above C_i, where the weights depend on the number of at-risk subjects. Portnoy (2003) and Wang and Wang (2009) used similar idea to fit a quantile regression with covariates X_i, where the conditional point masses $\mathrm{pr}(T_i > C_i|C_i, X_i)$ for censored subjects are re-distributed to the right. For quantile regression, the estimator only depends on the signs of residuals and thus the point masses for censored subjects are re-distributed to $+\infty$. Since there are covariates involved, the conditional masses to be estimated depend on the covariates and the unknown distribution function.

In the first part of this dissertation, to estimate $\beta(t_0)$ from (1.1) or (1.2), we fit a pseudo-logistic regression of $I(T_i \leq t_0)$ through redistributing weights to the right to account for censoring. We apply the procedure to estimate the coefficient function at distinct uncensored event times, and smooth the coefficient functions across the entire support of event times when necessary. This type of smoothing was found to be equivalent to applying local kernel smoothing directly (Ma and Wei, 2012). The proposed computational procedure is extremely easy to implement and can be handled by standard softwares. We investigate the asymptotic properties of the proposed estimator to show consistency and normality, and conduct simulation studies to examine its finite sample performance. The proposed methods are applied to estimating the cumulative risk of developing HD from subjects with IT15 gene mutation using the COHORT data and illustrate an inverse relationship between the cumulative risk of HD and the length of CAG repeats. We compare the estimates under model (1.2) with fully nonparametric Kaplan-Meier estimates using subjects with the same CAG values and reveal consistent results.

1.3 Introduction to the targeted local kernel support vector machine for age-dependent classification and prediction

An important research goal for chronic diseases is to develop effective early intervention to delay onset, slow disease progression, and provide different treatment or care management at each stage based on subject-specific characteristics (Paulsen et al., 2006). It is necessary to identify biological, behavioral and clinical markers that can be combined to distinguish premanifest subjects at high risk of a disease from those who are at low risk or free of risk. For many illnesses, existing clinical literature may suggest the risk of disease varies with some markers of biological and etiological importance. For example, it is well known that the risk of Alzheimer's disease increases with age (Celsis, 2000), and the predictive power of other markers and their relative importance often change over a subject's lifespan. It is beneficial to take advantage of the existing etiologic information on disease risk to develop age-sensitive diagnostic rules in conjunction with other markers with less clear prior biological information on disease risk to boost predictive power. Using age as an example of a salient marker to receive special care in the analysis, we develop methods to treat biologically important variables separately from other variables in the presence of some potential noise markers. The developed prediction rules have implications on prioritizing other markers and informing timing of therapeutic interventions to guide personalized medicine.

To predict binary outcomes such as disease status, regression-based methods including logistic regression and time-varying coefficient models are often used (Cai et al., 2000; Wang et al., 2009b). These models focus on estimating population-average association (e.g., odds ratio) instead of making subject-specific prediction or classification, thus may not be optimal (Pepe et al., 2004, 2006; Ware, 2006). For example, variables that are themselves not significant at certain levels may contribute to improving prediction in combination especially when they are highly correlated (Wei et al., 2009). To directly focus on classification and prediction, large margin-based statistical learning approaches (e.g., Vapnik, 1995; Shen et al., 2003; Zhang et al., 2006; Wang et al., 2009a; Wu and Liu, 2013) can be used. The geometric set up of these methods is to construct an optimal separating boundary between two classes by maximizing the margin from each class to the boundary. The equivalent statistical framework is to minimize a margin-based loss function subject to a regularization penalty. They are among the most successful nonparametric and robust classifiers in practice that can improve individual-specific prediction and classification problems especially in high-dimensional settings with correlated variables (Moguerza and Munoz, 2006; Orru et al., 2012). Among the large-margin based classifiers, support vector machine (SVM) is one of the most popular binary classifiers proven to exhibit some optimal theoretical properties (Lin, 2002). Recently, Ladicky and Torr (2011) and Zhang et al. (2011b) proposed a non-specific locally linear smoothing in the SVM context using all the features variables. However, what they considered is based on local affine approximation of the entire feature variable space involving variables in all dimensions. Their locality is defined by all the features variables in a neighborhood of a data point. When the dimension of the feature variable space is high, it may be difficult to perform smoothing in the entire feature space due to sparseness of data in any local neighborhood. In addition, since these approaches are based on linearization of a potentially high-dimensional nonparametric surface, stronger assumptions on the smoothness of separating boundary in all dimensions of the feature space are required.

One convenient approach to incorporate age information to classify a subject's at-risk status is to treat age as one of input variables interchangeably with other markers and learn classification rules using kernel machine (e.g., Gaussian kernel). However, such a strategy may not be optimal for several reasons. First, from a clinical point of view, age plays distinctive clinical and biological roles on disease risk. It is the easiest factor to measure to be used for indicating the timing of intervention and guiding choices of treatments, and thus it should call for some special attention. Second, from a statistical point of view, lumping age together with other markers exchangeably in a learning algorithm is very likely to dilute the age signal especially when the marker dimension is not small and some noise variables are included. Furthermore, since all variables are tuned by the same tuning parameter, age effect may be masked by the other markers which potentially introduce noise. This is observed in our subsequent numerical studies. Lastly, using fully nonparametric learning without distinguishing age from other markers makes it difficult to provide an interpretable and practical guideline for timely intervention.

In the second part of this dissertation, we develop a large-margin based classifier implemented with SVM for discriminating subjects at risk through solving a kernel weighted optimization problem to provide age-dependent prediction rules from markers collected in cross-sectional studies. Since disease risk for two subjects close in age is expected to be similar controlling for other characteristics, certain smoothness with respect to age is anticipated so it can be taken advantage of when classifying a subject's disease status. The proposed approach uses a local smoothing kernel to pool information across subjects similar in age and selects the tuning parameter for age separately from tuning parameter for other markers. Therefore, we adaptively estimate age effect and protect the age signal from being lost especially when noise markers are present. We first consider interpretable locally linear prediction rules where the age profile for each marker can be easily presented and used to assess importance of each marker. We then consider more general nonlinear prediction rules through kernel machines locally at each age. Our method differs from the literature (Ladicky and Torr, 2011; Zhang et al., 2011b) in that there exists a targeted variable with strong prior knowledge to be predictive or needs to be adjusted. We perform local smoothing along one targeted dimension of a well-motivated content-important variable (e.g., age) while leaving other variables intact. Our approach only requires data to be reasonably abundant along one targeted dimension.

The remainder of this work is organized as follows. In section 3.2, we describe the details of the proposed method and provide an easy computational algorithm supporting the method. In section 3.3, we study the theoretical properties of the risk bound as a function of local smoothing kernel bandwidth. In section 3.4, we perform extensive simulation studies to compare the proposed method with several alternative approaches and examine the finite sample properties of the fitted classification boundaries. In section 3.5, we apply the proposed methods to two Huntington's disease (HD) data examples (Dorsey and Huntington Study Group COHORT Investigators, 2012; Paulsen et al., 2008) to predict age-specific risk of developing HD or risk of pre-symptomatic subjects receiving HD diagnosis during study period using motor, cognitive and behavioral markers, and show the age-dependent profiles of several key markers. Some concluding remarks are given in section 3.6.

1.4 Introduction to the multiple kernel learning with random effects for predicting longitudinal outcomes and data integration

Accurate prediction of current and future clinical status of a patient based on subject-specific clinical and biological markers is an important goal for early diagnosis and monitoring diseases progression. Modern technologies offer opportunities to collect data from heterogeneous sources such as genetic data, imaging data, and clinical data including electronic health records. Furthermore, many cohort studies on natural history and etiology of chronic diseases often span years and data may be collected at multiple visits. It is thus important to develop time-sensitive prediction rules that not only integrate data from multiple sources but also make use of the longitudinal nature of the data collected from the same subjects.

There is an extensive body of literature on longitudinal data analysis exploring the association between candidate predictors and outcomes measured repeatedly over time (See for example, Diggle et al., 2002). In these association analyses, primary goals are estimation and hypothesis testing of regression parameters which may not necessarily yield powerful prediction rules. For the purpose of prediction with longitudinal data, a number of works focus on linear or quadratic discriminant analysis of longitudinal profiles or a sample of curves (e.g., James and Hastie, 2001; Marshall and Baron, 2000; Luts et al., 2012). These works aim to classify a functional curve into two groups and rely on either linear mixed effects models (Verbeke and Lesaffre, 1996; Marshall and Baron, 2000) or functional data analysis or their extensions (James and Hastie, 2001) to perform classification. In the past decades, there has been growing interest in using powerful machine learning methods to build effective predictive models for binary and continuous disease outcomes (Oquendo et al., 2012). Particularly, kernel-based methods such as support vector machine or support vector regression are proposed to classify longitudinal profile into groups (Pearce and Wand, 2009; Luts et al., 2012). However, disease outcomes in these approaches do not change with time so they are not applicable to classify clinical outcomes assessed repeatedly over time. Since most of the existing statistical learning methods assume the sample to be independent and identically distributed, there is a lack of literature on how to effectively incorporate within-subject dependence to improve prediction of future subjects' clinical outcomes or within-subject change especially when the outcomes are binary ones.

In the third part of this dissertation, we introduce a novel statistical learning method to predict longitudinal binary outcomes in the multiple kernel learning (Lanckriet et al., 2004; Bach and Lanckriet, 2004) framework. Our method not only uses observed feature variables but also introduces subject-specific unobserved latent variables to extract information from correlated outcomes and build time-sensitive prediction rules. More specifically, we use multiple additive kernels for observed feature variables, which can account for heterogeneous data sources taking advantage of the correlation within each data modality, while at the same time, we account for within-subject correlation of longitudinal measurements by introducing subject-specific short-term and long-term latent random effects modeled through a separate kernel. In many biomedical studies, the observed feature variables only explain some proportion of variability in outcomes, and the gain from using latent random effects to extract information from the remaining unexplained variability can be substantial. The weights used for each kernel are tuned based on minimizing the overall loss, therefore we optimally combine data across modalities in a data-driven fashion. In addition to methods for training model, we also develop methods for predicting future outcomes through observed features and unobserved latent effects when longitudinal training data are available.

On one hand, depending on the choice of kernels, the proposed method bears some similarity with semiparametric or nonparametric mixed effect models for longitudinal data. However, unlike traditional mixed models, our proposed method aims at prediction accuracy, allows greater flexibility through the use of kernel machines, and is relatively easy to scale up for large dimensional data. On the other hand, using different kernels for feature variables and latent variables shares the same advantages with multiple kernel learning methods which have been developed to handle challenges to integrate different data sources (Pavlidis et al., 2002; Lanckriet et al., 2004; Yu et al., 2010; Zhang and Shen, 2012). Specifically, the latter treats each data source component, for example, genetic data, imaging data or clinical data, as belonging to separate kernel spaces and finds an optimal way to combine them for prediction. The multiple kernel methods have been shown to yield much improved performance as compared to using one single kernel in various biomedical applications (Yu et al., 2010). Although our proposed method uses multiple kernels, one significant difference from the above literature is that separate kernels are also applied to unobserved latent variables in our method.

Next we summarize third part of this dissertation. In Section 4.2, we propose learning method to predict longitudinal binary outcomes based on the support vector machine with multiple kernels. In Section 4.3, extensive simulation studies are conducted to illustrate small-sample performance of the proposed method and compare with some existing approaches. In Section 4.4, we apply the developed method to two real data examples. In the first example, we predict Huntington's disease (HD) diagnosis in a large multi-site HD epidemiological study from various sources of clinical interviews and biomarkers and show that the proposed method outperforms single kernel approaches and multiple kernel approaches without accounting for subject-specific correlations in terms of both predicting future subjects and predicting future outcomes on the same subject. In the second example, we apply the proposed method to analyze the Alzheimer's Disease Neuroimaging Initiative (AD-NI) data, where a unique opportunity is presented to combine various modalities of imaging and genetic data to distinguish subjects with mild cognitive impairment (M-CI) from subjects with Alzheimer's Disease (AD), and we show a substantial gain in performance while accounting for the longitudinal correlation of data. The proposed multiple kernel fusion with random effects proves to be effective in both applications. Some remarks are provided in Section 5.

Chapter 2

Statistical modeling for disease risk prediction

2.1 Overview

In this chapter, we propose parametric and non-parametric statistical models for chronic diseases with approaches to deal with missing information and censored data. In the first section, we develop a parametric survival model with EM algorithm for predicting cumulative risk of disease onset in a mixed population with partial missing information. In the second section, we extend the parametric model to a non-parametric model and deal with censoring through re-distributing weights approach. In the third section, we summarize our findings and discuss possible extensions.

2.2 Parametric survival model with EM algorithm

2.2.1 Methods

We start by introducing some notations. For the ith subject, let T_i denote the age-at-onset of HD, let δ_i be the event indicator, let C_i denote the censoring time, and let $X_i = \min(T_i, C_i)$. Let A_i denote the CAG repeat length. Langbehn et al. (2004) models distribution of T_i given A_i by a logistic function. The cumulative distribution function (CDF) given A_i is

$$F(t|A_i) = \Pr(T_i \leq t|A_i) = \frac{1}{1 + e^{-[t-\mu(A_i)]/s(A_i)}}, \tag{2.1}$$

and the density function is

$$f(t|A_i) = \frac{e^{-[t-\mu(A_i)]/s(A_i)}}{s(A_i)\{1 + e^{-[t-\mu(A_i)]/s(A_i)}\}^2}.$$

Here $\mu(A_i)$ is a location parameter depending on the covariate A_i and $s(A_i)$ is a scale parameter depending on A_i. Let $S(t|A_i) = 1 - F(t|A_i)$ denote the survival function of HD onset. The location and scale parameters have the following relationship with the mean and variance of T_i given A_i:

$$E(T_i|A_i) = \mu(A_i), \quad \text{var}(T_i|A_i) = \frac{\pi^2}{3}s^2(A_i)$$

Various parametric functions for the location and scale parameters were compared in Langbehn et al. (2004, 2010b), and the exponential function provides the best fit. Therefore we use the same model where

$$\begin{aligned} \mu(A_i) &= \mu_1 + \exp(\mu_2 - \mu_3 A_i), \\ \text{and} \ \ \text{var}(A_i) &= \sigma_1 + \exp(\sigma_2 - \sigma_3 A_i). \end{aligned}$$

Substitute these into $F(t|A_i)$ and $f(t|A_i)$ to obtain a parametric model for the distribution of AAO of HD with six parameters, $\beta = (\mu_1, \mu_2, \mu_3, \sigma_1, \sigma_2, \sigma_3)^T$.

2.2.1.1 Proband-only analysis

First, consider proband's data where all A_i's are observed. Since a subject's AAO of HD is subject to the right censoring, the likelihood function is

$$L(\beta) = \prod_{i=1}^{n} f^{\delta_i}(X_i|A_i;\beta)S^{1-\delta_i}(X_i|A_i;\beta), \tag{2.2}$$

and the log-likelihood is

$$l(\beta) = \sum_{i=1}^{n} \left\{ -\delta_i \log[s(A_i)] - \frac{X_i - \mu(A_i)}{s(A_i)} - (1+\delta_i) \log \left[1 + e^{-\frac{X_i - \mu(A_i)}{s(A_i)}} \right] \right\}.$$

The maximum likelihood estimate (MLE) of the parameters, $\widehat{\beta}$, can be obtained via a general-purpose optimization algorithm such as Newton-Raphson or Nelder-Mead implemented in the 'optim' function of the R program version 2.13.1. The variance-covariance matrix of $\widehat{\beta}$ is estimated by the inverse of the estimated Hessian matrix,

$$\widehat{\text{cov}}(\widehat{\beta}) = [H(\widehat{\beta})]^{-1}.$$

The standard error of the survival function, $\widehat{S}(t|A_i)$, is then estimated by the Delta method, that is,

$$\widehat{\text{var}}[\widehat{S}(t|A_i)] = G^T(\widehat{\beta})\widehat{\text{var}}(\widehat{\beta})G(\widehat{\beta}),$$

where the gradient vector

$$G(\widehat{\beta}) = \frac{\partial S(t|A_i)}{\partial \beta}\Big|_{\beta=\widehat{\beta}}.$$

2.2.1.2 Incorporating family members

Next, we consider incorporating family members' AAO data. We do not observe whether a family member inherits the mutation in the HD gene from the proband, but we observe whether a subjects has developed HD based on a systematic interview with the proband. The likelihood of AAO of HD takes a mixture form. Let p_i denote the probability of the ith subject receiving a deleterious allele from a proband and therefore becoming a carrier. Such probabilities are calculated based on Mendelian transmission (Wang et al., 2008). For example, offsrping and siblings of a carrier proband have a probability of 50% of receiving the Huntingtin mutation. We assume that conditioning on a family member receiving Huntingtin allele, the CAG repeat length is the same as observed in the proband, although this is a simplification. For subjects who receive a wild type allele (CAG<36), their probability of developing HD is zero, thus $f(t|A_i < 36) = 0$, and $S(t|A_i < 36) = 1, \forall t$. For the family members, the likelihood is

$$L(\beta) = \prod_{i=1}^{n} [p_i f^{\delta_i}(X_i|A_i;\beta) S^{1-\delta_i}(X_i|A_i;\beta) + (1-p_i)(1-\delta_i)],$$

where the above second term follows from the assumption that non-carriers do not develop HD. Note that for all carrier probands we observe $p_i = 1$, thus the likelihood reduces to (2.2).

The above likelihood can be maximized by a combination of EM and Newton-Raphson algorithm. Let G_i denote the unobserved carrier status indicator for the ith family member (i.e., $G_i = 1$ indicates a family member receives a mutation and $G_i = 0$ indicates otherwise). Then the complete data log-likelihood is

$$\sum_{i=1}^{n} I(G_i = 1)\{\delta_i f(X_i|A_i;\beta) + (1-\delta_i)\log[S(X_i|A_i;\beta)]\}$$

At the $(k+1)$th iteration of the E-step, we compute the conditional expectation of the complete data log-likelihood, given the observed data. Essentially, we compute

$$\begin{aligned} w_i^{(k+1)} &= E[I(G_i = 1)|X_i, \delta_i, \beta^{(k)}] \\ &= \frac{p_i f^{\delta_i}(X_i|A_i;\beta^{(k)}) S^{1-\delta_i}(X_i|A_i;\beta^{(k)})}{p_i f^{\delta_i}(X_i|A_i;\beta^{(k)}) S^{1-\delta_i}(X_i|A_i;\beta^{(k)}) + (1-p_i)(1-\delta_i)}. \end{aligned}$$

In the M-step, we update $\beta^{(k+1)}$ by maximizing the weighted log-likelihood

$$\sum_{i=1}^{n} w_i^{(k+1)}\{\delta_i f(X_i|A_i;\beta) + (1-\delta_i)\log[S(X_i|A_i;\beta)]\}$$

using the Newton-Raphson algorithm developed for the proband data.

Since the parameters are estimated by the MLE, it is straightforward to carry out the likelihood ratio tests (LRTs) to compare the model fit from the COHORT data with the ones obtained in other studies such as Langbehn et al. (2004). Here, twice the difference in the log-likelihood follows a chi-square distribution with 6 degrees of freedom.

2.2.2 Simulation studies

We conducted two simulation studies closely related to the observed COHORT data to illustrate the performance of the Newton-Raphson optimization and the EM algorithm (Laird and Ware, 1982). In all our optimization procedures, we centered both A_i and X_i. Since the direct optimization and EM algorithm need reasonable initial values, we fitted two nonlinear least square (NLS) to the observed sample mean and variance of the AAO on subjects with $\delta_i = 1$. To be specific, we fit

$$m_1(a_i) = \mu_1 + \exp(\mu_2 - \mu_3 a_i), \quad s_1^2(a_i) = \sigma_1 + \exp(\sigma_2 - \sigma_3 a_i),$$

where $m_1(a_i)$ and $s_1^2(a_i)$ are the sample mean and variance for all subjects with $A_i = a_i$, respectively. The six NLS estimators were used as the initial values for further optimization. We denoted the estimated β from the centered data as $\widehat{\beta}_c$. For each simulation, the un-centered $\widehat{\beta}$ were then calculated based on $\widehat{\beta}_c$ and the sample mean of A_i and X_i.

We restricted simulations to CAG repeat lengths between 41 and 56 to guard against sensitivity to the extremely high or low CAG repeats to be consistent with Langbehn et al. (2004). For the analysis of proband data, we generated a sample of 2000 subjects, each with a CAG repeat length ranging from 41 to 56 that follows a multinomial distribution in which the probability $\text{pr}(A_i = a)$ equals to the observed proportion of $A_i = a$ in the COHORT proband data set (Table 2.5). The failure times T_i were simulated from the distribution (2.1), where the parameters β were fixed at the values fitted from the COHORT proband data (see next section for their values). The censoring times, C_i, were generated from a re-scaled Beta distribution.

For the analysis of the combined proband and family data, we generated a sample of 4000 subjects. The probabilities p_i were generated by re-sampling the observed p_i's in the COHORT data. With a given p_i for each subject, we simulated his or her mutation carrier status from a Bernoulli distribution with success probability p_i. For family members simulated to receive an expanded CAG repeat (carriers), their

CAG repeats A_i were set to be the same as the probands and their failure times were simulated from (2.2) with β fixed at estimates from the COHORT combined data. For non-carrier family members, their failure times were set to be infinity and their $X_i = C_i$. We used the same censoring distribution for generating C_i as in the first simulation study.

The results in Tables 2.1 and 2.2 report the mean estimated $\widehat{\beta}$ and $\widehat{\beta}_c$, their mean estimated standard errors, and empirical standard deviation in the two simulations. We see from these tables that the mean $\widehat{\beta}$ is very close to the true β in both studies. The mean estimated standard errors of $\widehat{\beta}_c$ are close to the empirical standard deviations, indicating that the estimation of variability is appropriate. Since one of our goals is to estimate the CDF of HD onset, we also examine the estimated $\widehat{F}(t|A_i)$ at typical A_i's. Figures 2.1 and 2.2 present three curves of $\widehat{F}(t|A_i)$ at $A_i = 41$, 46, 50 and their 95% empirical confidence intervals for the proband data and combined data, respectively. We see that $\widehat{F}(t|A_i)$ coincide with the circles representing true $F(t|A_i)$ at various ages. We provide numerical values of $F(t|A_i)$, mean $\widehat{F}(t|A_i)$, empirical standard deviation of $\widehat{F}(t|A_i)$, and the mean estimated standard error of $\widehat{F}(t|A_i)$ at various ages in Table 2.3 and 2.4.

2.2.3 COHORT data analysis results

We first describe the proband and family data in the COHORT study. Information on CAG repeat length and age was available for 1357 probands with CAG repeats varying from 36 to 100 (Table 2.5). There were 3409 first-degree relatives available from 675 probands. We show the descriptive statistics for the relatives stratified by relationship type in Table 2.6. A subset of 1151 subjects with CAG length between 41 and 56 was our proband data set (21 subjects whose self-reported and clinician-reported age at the onset of symptom differed by greater than 15 years were removed) and used for the proband-only analysis. Similar to Langbehn et al. (2004), we restricted the analysis to CAG repeat lengths between 41 and 56 to guard against sensitivity to the extremely high or low CAG repeats.

Information on CAG repeat length, age at time of evaluation and the probability of being a carrier (receiving Huntingtin mutation from the proband) was available for 2851 family members of all 1151 probands. In the proband data set, both individuals with manifest HD and presymptomatic carriers are included. Their age-at-diagnosis and age-at-first- motor sign were recorded. Among 1151 probands, 876 (76%) subjects had experienced HD onset and the average AAO of the HD diagnosis was 44 years of age. There were 54% females and 94% Caucasians. Our combined proband and family data has 4002 subjects. In this combined data set, 51%were females and 35% subjects had experienced HD onset. Among the 4002 subjects, 467 are singletons (probands with no family member included). Among the rest 3535 subjects, there are 623 families with average size 5.674 (sd=2.609). In the combined data, there are two different probabilities of being a carrier: $p_i = 1$ (1199 subjects) or $p_i = 0.5$ (2803 subjects). Among the 2851 family members, 966 are parents of the probands, 1095 are siblings of the probands and 790 are children of the probands.

When using the age-at-diagnosis in our proband data as T_i, the estimated cumulative risk of HD is

$$F(t|A_i) = \left(1 + \exp\left\{-\frac{\pi}{\sqrt{3}}\frac{[t - 16.284 - \exp(3.428 - 8.325A_i)]}{\sqrt{22.379 + \exp(15.657 - 0.284A_i)}}\right\}\right)^{-1}.$$

The estimated parameters for the CDF from the proband-only analysis are different from the ones obtained from Langbehn et al. (2004). Our estimated mean and standard deviation of the AAO of HD is about 1 to 3 years later than the ones obtained in Langbehn et al. (2004), and the standard deviation (SD) is slightly smaller (Table 2.7). In addition, the estimated CDF is lower at most A_i values using COHORT data. We ran a likelihood ratio test of

$$H_0 : \beta = \beta_0 \quad vs. \quad H_1 : \beta = \widehat{\beta}_1,$$

where β_0 are the values obtained in Langbehn et al. (2004) and the p-value was less than 0.001. When analyzing the age-at-first-symptom in our proband data, the estimated cumulative risk of HD is

$$\widehat{F}(t|A_i) = \left(1 + \exp\left\{-\frac{\pi}{\sqrt{3}}\frac{[t - 14.266 - \exp(7.987 - 0.104A_i)]}{\sqrt{28.933 + \exp(17.130 - 0.312A_i)}}\right\}\right)^{-1}.$$

We present $\widehat{F}(t|A_i)$ curves for age-at-diagnosis and age-at-symptom at various CAG lengths and their 95% confidence intervals for the proband data in Figures 2.3. It can be seen that with a given A_i, the estimated probability of having the first symptoms of HD is higher than the probability of a diagnosis of HD at the same age. This is consistent with the intuition that symptoms of HD will be observed before a diagnosis. The mean AAO of first-symptom is estimated to be about 2 years earlier than AAO of diagnosis (Table 2.7) and the standard deviation of the former is slightly larger, indicating that age-at-first-symptom is more variable.

As a sensitivity analysis, we compared the estimated CDF based on the parametric model with a nonparametric Kaplan-Meier estimator for subjects with a given A_i. Figure 2.4 presents this comparison using probands' age-at-diagnosis data. We show in the figure that the parametric model fit is consistent with the Kaplan-Meier fit. However, as expected, the confidence interval for the parametric model estimate at a given age is narrower than the Kaplan-Meier estimate (results not shown). The figure comparing age-at-symptom data is similar and therefore omitted.

We analyzed the AAO of the first symptom using the combined proband and family data, since the age-at-diagnosis was not available for family members. The estimated cumulative risk of HD at age t is

$$\widehat{F}(t|A_i) = \left(1 + \exp\left\{-\frac{\pi}{\sqrt{3}}\frac{[t - 18.832 - \exp(8.461 - 0.118A_i)]}{\sqrt{32.365 + \exp(14.823 - 0.248A_i)}}\right\}\right)^{-1}.$$

The corresponding $\widehat{F}(t|A_i)$ curves at various CAG lengths and their 95% confidence intervals are shown in Figure 2.5. In Table 2.7, we compare the estimated mean and SD of the AAO from the proband and combined data. We can see that the estimated mean AAOs for several CAGs are similar regardless of whether family members are included. The SD estimated from the model is larger for the combined data. This is a

Table 2.1: Simulation 1 (probands data). Estimated parameters and standard errors of the direct optimization of proband-only analysis, $n = 2000$, 1000 replications.

	β	Mean $\widehat{\beta}$	Mean $\widehat{\beta}_c$	Mean se($\widehat{\beta}_c$)	Empi sd($\widehat{\beta}_c$)
$\mu 1$	14.402	13.748	-29.574	2.777	3.362
$\mu 2$	8.197	8.168	3.439	0.090	0.107
$\mu 3$	0.108	0.107	0.107	0.011	0.013
$\sigma 1$	19.687	19.323	19.323	10.737	9.847
$\sigma 2$	12.354	13.497	3.484	0.360	0.358
$\sigma 3$	0.200	0.227	0.227	0.088	0.086

reflection of the observed data in that there is a wider range of AAO in the combined data than in the proband data. For example, the SD for CAG=41 of the former is 11 years, whereas it is 10 years in the probands, and the SD for CAG=42 is 10 in the combined and 8 in the probands.

One of the utilities of the estimated curves is to estimate the conditional probability of having an HD onset (or staying HD free) in the next five or ten years, given a subject has not had an onset by a given age. In Table 2.8, we present such conditional probabilities in five-year intervals for a subject without HD at age 40 and with given CAG repeat length. For example, a 40-year pre-symptomatic subject with a CAG of 40 has a probability of 11% (CI: 9%, 14%) of developing HD in the next 10 years (by age 50), while for a subject with a CAG of 50 this probability increases to 93% (CI: 91%, 95%).

Table 2.2: Simulation 2 (combined data). Estimated parameters and standard errors of the EM algorithm with combined proband and family analysis, $n = 4000$, 1000 replications.

	β	Mean $\widehat{\beta}$	Mean	Mean	Empi
			$\widehat{\beta}_c$	se($\widehat{\beta}_c$)	sd($\widehat{\beta}_c$)
μ1	18.943	18.735	-30.795	2.449	2.556
μ2	8.628	8.647	3.307	0.093	0.096
μ3	0.121	0.122	0.122	0.013	0.013
σ1	32.538	30.432	30.432	11.683	11.241
σ2	14.642	15.000	3.923	0.271	0.269
σ3	0.244	0.253	0.253	0.075	0.076

Table 2.3: Simulation 1 (probands data). Estimated CDF and standard errors from the direct optimization of proband-only analysis, $n = 2000$, 1000 replications.

	CAG = 41				CAG = 46				CAG = 50			
Age	$F(t\|A_i)$	Mean $\widehat{F}(t\|A_i)$	Empi sd	Mean $\widehat{\text{sd}}$	$F(t\|A_i)$	Mean $\widehat{F}(t\|A_i)$	Empi sd	Mean $\widehat{\text{sd}}$	$F(t\|A_i)$	Mean $\widehat{F}(t\|A_i)$	Empi sd	Mean $\widehat{\text{sd}}$
10	0.0001	0.0001	0.0000	0.0000	0.0003	0.0003	0.0001	0.0001	0.0011	0.0012	0.0005	0.0004
20	0.0006	0.0007	0.0002	0.0002	0.0049	0.0048	0.0008	0.0009	0.0301	0.0309	0.0066	0.0060
30	0.0046	0.0049	0.0011	0.0012	0.0717	0.0709	0.0066	0.0068	0.4560	0.4578	0.0253	0.0248
40	0.0322	0.0335	0.0051	0.0054	0.5492	0.5487	0.0171	0.0162	0.9577	0.9572	0.0084	0.0077
50	0.1944	0.1972	0.0162	0.0160	0.9505	0.9509	0.0052	0.0056	0.9984	0.9983	0.0007	0.0006
60	0.6368	0.6358	0.0227	0.0219	0.9967	0.9967	0.0006	0.0007	0.9999	0.9999	0.0000	0.0000
70	0.9272	0.9252	0.0102	0.0108	0.9998	0.9998	0.0001	0.0001	1.0000	1.0000	0.0000	0.0000
80	0.9893	0.9887	0.0025	0.0026	1.0000	1.0000	0.0000	0.0000	1.0000	1.0000	0.0000	0.0000
90	0.9985	0.9984	0.0005	0.0005	1.0000	1.0000	0.0000	0.0000	1.0000	1.0000	0.0000	0.0000

Table 2.4: Simulation 2 (combined data). Estimated CDF and standard errors from the EM algorithm with combined proband and family analysis, $n = 4000$, 1000 replications.

	CAG = 41				CAG = 46				CAG = 50			
Age	$F(t\|A_i)$	Mean $\widehat{F}(t\|A_i)$	Empi sd	Mean $\widehat{\text{sd}}$	$F(t\|A_i)$	Mean $\widehat{F}(t\|A_i)$	Empi sd	Mean $\widehat{\text{sd}}$	$F(t\|A_i)$	Mean $\widehat{F}(t\|A_i)$	Empi sd	Mean $\widehat{\text{sd}}$
10	0.0006	0.0006	0.0002	0.0002	0.0010	0.0010	0.0002	0.0002	0.0025	0.0026	0.0008	0.0008
20	0.0028	0.0029	0.0007	0.0007	0.0102	0.0102	0.0014	0.0014	0.0373	0.0374	0.0069	0.0068
30	0.0134	0.0137	0.0023	0.0023	0.0928	0.0928	0.0069	0.0070	0.3754	0.3751	0.0241	0.0238
40	0.0609	0.0616	0.0069	0.0069	0.5041	0.5042	0.0148	0.0143	0.9031	0.9030	0.0139	0.0132
50	0.2373	0.2378	0.0149	0.0146	0.9099	0.9100	0.0076	0.0074	0.9931	0.9930	0.0020	0.0019
60	0.5987	0.5979	0.0200	0.0188	0.9901	0.9901	0.0015	0.0014	0.9996	0.9995	0.0002	0.0002
70	0.8773	0.8761	0.0133	0.0125	0.9990	0.9990	0.0002	0.0002	1.0000	1.0000	0.0000	0.0000
80	0.9717	0.9710	0.0050	0.0047	0.9999	0.9999	0.0000	0.0000	1.0000	1.0000	0.0000	0.0000
90	0.9940	0.9937	0.0015	0.0014	1.0000	1.0000	0.0000	0.0000	1.0000	1.0000	0.0000	0.0000

Table 2.5: Descriptive statistics of the COHORT proband data

		Numbers and ages for a CAG repeat of																					
		36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57+	Total
At risk	Number	2	5	15	21	43	55	68	57	31	28	18	9	5	2	1	0	2	0	0	0	0	0	362
	Ave age	61	64	48	55	50	45	42	39	37	31	34	34	27	23	30		35						42
	Min age	60	61	26	37	25	21	21	18	19	19	20	21	20	21	30		23						18
	Max age	62	69	66	70	88	67	71	62	51	44	51	53	40	25	30		47						88
	sd	1	3	11	9	14	11	11	10	9	7	9	9	9	3	.		17						13
	%	0.6	1.4	4.1	5.8	11.9	15.2	18.8	15.7	8.6	7.7	5.0	2.5	1.4	0.6	0.3	0.0	0.6	0.0	0.0	0.0	0.0	0.0	
Affected	Number	2	1	6	7	67	128	148	144	143	93	83	47	34	21	18	9	7	10	6	3	3	15	995
	Ave age	54	68	55	53	60	55	51	48	44	41	38	36	33	31	31	30	28	23	26	26	23	20	45
	Min age	49	68	46	25	37	28	17	19	21	16	25	21	20	19	22	23	22	11	18	25	17	12	11
	Max age	59	68	67	77	82	76	76	67	67	58	53	48	46	44	39	35	35	29	31	29	28	27	82
	sd	7	.	7	19	10	9	9	8	8	8	6	6	6	6	5	4	5	6	5	2	6	4	12
	%	0.2	0.1	0.6	0.7	6.7	12.9	14.9	14.5	14.4	9.3	8.3	4.7	3.4	2.1	1.8	0.9	0.7	1.0	0.6	0.3	0.3	1.5	
Total	Number	4	6	21	28	110	183	216	201	174	121	101	56	39	23	19	9	9	10	6	3	3	15	1357

Table 2.6: Descriptive statistics of the first-degree relatives of COHORT proband subjects stratified by relationship

		Numbers and ages for relationship of			
		Parents	Siblings	Children	Total
Not affected	Number	739	1110	931	2780
	Ave age	70	50	26	42
	Min age	27	0	0	18
	Max age	111	93	62	88
	sd	13	15	14	13
	%	26.6	39.9	33.5	
Affected	Number	379	237	13	629
	Ave age	45	42	36	45
	Min age	18	7	23	11
	Max age	82	70	44	82
	sd	11	11	7	12
	%	60.3	37.7	2.1	
Total	Number	1118	1347	944	3409

Table 2.7: Mean and standard deviation of AAO estimated from the model (2.1) for four analyses.

	Langbehn data		COHORT data					
			Probands diagnosis*		Probands symptom**		Combined symptom†	
CAG	Mean	SD	Mean	SD	Mean	SD	Mean	SD
41	57.06	10.50	59.84	8.78	57.74	9.13	59.33	11.68
43	48.06	8.62	51.17	7.31	49.32	7.90	50.63	9.60
46	38.66	7.08	41.29	5.97	39.66	6.57	41.20	7.59
48	34.32	6.57	36.31	5.47	34.75	5.95	36.69	6.79
50	31.08	6.28	32.32	5.16	30.80	5.50	33.21	6.28

* : Using proband age-at-diagnosis data;

**: Using proband age-at-first-symptom data;

†: Using proband and relative combined age-at-first-symptom data.

Table 2.8: Conditional survival probabilities estimated from the COHORT combined data

CAG	45 years	50 years	55 years	60 years	65 years	70 years
36	0.01 (0.00 , 0.02)	0.02 (0.00 , 0.04)	0.04 (0.00 , 0.08)	0.07 (0.01 , 0.13)	0.11 (0.20 , 0.20)	0.17 (0.07 , 0.28)
37	0.01 (0.00 , 0.02)	0.03 (0.01 , 0.06)	0.06 (0.02 , 0.11)	0.11 (0.05 , 0.18)	0.18 (0.27 , 0.27)	0.28 (0.17 , 0.39)
38	0.02 (0.01 , 0.03)	0.05 (0.02 , 0.08)	0.10 (0.06 , 0.15)	0.18 (0.12 , 0.25)	0.29 (0.38 , 0.38)	0.43 (0.33 , 0.53)
39	0.03 (0.02 , 0.04)	0.08 (0.05 , 0.11)	0.17 (0.12 , 0.21)	0.29 (0.23 , 0.35)	0.44 (0.52 , 0.52)	0.60 (0.52 , 0.69)
40	0.05 (0.04 , 0.06)	0.14 (0.11 , 0.16)	0.27 (0.23 , 0.31)	0.44 (0.39 , 0.50)	0.62 (0.68 , 0.68)	0.77 (0.72 , 0.82)
41	0.08 (0.07 , 0.09)	0.22 (0.19 , 0.24)	0.41 (0.37 , 0.44)	0.61 (0.57 , 0.65)	0.78 (0.81 , 0.81)	0.88 (0.86 , 0.91)
42	0.13 (0.12 , 0.14)	0.34 (0.32 , 0.36)	0.57 (0.54 , 0.60)	0.77 (0.74 , 0.79)	0.89 (0.90 , 0.90)	0.95 (0.94 , 0.96)
43	0.21 (0.20 , 0.22)	0.48 (0.46 , 0.51)	0.72 (0.70 , 0.75)	0.87 (0.86 , 0.89)	0.95 (0.95 , 0.95)	0.98 (0.97 , 0.98)
44	0.31 (0.29 , 0.33)	0.63 (0.60 , 0.65)	0.83 (0.81 , 0.85)	0.93 (0.92 , 0.95)	0.97 (0.98 , 0.98)	0.99 (0.99 , 0.99)
45	0.43 (0.40 , 0.45)	0.74 (0.72 , 0.77)	0.90 (0.88 , 0.92)	0.96 (0.96 , 0.97)	0.99 (0.99 , 0.99)	>0.99 (0.99 , >0.99)
46	0.53 (0.50 , 0.56)	0.82 (0.80 , 0.85)	0.94 (0.93 , 0.95)	0.98 (0.97 , 0.99)	0.99 (>0.99 , >0.99)	>0.99 (>0.99 , >0.99)
47	0.61 (0.57 , 0.64)	0.87 (0.85 , 0.89)	0.96 (0.95 , 0.97)	0.99 (0.98 , 0.99)	>0.99 (>0.99 , >0.99)	>0.99 (>0.99 , >0.99)
48	0.66 (0.63 , 0.70)	0.90 (0.88 , 0.92)	0.97 (0.96 , 0.98)	0.99 (0.99 , >0.99)	>0.99 (>0.99 , >0.99)	>0.99 (>0.99 , >0.99)
49	0.70 (0.66 , 0.74)	0.92 (0.90 , 0.94)	0.98 (0.97 , 0.99)	0.99 (0.99 , >0.99)	>0.99 (>0.99 , >0.99)	>0.99 (>0.99 , >0.99)
50	0.73 (0.68 , 0.77)	0.93 (0.91 , 0.95)	0.98 (0.97 , 0.99)	>0.99 (0.99 , >0.99)	>0.99 (>0.99 , >0.99)	>0.99 (>0.99 , >0.99)
51	0.74 (0.69 , 0.80)	0.94 (0.91 , 0.96)	0.98 (0.98 , 0.99)	>0.99 (0.99 , >0.99)	>0.99 (>0.99 , >0.99)	>0.99 (>0.99 , >0.99)
52	0.76 (0.70 , 0.82)	0.94 (0.91 , 0.97)	0.99 (0.98 , >0.99)	>0.99 (0.99 , >0.99)	>0.99 (>0.99 , >0.99)	>0.99 (>0.99 , >0.99)
53	0.77 (0.70 , 0.83)	0.95 (0.92 , 0.98)	0.99 (0.98 , >0.99)	>0.99 (0.99 , >0.99)	>0.99 (>0.99 , >0.99)	>0.99 (>0.99 , >0.99)
54	0.77 (0.70 , 0.85)	0.95 (0.92 , 0.98)	0.99 (0.98 , >0.99)	>0.99 (0.99 , >0.99)	>0.99 (>0.99 , >0.99)	>0.99 (>0.99 , >0.99)
55	0.78 (0.70 , 0.86)	0.95 (0.92 , 0.99)	0.99 (0.98 , >0.99)	>0.99 (0.99 , >0.99)	>0.99 (>0.99 , >0.99)	>0.99 (>0.99 , >0.99)
56	0.78 (0.70 , 0.87)	0.95 (0.92 , 0.99)	0.99 (0.98 , >0.99)	>0.99 (0.99 , >0.99)	>0.99 (>0.99 , >0.99)	>0.99 (>0.99 , >0.99)

Figure 2.1: Estimated CDF of HD onset for $A_i = 41, 43, 46$, and 50 with simulated proband data.

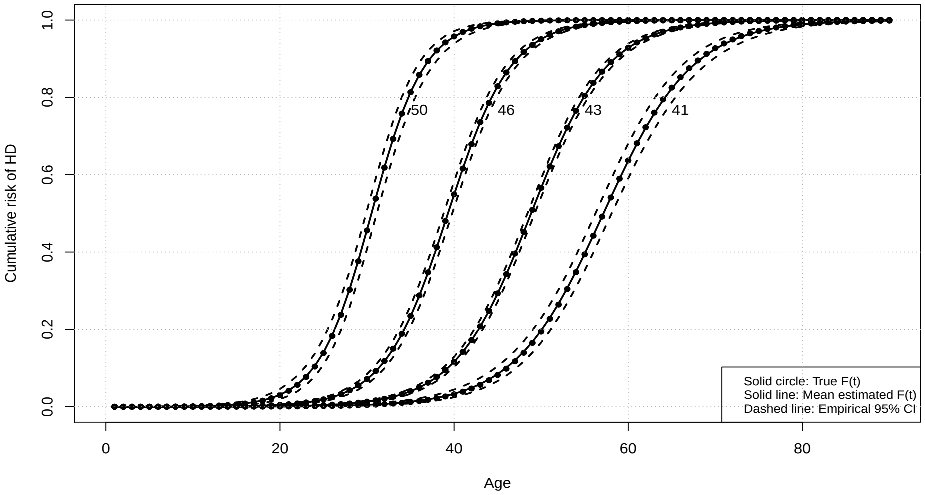


Figure 2.2: Estimated CDF of HD onset for $A_i = 41, 43, 46$, and 50 with simulated proband and family data.

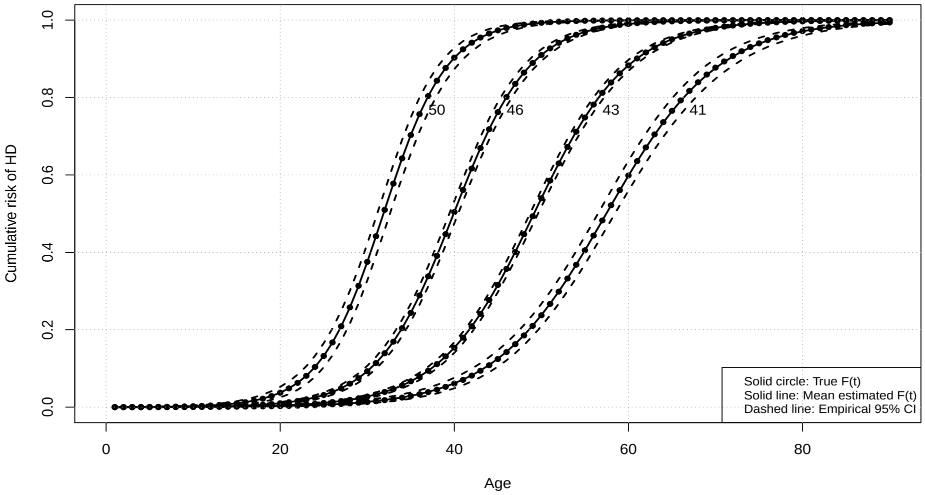


Figure 2.3: Estimated CDFs of age-at-diagnosis and age-at-first-symptom of HD for $A_i = 41, 43, 46$, and 50 with COHORT proband data.

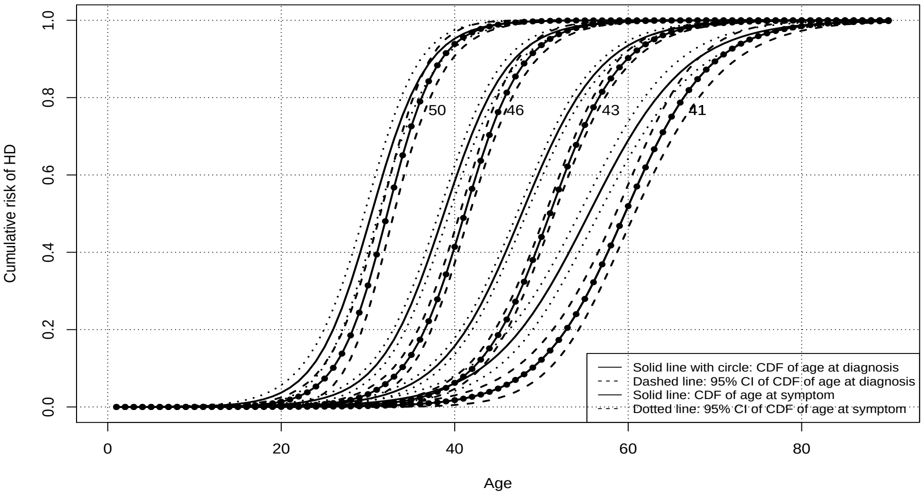


Figure 2.4: Kaplan-Meier curve and estimated CDF of age-at-diagnosis of HD for $A_i = 41, 46$, and 50 with COHORT proband data.

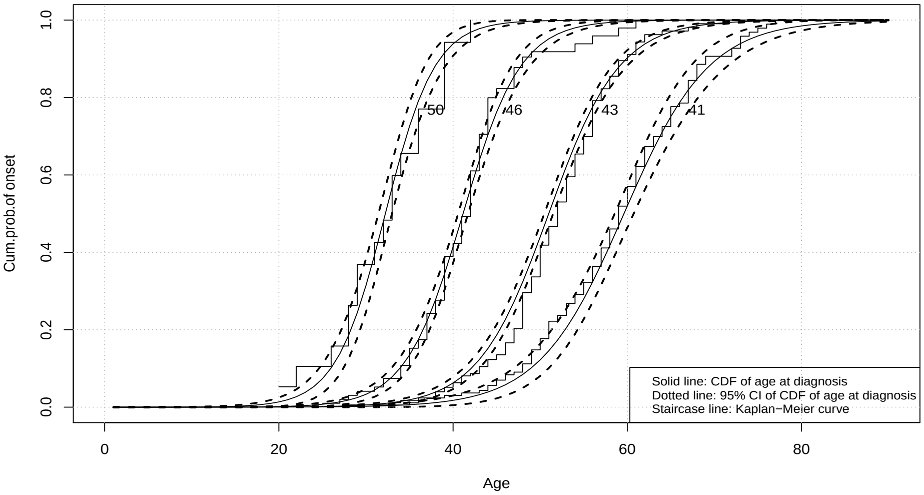
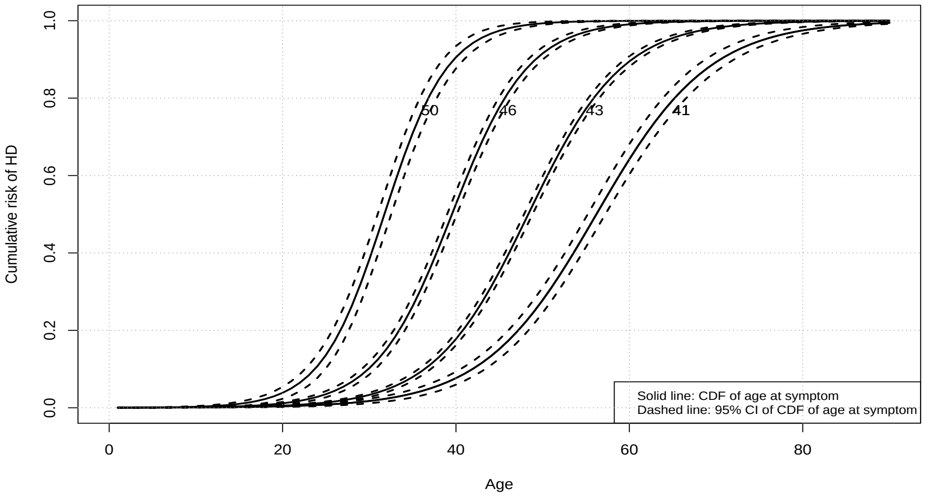


Figure 2.5: Estimated CDF of age-at-first-symptom of HD for $A_i = 41, 46,$ and 50 with COHORT proband and family data.



2.3 Non-parametric model with re-distributing weight-s

2.3.1 Methods

For the purpose of illustration, we mainly focus on the varying coefficient model (1.2). It is straightforward to extend to the more general model (1.1).

2.3.1.1 Uncensored data

First we investigate estimation at a fixed time point t_0 when the outcome is not subject to censoring. Let $\beta(t) = \{\beta_0(t), \beta_1(t)\}^T$, let $\theta = \beta(t_0)$ denote $\beta(\cdot)$ evaluated at t_0, and let $Z_i = (1, X_i)^T$. When there is no censoring, the likelihood for $\{I(T_i \leq t_0), X_i, i = 1, \cdots, n\}$ under a logistic link takes the standard form, $\prod_i \frac{\exp\{I(T_i \leq t_0) Z_i^T \theta\}}{1 + \exp\{Z_i^T \theta\}}$. To estimate θ, we solve the estimating equation

$$\sum_{i=1}^{n} m(X_i, T_i; t_0, \theta) = 0,$$

where $m(X_i, T_i; t_0, \theta) = \{I(T_i \leq t_0) - \mu(X_i; \theta)\} Z_i$, and $\mu\{X_i; \theta\} = \frac{\exp\{Z_i^T \theta\}}{1 + \exp\{Z_i^T \theta\}}$. The influence function for the estimate $\widehat{\theta}$ is

$$\phi(X_i, T_i; t_0, \theta) = A(X_i; \theta) \{I(T_i \leq t_0) - \mu(X_i; \theta)\} Z_i,$$

where $A(X_i; \theta) = \Big(E[\mu(X_i; \theta)\{1 - \mu(X_i; \theta)\} Z_i Z_i^T]\Big)^{-1}$. We fit a logistic regression of $I(T_i \leq t_0)$ on X_i and repeat this process while varying t_0 at all distinct values of observed T_i's. One can then smooth the estimates $\widehat{\beta}(t_0)$ as a function of t_0 (Ma and Wei, 2012) subject to the monotonicity constraint. An alternative is to fit a nonparametric regression (for example using splines) treating $I(T_i \leq t)$ as generalized outcomes. This method was shown to have similar performance as the post-hoc smoothing above (Ma and Wei, 2012), but is more difficult to implement under the monotonicity constraint, therefore we do not further explore here.

2.3.1.2 Censored data

When a subject is right censored (i.e., $T_i > C_i$) and $C_i \geq t_0$, we still observe $I(T_i \leq t_0) = 0$. Ambiguity occurs when a subject is censored and $C_i < t_0$. We propose an estimator that re-distributes weights to the right for ambiguous subjects based on self-consistency equations similar to Efron (1967) and Wang and Wang (2009). Let $O_i = \{X_i, T_i \wedge C_i, \Delta_i \equiv I(T_i \leq C_i)\}^T$ denote the ith observation. We solve the following weighted estimating equation

$$S_n(O_i; \theta, \beta) = n^{-1} \sum_{i=1}^{n} s(O_i; t_0, \theta, \beta) = 0, \tag{2.3}$$

where

$$s(O_i; t_0, \theta, \beta) \quad = \quad w\{O_i; t_0, \beta(\cdot)\} m(X_i, T_i \wedge C_i; t_0, \theta) + [1 - w\{O_i; t_0, \beta(\cdot)\}] m(X_i, +\infty; t_0, \theta),$$

and

$$w\{O_i; t_0, \beta(\cdot)\} = \begin{cases} 1, & \Delta_i = 1 \text{ or } (\Delta_i = 0 \text{ and } C_i \geq t_0) \\ \dfrac{F(t_0|X_i) - F(C_i|X_i)}{1 - F(C_i|X_i)}, & \Delta_i = 0 \text{ and } C_i < t_0. \end{cases} \tag{2.4}$$

Here $F(t|x) = \mu\{x; \beta(t)\}$ is the conditional distribution of T_i given X_i introduced in model (1.2), and the weight for the ith subject depends on $\beta(\cdot)$ evaluated at t_0 and C_i.

To gain insights on the weights, note that subjects with observed $I(T_i \leq t_0)$ will receive a weight of one for their contributions to the estimating equation. Subjects with missing $I(T_i \leq t_0)$ have conditional probability masses

$$E\{I(T_i \leq t_0) | T_i > C_i, X_i, C_i\} \quad = \quad \frac{F(t_0|X_i) - F(C_i|X_i)}{1 - F(C_i|X_i)}.$$

Treating (X_i, C_i) as pseudo-observations for censored subjects with censoring time less than t_0, they receive weights $w\{O_i; t_0, \beta(\cdot)\} = \text{pr}(T_i \leq t_0 | T_i > C_i, X_i)$. We re-distribute their complementary weights $1 - w\{O_i; t_0, \beta(\cdot)\} = \text{pr}(T_i > t_0 | T_i > C_i, C_i, X_i)$ to the right. Since the outcomes are binary variables, the complementary masses $1 - w\{O_i; t_0, \beta(\cdot)\}$ for pseudo-observations (X_i, C_i) can be re-distributed to any point that is greater than all observations that is not specific to any observation above C_i (see also Portnoy (2003); Wang and Wang (2009)). Thus, any point above C_i contributes the same information to the estimating equation. Without loss of generality, we re-distribute the complementary mass to $+\infty$, and the contribution from these observations to the estimating equation is $m(X_i, +\infty; t_0, \theta) = -\mu(X_i; \theta)Z_i$.

In practice, the weights $w\{O_i; t_0, \beta(\cdot)\}$ depend on unknown distribution function $F(\cdot|X)$ which needs to be estimated. An initial estimator for $\beta(\cdot)$ is easily obtained by the inverse probability of censoring weighting (IPW) proposed in Bang and Tsiatis (2000), which weights subjects having an event by the inverse of their probabilities of not being censored. To be specific, we can obtain an initial estimator by solving the estimating equation

$$\sum_{i=1}^{n} \frac{I(T_i \leq C_i)m(X_i, T_i; t_0, \theta)}{G(T_i)},$$

where $G(\cdot)$ is the survival function for the censoring times C_i. Estimating $G(\cdot)$ by the Kaplan-Meier of the censoring process, the estimating equation for θ is

$$\sum_{i=1}^{n} \frac{I(T_i \leq C_i)m(X_i, T_i; t_0, \theta)}{\widehat{G}(T_i)} = 0. \tag{2.5}$$

This process is repeated for t_0 on a grid $(u_1, \cdots, u_M)$ and denote the resulting estimator as $\widehat{\beta}(u_j), j = 1, \cdots, M$.

Substituting $\widehat{\beta}(\cdot)$ in (2.4) to obtain weights $w\{O_i; t_0, \widehat{\beta}(\cdot)\}$ to be redistributed, the final estimator $\widehat{\theta}_n$ then solves the weighted estimating equation

$$S_n(O; t_0, \theta, \widehat{\beta}) = n^{-1} \sum_{i=1}^{n} s(O_i; t_0, \theta, \widehat{\beta}) = 0. \tag{2.6}$$

It is extremely easy to implement this weighting scheme. Without loss of generality, assume the first n_0 subjects have unobserved outcomes $I(T_i \leq t_0)$. Create pseudo-observations $\widetilde{O}_1 = (X_1, +\infty, \Delta_1), \cdots, \widetilde{O}_{n_0} = (X_{n_0}, +\infty, \Delta_{n_0})$. Append all pseudo-observations to the original observations to obtain observations $(O_1, \cdots, O_n, \widetilde{O}_1, \cdots, \widetilde{O}_{n_0})$ with weights

$$[w\{O_1; t_0, \widehat{\beta}(\cdot)\}, \cdots, w\{O_n; t_0, \widehat{\beta}(\cdot)\}, 1 - w\{O_1; t_0, \widehat{\beta}(\cdot)\}, \cdots, 1 - w\{O_{n_0}; t_0, \widehat{\beta}(\cdot)\}].$$

Then $\widehat{\theta}_n$ is estimated by a weighted logistic regression. The weights $w\{O; t_0, \widehat{\beta}(\cdot)\}$ extract information at multiple time points simultaneously, and thus pool information across time points to estimate the distribution function at t_0.

2.3.2 Asymptotic properties

To show consistency and asymptotic normality of $\widehat{\beta}(t)$ at fixed t obtained from (2.6), we will need the following technical conditions:

A1. Assume that $\beta(t)$ is right continuous with left-hand limits (cadlag) componentwise.

A2. Assume that for $t \in [a, b]$, $\beta(t)$ is uniformly bounded on $[a, b]$ componentwise, that is, $\sup_{t\in[a,b]} |\beta(t)| \leq c < \infty$ componentwise.

A3. Assume that the covariates X_i are not degenerate, i.e., $\text{pr}(X_i = x_0) \neq 1$ and are bounded in probability, i.e., $\text{pr}(|X_i| < c) = 1$.

A4. Assume that the censoring times are bounded, i.e., $\text{pr}(C_i < c) = 1$.

A5. Assume that $E\big(Z_i Z_i^T \exp\{Z_i^T \beta(t)\}/[1 + \exp\{Z_i^T \beta(t)\}]^2\big)$ is positive definite.

The conditions A1-A2 control the size of the parameter space. The conditions A3-A4 exclude some degenerate cases. The condition A5 ensures a unique solution to the estimating equation. The following theorem establishes the consistency of the estimator $\widehat{\theta}_n$.

<u>Theorem</u> 1. *Assume that $\{O_i, i = 1, \cdots, n\}$ are i.i.d. random samples, and T_i and C_i are independent. Then under model (1.1) and assumptions A1-A5, $\widehat{\theta} \to \theta$ in probability as $n \to \infty$ for any $t_0 \in (a, b)$.*

The proof of this theorem is in Appendix A and it uses the semiparametric asymptotic results developed in Newey (1994) and Chen et al. (2003).

Since the final estimator involves estimates $\widehat{\beta}(\cdot)$ in the entire range of T_i, uniform consistency of the initial estimator is required. The next theorem establishes the asymptotic normality of $\widehat{\theta}_n$.

Theorem 2. *Under the assumptions of Theorem 1, as $n \to \infty$,*

$$\sqrt{n}(\widehat{\theta} - \theta) \to N(0, A^{-1}VA^{-1})$$

in distribution, where $A = E[\mu(X_i;\theta)\{1-\mu(X_i;\theta)\}Z_iZ_i^T], V = cov\{s(O_i;t_0,\theta,\beta) + \xi(T_i;t_0,\theta,\beta)\}$,

$$\begin{aligned}\xi(T_i;t_0,\theta,\beta) \;&=\; \int_0^{t_0} g(u)\int h(x)zz^T\Big[F(t_0|x)\{1-F(t_0|x)\}\psi(x,T_i;t_0,\theta) \\ &\quad -F(u|x)\{1-F(t_0|x)\}\psi\{x,T_i;u,\beta(u)\}\Big]dxdu,\end{aligned}$$

$g(u)$ is the density function for C_i, $h(x)$ is the density function for X_i, $z = (1,x)^T$, and $\psi\{x,T_i;u,\beta(u)\}$ is defined in the appendix.

The proof of this theorem is in Appendix A and it also uses the results in Newey (1994).

2.3.3 Simulation studies

To study the finite sample performance of the proposed estimator, we ran two sets of simulation studies. In each set, the true survival times were generated from the distribution (1.1) with $\beta_0(t) = \beta_{00} + \beta_{01}\log(t)$, $\beta_1(t) = \beta_{10} + \beta_{11}\log(t)$ and $(\beta_{00},\beta_{01},\beta_{10},\beta_{11})^T = (-80, 21.5, -1.4, 0.7)^T$. The parameters were designed such that the cumulative risk functions resembles the fit from COHORT data in section 2.3.4. We simulated X_i from a multinomial distribution with support on integer values between 41 and 50 representing CAG repeats. The censoring times were generated from a Beta distribution where the overall censoring rate is about 25%, similar to the COHORT data. We simulated two samples sizes $n = 1000$ and $n = 2000$ since the real data has a sample size of 1151.

We compared two estimators. The first is the initial inverse probability weighted estimator (IPW) $\widehat{\beta}^0(t)$ from (2.5) and the second is the proposed redistribution to the right weighted estimator (REW) $\widehat{\beta}(t)$ from (2.6). We summarized the numerical results in Tables 2.9 and 2.10, where we presented the estimated culumative distribution functions (CDFs) obtained from the two estimators at various ages and CAG values (42 and 46). It can be seen that both IPW and REW estimators have small finite sample biases. The estimated standard errors and empirical standard errors are close to each other over the entire age range. The empirical standard error of REW is smaller than that of IPW, especially at older ages. For example, the efficiency gain of REW over IPW is 10% at age 50 for CAG=42 and $n = 1000$. The empirical coverage of the 95% confidence interval is close to the nominal level when age is below 60 for both IPW and REW. At age 60, since censoring is heavier, the coverages of both IPW and REW are lower than the nominal level, with the performance of REW slightly better between the two.

We presented the true CDFs and the mean CDFs obtained from REW at various CAGs in Figure 2.6. The mean estimated distribution function coincides with the true function in most cases. When CAG=42 and $n = 1000$, there appears to be a small bias at the tail area for IPW estimator, for example, at $t = 65$ (bias=0.0055, SE=0.0009). However, this bias is within the variability range, which may be explained by the higher censoring rate within this range for subjects with CAG=42 (about 45%). When we increase the sample size to $n = 2000$, the bias decreased to almost zero.

In addition to the above estimators, we also investigated a smoothed REW estimator, where $\widehat{\beta}(t)$ were smoothed across the range of t subject to monotone constraint using a Generalized Pooled-Adjacent-Violators Algorithm (de Leeuw et al., 2009). The mean estimated cumulative distribution function and empirical standard error are almost identical to the non-smoothed estimator. The maximum absolute difference in the mean of the two estimators averaged across simulations was very small. Therefore we omit the results of the smoothed estimator here.

2.3.4 Application to COHORT data

As introduced in introduction, despite identification of the causative gene for HD, there is currently no effective treatment that delays HD onset or stops disease progression. To improve the care of HD patients and inform the development of effective treatment, a large genetic epidemiological study on HD, the Cooperative Huntington's Observational Research Trial (COHORT), was started in 1996. This is a study organized by 42 Huntington Study Group research centers in North America and Australia (Kieburtz and Huntington Study Group, 1996b; Dorsey et al., 2008). Participants in COHORT underwent a clinical evaluation where blood samples are genotyped for Huntingtin gene mutation and their CAG repeats lengths were obtained. Modeling the inverse association between the CAG repeats length and age-at-onset of HD accurately is important.

In this section, we fit the COHORT data by the model (1.1) where we do not assume a parametric form of $\beta_0(t)$ or $\beta_1(t)$. In our analysis, information on CAG repeats length, age at the time of evaluation, and age at diagnosis of HD onset (if a subject had been diagnosed) were available for 1151 subjects recruited in COHORT. In the study, both HD affected carriers and pre-symptomatic carriers were included. Their ages-at-first-motor-symptom were also recorded. Among 1151 subjects, 876 (76%) subjects had experienced HD motor sign onset and the average age of the diagnosis was 44 years of age.

To estimate the distribution of age-at-onset of HD given a subject's CAG repeats length, we fit three estimators: the initial IPW, REW, and Kaplan-Meier estimate using only subjects with a particular CAG repeats length. Figure 2.7 presents the estimated CDF at various CAG values. The results show a positive correlation between the onset probability and the CAG repeats, that is, the cumulative risk of HD onset by a given age increases with increasing number of CAG repeats. Subjects with longer CAG repeats have a higher probability of developing HD by a certain age, which is consistent with the literature (Langbehn et al., 2004). We summarize numerical results of estimated CDF at a few CAGs and age in Table 2.11. As a comparison, we see that IPW and REW provides point estimates of CDF similar to Kaplan Meier using only subjects with the same CAG values. However, the standard errors of REW at different age and CAGs are smaller than both Kaplan-Meier and IPW, suggesting efficiency gain. For example, at CAG=42 and age 50, the standard error of the cumulative risk estimated by IPW is 18% larger than REW and KM is 40% larger than REW. The post-hoc smoothing of $\widehat{\beta}(t)$ leads to CDF close to the non-smoothed CDF and therefore not reported here.

Table 2.9: Mean estimated IPW and REW estimators, their empirical SE, and 95% coverages. $n = 1000$, 400 simulations, CAG=42 and 46

						CAG=42					
Age	TRUE	IPW[‡]	SE(IPW)	SE(IPW)em	MSE(IPW)	REW[*]	SE(REW)	SE(REW)em	MSE(REW)	Cov(IPW)[$]	Cov(REW)
30	0.0001	0.0001	0.0007	0.0002	0.00000	0.0001	0.0007	0.0001	0.00000	0.8475	0.8425
35	0.0011	0.0012	0.0006	0.0006	0.00000	0.0011	0.0006	0.0006	0.00000	0.8950	0.8800
40	0.0143	0.0141	0.0043	0.0042	0.00002	0.0135	0.0041	0.0039	0.00002	0.9350	0.9200
45	0.1245	0.1250	0.0225	0.0223	0.00051	0.1224	0.0215	0.0213	0.00047	0.9525	0.9475
50	0.5233	0.5270	0.0469	0.0431	0.00221	0.5215	0.0400	0.0379	0.00160	0.9675	0.9600
55	0.8746	0.8905	0.3209	0.0408	0.10300	0.8845	0.0661	0.0330	0.00447	0.9900	0.9725
60	0.9742	0.9922	0.0780	0.0148	0.00640	0.9889	0.0644	0.0138	0.00436	0.2800	0.6675
						CAG=46					
Age	TRUE	IPW[‡]	SE(IPW)	SE(IPW)em	MSE(IPW)	REW[*]	SE(REW)	SE(REW)em	MSE(REW)	Cov(IPW)[$]	Cov(REW)
30	0.0027	0.0028	0.0121	0.0022	0.00015	0.0027	0.0114	0.0022	0.00013	0.9550	0.9500
35	0.0779	0.0783	0.0159	0.0161	0.00025	0.0764	0.0156	0.0158	0.00024	0.9375	0.9325
40	0.6208	0.6244	0.0356	0.0380	0.00128	0.6221	0.0348	0.0371	0.00121	0.9200	0.9275
45	0.9572	0.9567	0.0123	0.0121	0.00015	0.9570	0.0116	0.0116	0.00014	0.9550	0.9300
50	0.9957	0.9952	0.0029	0.0028	0.00001	0.9953	0.0026	0.0024	0.00001	0.8900	0.8950
55	0.9995	0.9991	0.0079	0.0013	0.00006	0.9992	0.0014	0.0010	0.00000	0.9300	0.8675
60	0.9999	0.9997	0.0053	0.0009	0.00003	0.9998	0.0051	0.0005	0.00003	0.2800	0.5850

[†]: Oracle estimator using true $F(t|x)$ to compute $w_i(F)$ in (2.3).

[‡]: Initial inverse probability weighting (IPW) estimator, $\widehat{\beta}_t^0$, solving (2.5).

[*]: Re-distributed to right (REW) weighted estimator, $\widehat{\beta}_t^W$, using IPW as initial estimator to solve (2.6).

[$]: 95% coverage probability of IPW estimator.

Table 2.10: Mean estimated IPW and REW estimators, their empirical SE, and 95% coverages. $n = 2000$, 400 simulations, CAG=42 and 46

						CAG=42					
Age	TRUE	IPW‡	SE(IPW)	SE(IPW)em	MSE(IPW)	REW*	SE(REW)	SE(REW)em	MSE(REW)	Cov(IPW)$	Cov(REW)
30	0.0001	0.0001	0.0001	0.0001	0.00000	0.0001	0.0001	0.0001	0.00000	0.8175	0.8050
35	0.0011	0.0011	0.0004	0.0004	0.00000	0.0011	0.0004	0.0004	0.00000	0.9475	0.9225
40	0.0143	0.0142	0.0031	0.0028	0.00001	0.0136	0.0029	0.0027	0.00001	0.9650	0.9275
45	0.1245	0.1248	0.0156	0.0154	0.00024	0.1225	0.0151	0.0153	0.00023	0.9575	0.9425
50	0.5233	0.5243	0.0305	0.0267	0.00093	0.5210	0.0272	0.0252	0.00075	0.9700	0.9725
55	0.8746	0.8789	0.0538	0.0245	0.00292	0.8762	0.0252	0.0211	0.00064	0.9800	0.9700
60	0.9742	0.9851	0.1265	0.0146	0.01610	0.9828	0.0695	0.0130	0.00491	0.6450	0.8850
						CAG=46					
Age	TRUE	IPW‡	SE(IPW)	SE(IPW)em	MSE(IPW)	REW*	SE(REW)	SE(REW)em	MSE(REW)	Cov(IPW)$	Cov(REW)
30	0.0027	0.0027	0.0025	0.0016	0.00001	0.0026	0.0024	0.0015	0.00001	0.9125	0.9000
35	0.0779	0.0784	0.0111	0.0104	0.00012	0.0764	0.0109	0.0102	0.00012	0.9525	0.9500
40	0.6208	0.6229	0.0247	0.0248	0.00062	0.6212	0.0242	0.0240	0.00058	0.9350	0.9500
45	0.9572	0.9565	0.0087	0.0087	0.00008	0.9566	0.0082	0.0083	0.00007	0.9550	0.9600
50	0.9957	0.9954	0.0020	0.0020	0.00000	0.9954	0.0018	0.0018	0.00000	0.9125	0.9225
55	0.9995	0.9993	0.0012	0.0007	0.00000	0.9993	0.0006	0.0006	0.00000	0.8925	0.8950
60	0.9999	0.9997	0.0059	0.0008	0.00004	0.9998	0.0036	0.0005	0.00001	0.6250	0.7500

†: Oracle estimator using true $F(t|x)$ to compute $w_i(F)$ in (2.3).

‡: IPW estimator, $\widehat{\beta}_t^0$, solving (2.5).

*: Re-distributed to right weighted estimator, $\widehat{\beta}_t^W$, using IPW as initial estimator to solve (2.6).

$: 95% coverage probability of IPW estimator.

Table 2.11: COHORT data: Estimated KM, IPW and REW estimators and their estimated SE for age at diagnosis at CAG=42, 44, 46 and 48.

CAG=42

Age	KM	SE(KM)	IPW‡	SE(IPW)	REW*	SE(REW)
30	0.0000	NA	0.0042	0.0017	0.0032	0.0014
35	0.0101	0.0071	0.0120	0.0033	0.0103	0.0030
40	0.0421	0.0146	0.0392	0.0070	0.0342	0.0063
45	0.1052	0.0229	0.0952	0.0135	0.0875	0.0123
50	0.2677	0.0344	0.2865	0.0242	0.2429	0.0199
55	0.5337	0.0402	0.5467	0.0358	0.5009	0.0257
60	0.7429	0.0369	0.7874	0.3009	0.7466	0.0234

CAG=44

Age	KM	SE(KM)	IPW	SE(IPW)	REW	SE(REW)
30	0.0246	0.0121	0.0135	0.0039	0.0105	0.0034
35	0.0566	0.0183	0.0475	0.0083	0.0375	0.0073
40	0.1564	0.0294	0.1591	0.0148	0.1367	0.0135
45	0.3649	0.0398	0.3750	0.0214	0.3412	0.0196
50	0.7322	0.0377	0.6859	0.0226	0.6482	0.0225
55	0.9184	0.0244	0.8840	0.0174	0.8684	0.0173
60	0.9674	0.0160	0.9601	0.0644	0.9544	0.0108

CAG=46

Age	KM	SE(KM)	IPW	SE(IPW)	REW	SE(REW)
30	0.0410	0.0201	0.0420	0.0087	0.0342	0.0078
35	0.1522	0.0375	0.1705	0.0184	0.1273	0.0156
40	0.4233	0.0523	0.4677	0.0286	0.4147	0.0265
45	0.8230	0.0413	0.7739	0.0278	0.7366	0.0284
50	0.9183	0.0310	0.9224	0.0142	0.9136	0.0149
55	0.9387	0.0292	0.9796	0.0063	0.9775	0.0064
60	0.9796	0.0193	0.9936	0.0101	0.9933	0.0030

CAG=48

Age	KM	SE(KM)	IPW	SE(IPW)	REW	SE(REW)
30	0.1389	0.0576	0.1231	0.0208	0.1053	0.0188
35	0.4783	0.0841	0.4585	0.0420	0.3535	0.0359
40	0.8344	0.0647	0.8031	0.0306	0.7603	0.0334
45	0.9337	0.0446	0.9513	0.0127	0.9379	0.0149
50	0.9669	0.0323	0.9848	0.0047	0.9838	0.0048
55	1.0000	NA	0.9967	0.0016	0.9965	0.0015
60	1.0000	NA	0.9990	0.0015	0.9991	0.0006

‡: IPW estimator, $\widehat{\beta}_t^0$, solving (2.5).

*: Re-distributed to right weighted estimator, $\widehat{\beta}_t^W$, using IPW as initial estimator to solve (2.6).

Figure 2.6: True and REW estimated cumulative distribution curves at CAG=50, 48, 46, 44, 42 (left to right). The mean and true cumulative distribution curves are indistinguishable in IPW and REW estimators for most cases. $n = 1000$ (top) and $n = 2000$ (bottom), 400 simulations.

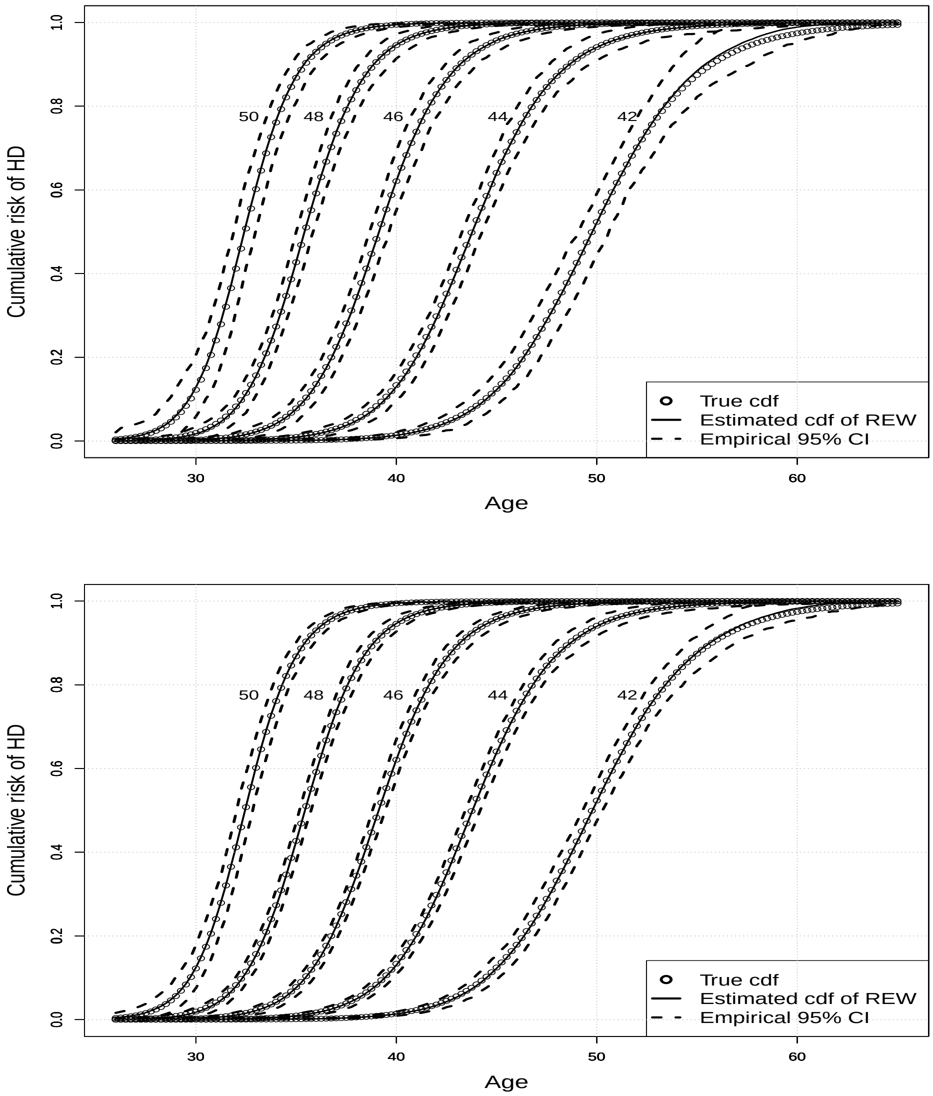
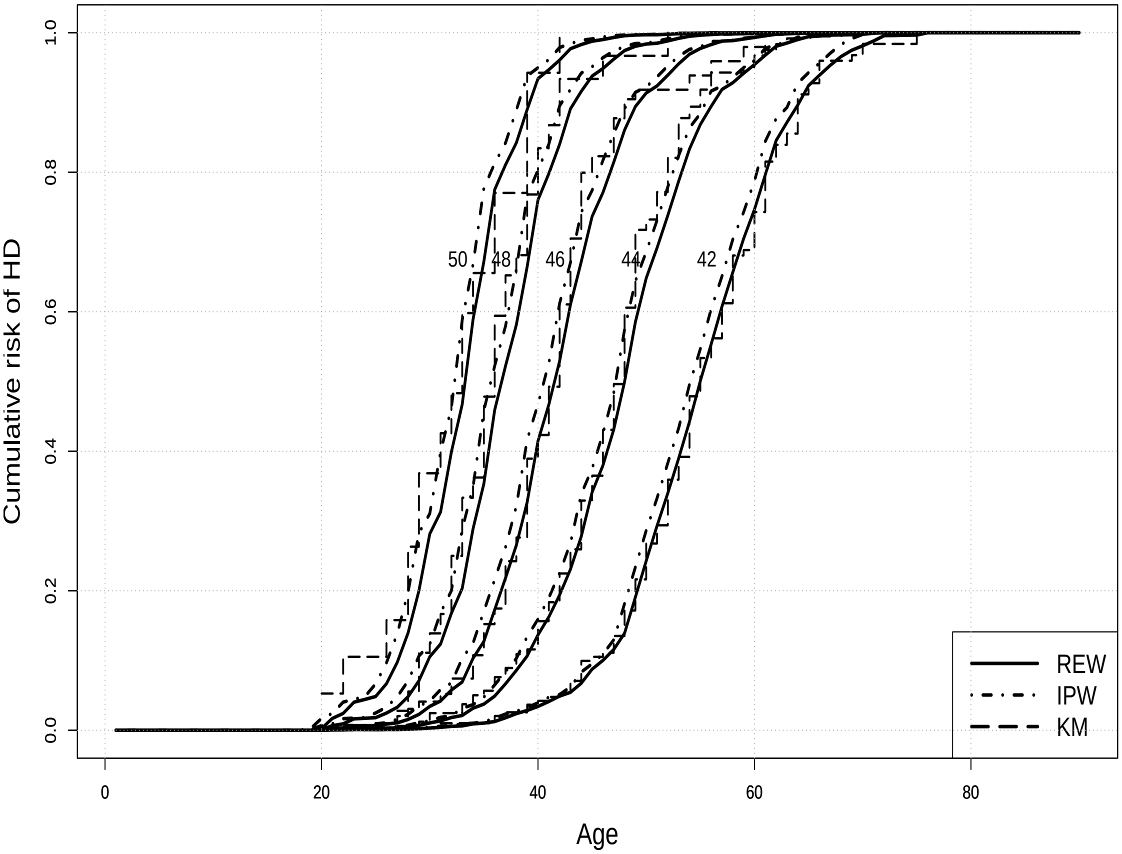


Figure 2.7: Estimated cumulative distribution curves (KM, IPW and REW) with COHORT proband data (n = 1151) evaluated at CAG=50, 48, 46, 44, 42 (left to right).



2.4 Discussion

We propose methods to predict disease risk from a known mutation (or to estimate the penetrance function). For most complex diseases, predicting the AAO of a disease from genetic markers such as single-nucleiotide polymorphisms (SNPs) continue to be a challenging issue (Kanga et al., 2010). Even with diseases like HD where the gene is identified, the predictive model can be complicated: a special feature of the HD data is that the mutation is defined as a continuous variable (CAG repeats) instead of a categorical variable, as it is in most genome-wide association studies.

One of the contributions of this work is to use the family data as well as the proband data to maximize available information in building a model. Our results reveal that the estimated risk obtained from the combined proband and family data is slightly lower than the risk estimated from the proband data alone. It is possible that the proband data consists of a biased clinical sample of gene positive or HD affected subjects, and is therefore not a fair representative sample of the entire HD population, especially under-representing subjects at-risk. The family data may be a better representative of the population since the family members are included in the analysis only through the inclusion of the probands. Although proband may participate the study because they had HD or they had more servere symptoms of HD, the relatives were not included based on their CAG repeat length or affection status. Of course, some of the family members will not receive an allele with expanded CAG repeats from the probands and therefore are non-carriers who will never develop HD. The expected number of carriers in the combined data is 2601 and the proportion of affecteds (n=1496) out of the expected carriers 58%, which is lower than the proportion of affected subjects in the proband data.

Note that our estimated cumulative risk of onset of a positive HD diagnosis in the proband data is also slightly lower than Langbehn et al. (2004) which also examined age-at-HD-diagnosis. We observe later mean AAO for each CAG repeat length for COHORT data than Langbehn et al. (2004). For example, the mean AAO of HD diagnosis for probands with a CAG of 42 in the former data was 3 years later and for a CAG of 43, it was 4 years later (Table 2.5). On average, for all subjects with a CAG between 41 and 50, the mean AAO in Langbehn data was 2 years earlier than in the COHORT data. This is consistent with the prediction from the model shown in Table 2.7. Another possible explanation for the difference is that the CAG repeat lengths in Langbehn study were measured in different laboratories while in the COHORT they were all measured in the same lab.

Here, we assumed Mendelian transmission of the mutation without interference so that the CAG length does not change from parents to offspring. There are several possible violations of these assumptions. For example, an extremely elongated HD mutation (CAG $>>$ 56) may occasionally cause miscarriages, perhaps even in very early development before a mother necessarily knows that she was pregnant. Therefore, the transmission probability from a mother with extremely high CAG to a child can be less than 0.5. Another possible violation of Mendelian law is that those inheriting the gene from their father may have a higher probability of longer CAG repeat length than their father. The probability of this occurring is much lower if inheritance is from the mother. An explanation is that there are many more biological opportunities for the CAG length to change in the father's process of sperm formation than in the mother's process of egg formation. Under normal conditions, the CAG length does not change, but there is a slightly higher probability that the CAG repeat length will increase (expand) rather than decrease at each generation of new germ cells. However, there are no reasonable dynamic population genetics models for these effects and appropriate assumptions are rather complicated.

Consistent with Langbehn et al. (2004) and other studies (Brinkman et al., 1997; McNeil et al., 1997), we estimated reduced penetrance for lower CAG repeat lengths (≤ 40). We point out that the parameter estimates from the current model do not include subjects with CAG less than 41, therefore the risk estimates for these subjects are extrapolations. However, it is conceivable that as long as the inverse relationship between AAO and CAG still holds for the lower CAGs, the life time disease risk for these subjects will be less than 100%, since the life time risk for a CAG of 41 is about 100%.

Although the CAG repeat explains the majority (about 75%) of variability in AAO of HD, there are other potential variables contributing to the distribution of AAO. Our current model does not include other observed covariates, such as gender, nor does it account for unobserved residual familial aggregation, aside from sharing HD mutation in a family. Future research will focus on incorporating observed covariates and adding family-specific random effects to account for residual familial aggregation.

We also propose methods to estimate cumulative disease risk from a nonparametric varying-coefficient model. The proposed method explores a pseudo-logistic regression and redistributes the probability mass at the censored outcomes to the right. The procedure has desirable numerical and asymptotic properties and is extremely easy to implement. Although we focused on assessing the effect of CAG repeats on HD onset, it is easy to include other covariates with time-invariant effect through a backfitting procedure for models such as

$$\text{logit}\{\text{pr}(T_i \leq t|X_i)\} = \beta_0(t) + \beta_1(t)X_i + \gamma^T Z_i.$$

The proposed methods have computational advantages compared to, for example, Peng and Huang (2007). In addition to the logistic link as discussed here, the developed methods can be adapted to transformation models with a known link function.

Chapter 3

Targeted local kernel support vector machine for age-dependent classification and prediction

3.1 Overview

In this chapter, we propose statistical learning methods for age-dependent disease classification and prediction. In the first section, we propose a targeted local support vector machine. In the second section, we show the asymptotic properties. In the third section, we conduct two simulation studies to investigate performance of the proposed methods. In the fourth section, we apply the method to two study data sets on Huntington's disease, the COHORT data and the PREDICT-HD data. Finally, we summarize our findings and discuss possible extensions in the fifth section.

3.2 Targeted local smoothing for large margin classifiers

Let D be the dichotomous at-risk status coded as 1 and -1 for subjects at risk of a disease and not at risk, respectively. Let W denote a subject's age and let $\boldsymbol{X}$ be a vector of the other potential risk-altering markers for this subject. The goal is to determine an age-dependent classification rule using $\boldsymbol{X}$ to predict D at each age W (the target variable). For this purpose, we first consider the following composite predictive score

$$\alpha(W) + \boldsymbol{X}^T\boldsymbol{\beta}(W), \tag{3.1}$$

where $\alpha(W)$ is an unspecified baseline function, and $\boldsymbol{\beta}(W)$ is a vector of unspecified age-dependent coefficients for markers $\boldsymbol{X}$. A subject with a positive fitted score will be classified as at risk of disease, and as risk free if the subject has a negative fitted score. Note the score in model (3.1) has a nonparametric form with respect to age effect, while at each given age it is linear in terms of markers $\boldsymbol{X}$. This formulation allows decomposition of the diagnostic score as the sum of a component due to normal aging, $\alpha(W)$, and a component due to the other markers, $\boldsymbol{X}^T\boldsymbol{\beta}(W)$. The unrestricted form of $\boldsymbol{\beta}(w)$ allows the age-dependent effect to change freely. Since age may serve as a surrogate for many unmeasured physiological factors, for subjects close in age and with the same values of other markers, the disease risk is expected to be similar, and thus certain smoothness is expected for functions $\alpha(w)$ and $\boldsymbol{\beta}(w)$.

The age-dependent classification boundary in (3.1) has several features. First, although the score is allowed to change from one age to another in an unspecified fashion, at a given age the prediction is a linear combination of markers to facilitate interpretation. It is easy to tell which markers are effective at which age by examining coefficient functions $\boldsymbol{\beta}(w)$. When varying age smoothly, the corresponding separating hyperplane constructed from other markers also changes smoothly. Second, since the coefficient function $\boldsymbol{\beta}(w)$ is age-adaptive, it captures the age-dependent effects of markers. As introduced later in section 3.5, there might be markers informative for younger subjects but not older subjects or vice versa, which suggests different sets of markers would be considered as effective depending on a subject's age. Third, some cumulative summary of $\boldsymbol{\beta}(w)$, for example, the vector $\int |\boldsymbol{\beta}(w)|dw$, can be used to rank the overall importance of markers under model (3.1).

In a standard classification problem with predictive score $\alpha + \boldsymbol{X}^T\boldsymbol{\beta}$, a large-margin based classifier would minimize a penalized loss function,

$$\min_{\alpha,\boldsymbol{\beta}} \sum_i \mathcal{L}\{D_i, \boldsymbol{X}_i; \alpha, \boldsymbol{\beta}\} + \lambda_n ||\boldsymbol{\beta}||^2,$$

where λ_n is a tuning parameter depending on the sample size, and $\mathcal{L}(\cdot)$ belongs to a class of margin-based loss functions. Examples of margin-based loss functions include hinge loss, i.e., SVM loss, $\{1 - df(\boldsymbol{x})\}_+$; its variations such as $\psi-$loss which satisfies $U \geq \psi(z) > 0$ where $z = df(\boldsymbol{x})$, if $z \in [0, \tau]$; $\psi(z) = U(1 - \text{sign}(z))$, otherwise for some constants U and $0 < \tau < 1$ (Shen et al., 2003); and logistic loss, $\log\{1+\exp(-df(\boldsymbol{x}))\}$. To fit the age-dependent predictive score in model (3.1) taking advantage of the smoothness effect in age, we introduce a local smoothing kernel weighted support vector machine (KSVM). Essentially, the KSVM solves an SVM at each w_0 where the ith subject is weighted by a local smoothing kernel function $K_{h_n}(W_i - w_0)$, so we pool information across subjects whose ages are close to w_0. Here $K_{h_n}(\cdot)$ is a symmetric kernel density and h_n is its bandwidth. Specifically, we fit (3.1) by solving

$$\min_{\alpha(w_0),\boldsymbol{\beta}(w_0)} \sum_i K_{h_n}(W_i - w_0)\mathcal{L}\{D_i, \boldsymbol{X}_i; \alpha(w_0), \boldsymbol{\beta}(w_0)\} + \lambda_n ||\boldsymbol{\beta}(w_0)||^2, \tag{3.2}$$

where w_0 varies across the support of age W_i. The loss function in the minimization problem (3.2) can be considered as a locally weighted loss where the subjects closer to age w_0 contribute larger weights.

In the subsequent implementation of KSVM, we choose the hinge loss. Computationally, the optimization problem is solved by

$$\min_{\alpha(w_0),\boldsymbol{\beta}(w_0)} \sum_i K_{h_n}(W_i - w_0)\xi_i + \lambda_n \|\boldsymbol{\beta}(w_0)\|^2,$$
$$\text{subject to } \; D_i\{\alpha(w_0) + \boldsymbol{X}_i^T \boldsymbol{\beta}(w_0)\} \geq 1 - \xi_i, \quad \xi_i \geq 0.$$

This alternative form provides some insights to the locally weighted objective function (3.2). Treating the slack variables ξ_i as serving similar roles as residuals in a regression model, problem (3.2) can be thought as minimizing a penalized locally weighted "residual" subject to linear constraints. Using the Lagrange multipliers, we can derive the corresponding dual form as

$$\max_{\boldsymbol{\gamma}\in\mathbb{R}^n} \sum_i \gamma_i - \frac{1}{2}\sum_{i,j} \gamma_i\gamma_j D_i D_j \boldsymbol{X}_i^T \boldsymbol{X}_j,$$
$$\text{subject to } 0 \leq \gamma_i \leq K_{h_n}(W_i - w_0)C_n, \;\text{ and } \; \sum_i \gamma_i D_i = 0.$$

Note that by reparametrizing γ_i as $\gamma_i K_{h_n}(W_i - w_0)$, the dual form is equivalent to

$$\max_{\boldsymbol{\gamma}\in\mathbb{R}^n} \sum_i \gamma_i K_{h_n}(W_i - w_0) - \frac{1}{2}\sum_{i,j} \gamma_i\gamma_j D_i D_j K_{h_n}(W_i - w_0) K_{h_n}(W_j - w_0) \boldsymbol{X}_i^T \boldsymbol{X}_j,$$
$$\text{subject to } 0 \leq \gamma_i \leq C_n, \;\text{ and } \; \sum_i \gamma_i K_{h_n}(W_i - w_0) D_i = 0. \tag{3.3}$$

This is a locally weighted quadratic programming problem with linear constraints which can be solved conveniently using existing quadratic programming packages in R or MatLab. The resulting prediction of disease status for a w-year-old subject with markers $\boldsymbol{x}$ is

$$\widehat{d}(\boldsymbol{x}, w) = \text{sign}\{\widehat{f}(\boldsymbol{x}; w)\}, \quad \widehat{f}(\boldsymbol{x}; w) = \widehat{\alpha}(w) + \boldsymbol{x}^T \widehat{\boldsymbol{\beta}}(w). \tag{3.4}$$

When at a given age the disease risk groups cannot be adequately separated by a linear function of marker, it may be useful to perform prediction in the reproducing kernel Hilbert space (RKHS, Wahba, 1990) feature space instead of the original marker space. Consider a nonparametric predictive score, $f(\boldsymbol{X}_i; w_i)$, which is a completely unspecified function of age and markers. The age-dependent decision boundary (3.1) corresponds to a special case of taking a linear combination of all components of $\boldsymbol{X}_i$ at each age point w, i.e., $\alpha(w_i) + \boldsymbol{X}_i^T \boldsymbol{\beta}(w_i)$. The nonlinear classification boundary relaxes the linear form (in terms of markers) at each age. To fit this nonlinear predictive score, we smooth age effect by a local smoothing kernel while mapping other markers to a RKHS feature space through Mercer kernels. To be specific, denote a Mercer kernel $H(\boldsymbol{x}, \boldsymbol{y})$, through an appropriate inner product in the RKHS. Commonly used Mercer kernels include the Gaussian kernel, where $H(\boldsymbol{x}, \boldsymbol{y}) = \exp(-\gamma \|\boldsymbol{x} - \boldsymbol{y}\|^2)$, and the kth order polynomial kernel, where $H(\boldsymbol{x}, \boldsymbol{y}) = (1 + \boldsymbol{x}^T \boldsymbol{y})^k$. At a given age w_0, the general decision boundary can be expressed as a function in the RKHS associated with $H(\cdot, \cdot)$ as

$$f(\boldsymbol{x}; w_0) = \eta_0(w_0) + \sum_i \eta_i(w_0) H(\boldsymbol{X}_i, \boldsymbol{x}).$$

Comparing with the age-dependent model in (3.1), we see the methodology developed there can be implemented similarly. To pool information from subjects with similar ages, we use local smoother to weight observations around w_0, and the resulting local optimization problem is

$$\min_{\eta_0(w_0), \boldsymbol{\theta}(w_0)} \sum_i K_{h_n}(w_i - w_0)\{1 - D_i[\eta_0(w_0) + \sum_j \eta_j(w_0) H(\boldsymbol{X}_j, \boldsymbol{X}_i)]\}_+ + \lambda_n ||f(\cdot; w_0)||^2_{\mathcal{H}},$$

where $\|f\|_{\mathcal{H}}$ is the norm of f in the RKHS. This locally weighted problem is solved in the dual space by replacing $\boldsymbol{X}_i^T \boldsymbol{X}_j$ in (3.3) by $H(\boldsymbol{X}_i, \boldsymbol{X}_j)$ associated the RKHS. The predicted at-risk status for a subject with marker $\boldsymbol{x}$ at age w using a fully nonparametric boundary is

$$\widehat{d}(w, \boldsymbol{x}) = \text{sign}\{\widehat{f}(\boldsymbol{x}; w)\}.$$

Note the distinct roles of the smoothing kernel $K_{h_n}(\cdot)$ and Mercer kernel $H(\cdot, \cdot)$: the former is used to pool information across age and the later for producing nonlinear decision boundary and dimension reduction with respect to the markers $\boldsymbol{X}$. The tuning parameters h_n and λ_n are chosen over a grid in a range, respectively, by minimizing the five-fold cross validated misclassification error. By using a different kernel and a separate tuning parameter for age, the age effect can be better accommodated. In summary, the proposed method can be viewed as a splice of local smoothing and the RKHS framework for the SVM.

3.3 Theoretical results

In this section, we provide general theoretical results for the prediction errors using the fitted rule $\widehat{f}(\boldsymbol{X}; w)$ as compared to the true optimal rule based on $f_0(\boldsymbol{X}; w) = 2P(D = 1|\boldsymbol{X}, W = w) - 1$. Our results require the following assumptions:

(C.1) Markers $(\boldsymbol{X}, W)$ have a bounded support and the conditional density of $(D, \boldsymbol{X})$ given $W = w$ and is twice-continuously differentiable with respect to w. Moreover, the marginal density for W is twice-continuously differentiable and bounded away from zero;

(C.2) The conditional distribution of $P(\boldsymbol{X}|W = w)$ has a uniform geometric noise exponent $\alpha > 0$; that is, there exists a constant C independent of w such that

$$\int |f_0(\boldsymbol{x}; w)| \exp\left\{-\frac{\tau_{\boldsymbol{x}}(w)^2}{t}\right\} dP(\boldsymbol{x}|w) \leq Ct^{\alpha d/2},$$

where d is the dimension of $\boldsymbol{X}$, and $\tau_{\boldsymbol{x}}(w)$ is the minimum distance from $\boldsymbol{x}$ to set $\{\boldsymbol{z} : f_0(\boldsymbol{z}; w) \leq 0\}$ for $\boldsymbol{x}$ with $f_0(\boldsymbol{x}; w) > 0$ while it is the minimum distance from $\boldsymbol{x}$ to set $\{\boldsymbol{z} : f_0(\boldsymbol{z}; w) \geq 0\}$ for $\boldsymbol{x}$ with $f_0(\boldsymbol{x}; w) < 0$;

(C.3) The kernel function $K_{h_n}(x) = h_n^{-1}K(x/h_n)$, where $K(\cdot)$ is symmetric and has finite second moments. The reproducing kernel Hilbert space used to fit the general decision boundary in (3.4) is generated from a Gaussian kernel with the bandwidth σ_n^{-1}.

(C.4) $h_n, \lambda_n \to 0$, $\sigma_n = \lambda_n^{-1/((\alpha+1)d)}$ and $\sqrt{n}h_n^2 \to \infty$.

Condition (C.1) ensures the smoothness of the distribution of the data over age W, so that we can borrow neighboring information to infer an age-dependent rule. Condition (C.2) is given in Steinwart and Scovel (2007), where they discussed a list of examples that satisfy the geometric noise exponent condition. In particular, if the distribution satisfies that $|f_0(\boldsymbol{x}; w)| \leq c\tau_{\boldsymbol{x}}(w)^{\gamma_1}$ and $P(|f_0(\boldsymbol{x}; w)| \leq t|W = w) \leq Ct^q$ (Tsybakov noise exponent q), then condition (C.2) holds for $\alpha = (q+1)\gamma_1/d$ if $q \geq 1$. In condition (C.4), as indicated in the proof and also in Steinwart and Scovel (2007), the choice of σ_n is optimal in terms of approximating the Bayesian error bound using the decision function for the reproducing kernel Hilbert space. Our main theoretical result is the following.

Theorem 1. Define $Err(f; w)$ as the prediction error at age w, i.e., $Err(f; w) = P(Df(\boldsymbol{X}; w) < 0|W = w)$. Under conditions (C.1)-(C.4), there exists a constant c_d such that for any $t > t_0$ where t_0 is a constant that depends on d, with probability at least $1 - e^{-t}$, it holds

$$\sup_{w \in \mathcal{W}} \left\{ \left| Err(\widehat{f}; w) - Err(f_0; w) \right| \right\} \leq c_d(h_n^2 \lambda_n^{-1} + \lambda_n^{\alpha/(\alpha+1)} + r_n t),$$

where $r_n = n^{-1/2} h_n^{-2} \lambda_n^{-1-(d+2)/[(\alpha+1)d]}$ and is assumed to vanish as n goes to infinity.

Note the rate of risk bound is characterized through the geometric noise exponent α, local kernel smoothing parameter h_n, and the regularization parameter λ_n for SVM. In addition, we obtain the supreme norm risk bound over the support of age. The proof of Theorem 1 uses the embedding properties of the reproducing kernel Hilbert space, the large deviation results of empirical processes and the approximation using the kernel function. In the proof, we first note that $Err(\widehat{f}; w) - Err(f_0; w)$ can be bounded by the corresponding risk based on the hinge loss $E[(1 - D\widehat{f})_+|W = w] - E[(1 - Df_0)_+|W = w]$. We then decompose the latter into

$$\begin{aligned} E[(1 - D\hat{f})_+|W = w] \quad &- \quad E[(1 - D\widehat{f})_+ K_{h_n}(W - w)]/f_W(w) \\ -\Big\{ E[(1 - Df_0)_+|W = w] &- E[(1 - Df_0)_+ K_{h_n}(W - w)]/f_W(w) \Big\} \end{aligned}$$

and

$$\Big\{ E[(1 - D\hat{f})_+ K_{h_n}(W - w)] - E[(1 - Df_0)_+ K_{h_n}(W - w)] \Big\} / f_W(w),$$

where f_W is the marginal density of W. Note that the first part is the bias due to the kernel smoothing so can be controlled using the kernel bandwidth. The latter part is a weighted version of the hinge loss; therefore, we will adapt the existing theory for the support vector machine (Steinwart and Scovel, 2007) but with careful modification due to the local smoothing kernel weights. The main challenge is to control the complexity of the kernel weighted functions from the reproducing kernel Hilbert space and assess the tail bound of some kernel weighted empirical processes. The detail of the proof is given in Appendix B.

From Theorem 1, we conclude

$$\sup_{w\in\mathcal{W}}\left\{\left|Err(\widehat{f};w)-Err(f_0;w)\right|\right\}=O(h_n^2/\lambda_n+\lambda_n^{\alpha/(\alpha+1)})+O_p(r_n).$$

Therefore, the optimal h_n is $[n^{-1/2}\lambda_n^{-1-(d+2)/[(\alpha+1)d]}]^{1/4}$ and the derived rate becomes

$$O_p(\lambda_n^{\alpha/(\alpha+1)}+[n^{-1/2}\lambda_n^{-1-(d+2)/[(\alpha+1)d]}]^{1/2}/\lambda_n).$$

This further gives the optimal choice of λ_n to be $\lambda_n^{opt}=n^{-\gamma}$ where $\gamma=1/[6+2(d+2)/[(\alpha+1)d]+4\alpha/(\alpha+1)]$ so it results in the optimal rate as

$$\sup_{w\in\mathcal{W}}\left\{\left|Err(\widehat{f};w)-Err(f_0;w)\right|\right\}=O_p(n^{-\gamma\alpha/(\alpha+1)}).$$

Clearly, these optimal rates depend on the unknown α, so they cannot be estimated. Instead, we suggest using the cross-validation to estimate the optimal choices of (h_n,λ_n) in practice. Under the special case when $f_0(\boldsymbol{x};w)=\boldsymbol{X}^T\boldsymbol{\beta}_0(w)$, if we choose $h_n^4/\lambda_n=n^{-1/2}$, then Theorem 1 can be modified to obtain

$$\sup_{w\in\mathcal{W}}|Err(\widehat{f};w)-Err(f_0;w)|=O_p(n^{-1/4}).$$

This rate gives an supreme bound of the classification error over the range of age when the underlying true classification boundary is linear.

3.4 Simulation studies

In this section, we conducted two sets of simulation studies to compare the empirical performance of KSVM with several alternatives. We generated samples with a size of $n = 500$ or 1000. For each setting we carried out 200 simulation runs. The standardized ages W_i were generated from a uniform distribution with support $(0, 1)$. In the first set of experiments, we simulated data retrospectively. We generated dichotomous outcomes

$$Y_i = \mathrm{sign}(W_i^2 + W_i - 1 + \epsilon_i), \quad \epsilon_i \sim N(0, 1),$$

and given Y_i and W_i, we generated markers $\boldsymbol{X}_i = (X_{i1}, X_{i2})^T$ as

$$\boldsymbol{X}_i | Y_i, W_i \sim MVN\{\boldsymbol{\beta}(W_i)Y_i, \sigma^2\mathbf{I}\},$$

and $\boldsymbol{\beta}(w) = (\sin(4\pi w), 2\exp\{-20(w - 0.5)^2\})^T$.

We compared several alternative methods of handling age and other markers. For the handling of age effect we compared three approaches: (1) Using $\boldsymbol{X}_i$ but no W_i to train a standard SVM (SVM$_0$); (2) Using $\boldsymbol{X}_i$, W_i and $\boldsymbol{X}_i W_i$ as input variables to train a standard SVM (SVM$_1$); and (3) the proposed local smoothing SVM (KSVM). Within each of these methods, we compared using a linear kernel for input variables versus using a Gaussian kernel. To evaluate the performance of different approaches, we recorded the misclassification rate and area under the receiver operating characteristic (ROC) curve (AUC) at each age point, and computed an overall AUC and mean misclassification rate pooling data across all age points. For KSVM, the bandwidth h_n and the tuning parameter λ_n were chosen by 5-fold cross validation separately. For the multiple marker case, we included both markers X_{i1} and X_{i2}, and two other noise markers that do not contribute to disease risk.

Table 3.1 records the mean overall misclassification rate and AUC averaged over simulations for SVM$_0$, SVM$_1$ and KSVM with two choices of Mercer kernels when using X_{i1} alone, using X_{i2} alone, or using both plus two noise markers generated from a standard uniform distribution. From Table 3.1, when a single marker is used and the true classification boundary is more complex, such as a sine function, the locally weighted KSVM has much lower average misclassification rate and much higher overall AUC than fitting SVM$_0$ or SVM$_1$. Using a Gaussian kernel improves overall performance for SVM_1 and KSVM but not for SVM_0. As expected, larger sample size improves AUC and decreases misclassification rate. When the underlying separating boundary is a simpler function such as a Gaussian function, the difference between KSVM and SVM_0 is still substantial while the difference between KSVM and SVM_1 is smaller. KSVM performs better than both SVM_0 and SVM_1 when a linear Mercer kernel is used. With a Gaussian kernel, the overall performance of SVM_1 and KSVM is similar due to the true coefficient function $\beta(w)$ being Gaussian (nonlinear) and the ability of Gaussian Mercer kernel to fit nonlinear separating boundaries. Furthermore, from this table, we observe that using all the markers compared to using single markers improves the prediction accuracy. In this case, comparing three approaches in treating the age effect, KSVM still has the overall performance superior to SVM_0 or SVM_1. The decrease in misclassification rate of KSVM over alternatives averaged across age and simulations is up to 50% (0.133 versus 0.265), and the increase in AUC is up to 13% (0.941 versus 0.817), which is substantial. Comparing different treatment of Mercer kernels, using a Gaussian kernel does not improve performance of either SVM_0, SVM_1 or KSVM.

Figure 3.1 presents more detailed information on the age-specific misclassification rate and AUC as a function of w averaged across simulation repetitions. When the true coefficient function is a sine function, KSVM dominates the alternatives over the entire range of w: it has lower misclassification rate and higher AUC at each age. For Gaussian coefficient function, KSVM improves upon SVM_0 and SVM_1 at the tail area. For the multiple marker case, Figure 3.1 (bottom panels) shows that while the age-specific AUC and misclassification rate indicates a superior performance of KSVM over the alternatives in the entire range of age, the improvement is much more significant at the tail area and at places where the two classes have large overlap. For example, SVM_1 fails to accommodate the decision boundary around about $w = 0.15$ and $w = 0.85$ (high misclassification rate and low AUC) as shown by two subfigures.

In the second set of simulations, we simulated data prospectively based on a known true decision boundary thus we could assess the performance of the fitted decision boundary through its mean squared error. First, we generated the standardized ages W_i from a uniform distribution with support (0,1). The markers $\boldsymbol{X}_i = (X_{i1}, X_{i2})^T$ are generated as $\boldsymbol{\beta}(W_i) + \boldsymbol{\epsilon}_i$, where $\boldsymbol{\epsilon}_i$ follows $MVN(0, 1.5^2\boldsymbol{I})$ and the true $\boldsymbol{\beta}$'s are the same as in the first set of simulations. We further considered three different scenarios where $Z_{i1} = X_{i1} - \beta_1(W_i)$, $Z_{i2} = X_{i2} - \beta_2(W_i)$, and $Z_{i3} = Z_{i1} + Z_{i2}$, and the true margin had a width of $\delta = 0.3$. Then the class labels were generated as

$$Y_{ik} = \begin{cases} \text{1; if } Z_{ik} > \delta \\ \text{-1; if } Z_{ik} < -\delta \\ \text{1 or -1 with probability of 0.5; otherwise,} \end{cases}$$

for $k = 1, 2, 3$. We show a scatter plot of data generated in a typical simulation and the true discriminant boundary which depends on the age in Figure 3.2.

We computed the mean squared error (MSE) of the fitted classification boundary averaged across age for SVM$_0$, SVM$_1$ and KSVM. When a linear Mercer kernel is used and with a sample size of 500, the MSE of KSVM under sine or Gaussian coefficient function is much smaller than either SVM$_0$ or SVM$_1$: for the sine coefficient, MSE(×100)=49.8, 43.5, 6.24, respectively for SVM$_0$, SVM$_1$ and KSVM; for the Gaussian coefficient, MSE(×100)=50.7, 51.4, 2.59, respectively. This reflects the inflexibility of SVM$_1$ in fitting nonlinear age boundaries. When we increase the sample size to 1000, the bias in SVM$_1$ persists for all three scenarios. In Table 3.2, we summarized overall AUC and misclassification under all settings with linear kernel and Gaussian kernel. The trend in these indices is similar to the first set of simulations. That is, for more complicated functions, KSVM noticeably improves upon SVM$_1$ and SVM$_0$ with either linear or Gaussian kernel. For simpler functions such as Gaussian, using a Gaussian kernel combining age and markers improves overall performance of SVM$_1$.

3.5 Applications to two clinical studies on Huntington's disease

HD is an autosomal dominant disease caused by an expansion of CAG trinucleotide repeats in IT15 gene on chromosome 4 (Huntington's Disease Collaborative Research Group, 1993). The disease is considered nearly fully penetrant. The inheritance of an expansion of CAG trinucleotide repeats (mutation) from a father is associated with increased penetrance to a greater extent in younger subjects than older subjects, while the effect of inheritance from a mother slightly increases over age range of their children. Majority of subjects with an expansion of CAG repeats in IT15 gene (CAG repeats ≥ 36) on one allele will develop HD if not censored by death (Kieburtz and Huntington Study Group, 1996b). It is well established that the risk of HD diagnosis increases with age and CAG repeats length (Zhang et al., 2011a). A range of cognitive and behavioral markers may have age-varying effect on the risk of HD as well. For example, the symbol digit modality test score (SDMT, a neuropsychological measure of attention, Smith, 1982) may be more sensitive than the total motor score (a measure of motor impairment in HD, Kieburtz and Huntington Study Group, 1996b) in identifying younger subjects at risk of HD while total motor score maybe more sensitive for older adults (Figure 3.6). In this section, various methods are applied to two large genetic epidemiological studies on HD to investigate these issues.

3.5.1 COHORT study results

COoperative Huntington's Observational Research Trial (COHORT, Dorsey and Huntington Study Group COHORT Investigators, 2012) is a large multi-site study that includes 42 Huntington Study Group research centers in North America and Australia. In the COHORT study, standard demographic, neurological, cognitive and behavioral instruments were administered. Individuals who met criteria for Huntington's disease (receiving a diagnostic confidence level, DCL, of 4 on the UHDRS assessment) as well as individuals at risk for HD by virtue of having a first degree relative with HD were assessed. In this example baseline data was used, and there were 338 premanifest cases and 670 controls.

Although genetic testing is available to determine whether a premanifest subject (individuals who have not been diagnosed) carries an expansion of CAG repeats, most individuals with a known family history of HD choose not to be tested since there is currently no efficacious treatment to prevent or delay onset of disease (Williams et al., 2010). Therefore, an important research goal is to develop personalized classification to distinguish pre-symptomatic subjects who will develop HD from controls who will never develop HD without taking a genetic test. In clinical practice, HD diagnosis is based on motor symptoms, and clinicians assign a diagnostic confidence level (DCL) from UHDRS motor exam. A lower DCL category indicates lower confidence of HD, and a level of "4" indicates confirmed HD and these subjects are no longer premanifests (Paulsen et al. 2008). For a neurodegenerative disease such as HD or Alzheimer's disease (Celsis, 2000), age is one of the most important variable to control for. The goal of this analysis is to develop age-sensitive prediction to determine whether a subject who has not received a diagnosis of HD (e.g., did not receive a UHDRS DCL of 4) at the baseline visit is a pre-manifest HD case (i.e., carrying an expansion of CAG repeats, gene-positive) or a control who will not develop HD (no CAG expansion, gene-negative, will not develop HD).

To this end, we first show some descriptives of the COHORT data. In Figure 3.3, we present the scatter plots of a few continuous variables reported in the literature (Langbehn et al., 2007) associated with the risk of HD such as total motor score of the UHDRS (higher is more severe) and symbol digit modality test, SDMT (higher score is better). We overlay the LOWESS smoothing of the average scores in the premanifest case group and control group on the scatter plot. It is clear that none of the markers alone can discriminate the groups based on a linear boundary. We tested for nonlinearity through a regression spline model with two knots and found a significant nonlinear effect for total motor score, SDMT, and verbal fluency test. It is desirable to combine markers and create nonlinear classification boundary.

We applied KSVM with Gaussian kernel to combine 19 markers in COHORT to capture the nonlinear age trend and develop an age-sensitive prediction rule. There were 6 continuous markers (e.g., body mass index (BMI), UHDRS total motor score, SDMT, verbal fluency test (Mitrushina et al., 2005), and stroop test (Stroop, 1935)) and 13 binary markers (e.g., history of alcohol abuse, history of drug abuse, significant history of depression, current depression, mother affected by HD, father affected by HD). To compute an honest AUC and misclassification rate, we randomly splitted samples into a training set ($n = 700$, approximately 34% of which are premanifest cases) and a testing set ($n = 308$) 100 times and reported the average performance indices when applying fitted model to the testing set. We compared the overall AUC and average misclassification rate over age for KSVM using all 19 markers with using a single marker for several selected markers. We compared with the penalized logistic regression with varying-coefficient age effect and accounting for interactions among markers (Paik and Hastie 2009). The varying-coefficient of age takes a nonparametric form fitted by a fourth order B-spline basis with 10 knots, and the tuning parameter was selected by five-fold cross validation. Lastly, we also compared KSVM with SVM$_1$ as described in section 3.4.

We summarize the overall sensitivity, specificity, AUC and misclassification rate using all 19 markers and several examples of using each individual marker alone in Table 3.3. KSVM with all 19 markers significantly improves the overall AUC (0.88) and decreases the average misclassification rate (0.19) comparing to using a single marker alone. It is clear that combining all the markers greatly improves the prediction performance distinguishing carriers of an expansion of CAG repeats from non-carriers (controls). Among the single marker models, total motor score has the highest AUC, and the other markers have similar predictive powers that are weaker than the total motor score. The average overal AUC and sensitivity are higher than penalized logistic regression with varying coefficients and SVM_1, and the misclassification rate for KSVM is lower than these two competing methods. In Figure 3.4, we present a boxplot of four performance measures obtained from 100 cross validations comparing three methods to demonstrate superior performance of KSVM. The mean AUC, sensitivity and missclassification rate of KSVM are better than the other two methods, while the specificty is similar. The variability of specificity and other measures of KSVM is smaller than the competing methods, suggesting KSVM to be more robust.

In the top panels of Figure 3.5 we show the age-specific sensitivity and specificity. We see a decreasing age trend in sensitivity which suggests it is easier to screen presymptomatic cases from the population for younger subjects than for older subjects, i.e., the predictive score is more sensitive for younger subjects. When a subject shows subtle motor signs or cognitive decline at an early age, it is an indication of increased likelihood of developing HD in the future since such signs may be rarely present in controls of similar age. When a subject shows signs of clinical symptoms at an older age, however, it is less predicative of HD disease status since controls at older age may also show similar signs.

Combining all markers significantly improves over using single marker. For example, total motor score and SDMT have sensitivities decreasing to zero for older ages (non-age-corrected raw SDMT was used). We show the specificity in the upper right panel of Figure 3.5. As expected, specificity increases with age, which suggests it is easier to screen controls from the sample for older subjects. When the clinical markers are absent by an old age, it is more likely a subject will never develop the disease, and therefore the score is more specific for older subjects. Furthermore, since a subject at-risk for HD is mostly likely to develop HD between age 30 and 50 (Foroud et al., 1999), the increasing trend in specificity is consistent with the clinical observation that an older subject who does not develop HD by a certain age is more likely to be in the control group. When compared to the penalized logistic regression, we see an improvement in sensitivity especially at the younger age.

In the bottom panels of Figure 3.5, we show trajectories of the age-specific AUC and misclassification rate. Again, we see at each age, using multiple markers has superior performance than using each single marker. The general trend shows that considering both sensitivity and specificity, it is easier to predict the risk status of HD in an older subject than a younger subject since the AUC increases with age and the misclassification status decreases with age. We can also see from the figure that the combined predictive score maybe more accurate in the older age range, for example, the AUC>0.85 for subjects with age$>$ 38. When splitting samples by the median age (47), the AUC is 0.84 for younger subjects and 0.89 for older subjects. The AUC of the KSVM is higher than the logistic regression from age 20 to 55, and similar from 55 to 70. Same trend is observed for the misclassification rate.

To further investigate the relative ranking of markers, the first two subfigures in Figure 3.6 present the age-specific predictive effect of several markers from age 20 to 70. These effects are computed as differences in the fitted discriminant functions between values 1 and 0 of a particular binary marker or as differences of 1/4 standard deviation units increase of a particular continuous marker with other markers fixed at sample means in the local age window (5-year). It shows the markers expressing different trends: some with increasing age effect (seeing a mental health professional) and decreasing effect (father's HD status). More importantly, we see that the relative magnitude of the marker effect changes across age and the ranking of the importance of markers based on the magnitude of shifts in their classification function also varies with age. For example, SDMT score is more important than the total motor score for younger subjects, while the total motor score dominates other markers for older subjects (age 45 or above).

In summary, this analysis shows that markers' sensitivity and specificity vary in predicting at risk for HD according to age. Combining informative markers significantly improves prediction accuracy. The most important marker for younger subjects is SDMT while it is total motor score for older subjects.

3.5.2 PREDICT-HD study results

We illustrate our methods through a second example, PREDICT-HD (Paulsen et al., 2008), a 32-site observational study of HD focusing on premanifest subjects followed from the prodromal phase through to post-diagnosis. To date, the main study has 1314 total participants, 1013 of whom were gene-expanded cases and 301 of whom were non-expanded controls. The individual follow up period spans 10 years with annual or biennial measurements on variables in important domains of motor, cognitive, psychiatric as well as brain imaging. The number of subjects at each visit ranges from 43 to 380. One of the major goals of PREDICT-HD is to discover markers for predicting onset of HD diagnosis based on motor symptoms in a short study period in premanifests subjects. Such information is valuable for planning recruit of a future clinical trial on HD. Thus, here our outcome of interest is the risk of a pre-symptomatic subject at baseline receiving HD diagnosis during the study period. That is, to predict risk of conversion: risk of a subject with DCL<4 (no diagnosis) at the baseline converting to DCL=4 (receive a confirmed clinical diagnosis) in the study period. This outcome of interest in this section is conversion status distinguishes PREDICT analysis from COHORT analysis in the previous section (outcome mutation carrier status).

Our analysis included a subsample of 671 gene-expanded cases from PREDICT-HD study who were not diagnosed with HD at the baseline. There were 107 converters who received a disease diagnosis during the study period. Five markers (gender, CAG repeats, total motor score, TFC and stroop color score) were used to predict the age-specific conversion status in the age range from 25 to 65. We applied both KSVM (with Gaussian kernel) and penalized logistic regression (Paik and Hastie 2009) with nonparametric varying coefficient (B-spline basis expansion with 10 knots) to the data for comparison similar to the COHORT study. The tuning parameter was selected by five-fold cross-validation.

We show some descriptives of the markers included in the analyses in the top panels of Figure 3.7. We present the scatter plots of baseline total motor score and stroop color score with overlaid LOWESS plots as examples. Although the figure hints the mean total motor score to be different in converters and non-converters, a linear separation boundary does not appear to be adequate. Similar pattern can be seen for the stroop score. We therefore combine all five markers to perform classification with a nonlinear boundary. The bottom panels of Figure 3.7 show the results. From bottom left subfigure, we see that the age-specific sensitivity of KSVM is much higher compared to penalized logistic regression in the younger age range (before 43 years old). The specificity of the two methods is similar (results not shown). For the older age range, their performance is similar. The right panel shows the standardized effects of four continues markers (measured in 1/4 standard deviation unit of each marker). Baseline total motor score has the largest effect across all age range, suggesting the importance of this marker in tracking disease progression. Among the other markers, total functional capacity has larger effect for younger subjects (less than age 37), while these markers have similar magnitude of effect for older age range.

In summary, this analysis shows that KSVM creates much more sensitive predictive score especially for younger subjects. In predicting conversion status during a fixed time period, baseline total motor score has dominating effect over other markers.

3.6 Discussion

We have proposed a local smoothing classification method to predict disease risk accounting for its age-dependent effect. Age has clear clinical interpretation and represents a constellation of underlying unobserved biological and physiological factors. Constructing age-specific prediction rules facilitates studying the timing of intervention and discovering markers useful to guide personalized treatments. The fitted coefficients $\boldsymbol{\beta}(w)$ depict age-sensitive profiles of the markers on disease risk. Furthermore, the obtained age-dependent predictive scores can be used to allocate patients into risk groups. Therefore the developed methods can be used to recruit high-risk patients for clinical trials based on a subject's age and marker values to improve efficiency of the trial. In the application example, we classified HD premanifest case/control status for presymptomatic individuals where all subjects with CAG$\geq$ 36 belong to the case group (they will develop HD a future time point). It would be interesting to use the actual CAG repeat length in a future work and to classify more refined groups of cases (e.g., close or far to disease onset). It may also be desirable to examine predictive powers of other markers such as brain imaging measures in a future analysis.

Here we considered markers with age-dependent effects, but it is easy to incorporate markers with constant effects. For example, an iterative backfitting procedure can be used to include markers $\boldsymbol{Z}$ with age-invariant effects and fit decision boundaries such as

$$\alpha(w) + \boldsymbol{X}^T\boldsymbol{\beta}(w) + \boldsymbol{Z}^T\boldsymbol{\gamma}.$$

Specifically, at a given $\boldsymbol{\gamma}$, $\alpha(w)$ and $\boldsymbol{\beta}(w)$ will be fitted through the developed approaches. Then fixing these functions at their fitted values, an update of $\boldsymbol{\gamma}$ is obtained through a regular SVM procedure without smoothing. These two steps will be iterated until convergence. We can extend the current approach when there is an additional marker that needs special attention (e.g., BMI or CAG repeats length). We can then extend our method to incorporate a two-dimensional coefficient function, i.e. $\boldsymbol{\beta}(w, u)$, and apply two-dimensional local kernel smoothing. It is also easy to extend the current methods to multi-category outcomes and to continuous outcomes.

Large margin classification with other penalty functions are discussed in Zhu et al. (2003) (i.e., 1-norm SVM) and Zou and Yuan (2008) (i.e., F_∞-norm SVM). We have not considered marker selection in the current local smoothing setting. It may be possible to use some of the other penalty functions to perform marker selection so that the marker without any effect at the entire range of age will be automatically excluded. We do not discuss effective handling of correlated markers here. Lastly, our simulation results show that different choices of Mercer kernel may lead to slight difference in prediction accuracy. A procedure that maximizes performance over a class of Mercer kernels is conceivable. These topics worth some future research.

Table 3.1: Summary of simulation results from Simulation 1 (retrospective data generation)

Marker used	Mercer kernel	Index	SVM$_0^\dagger$	SVM$_1^\ddagger$	KSVM§	SVM$_0^\dagger$	SVM$_1^\ddagger$	KSVM§
				n=500			n=1000	
X_{i1}	Linear	Miss*	0.439	0.334	0.240	0.440	0.334	0.228
		AUC**	0.500	0.738	0.833	0.498	0.743	0.848
	Gaussian	Miss	0.446	0.277	0.233	0.442	0.260	0.223
		AUC	0.500	0.760	0.825	0.501	0.783	0.838
X_{i2}	Linear	Miss	0.262	0.174	0.164	0.264	0.170	0.161
		AUC	0.819	0.884	0.899	0.818	0.885	0.903
	Gaussian	Miss	0.263	0.165	0.166	0.265	0.161	0.161
		AUC	0.782	0.899	0.890	0.780	0.902	0.897
Multiple	Linear	Miss	0.267	0.174	0.143	0.265	0.168	0.133
		AUC	0.813	0.897	0.932	0.817	0.902	0.941
	Gaussian	Miss	0.270	0.172	0.151	0.265	0.158	0.140
		AUC	0.792	0.896	0.915	0.799	0.909	0.925

*: Overall misclassification rate averaged over age; **: Overall AUC averaged over age;

† : Ignoring age effect; ‡ : A parametric linear age effect; §: Local smoothing of age effect.

Table 3.2: Summary of simulation results from Simulation 2 (prospective data generation)

Marker used	Mercer kernel	Index	$\text{SVM}_0^\dagger$	$\text{SVM}_1^\ddagger$	$\text{KSVM}^\S$	$\text{SVM}_0^\dagger$	$\text{SVM}_1^\ddagger$	$\text{KSVM}^\S$
				n=500			n=1000	
X_{i1}	Linear	Miss†	0.168	0.153	0.084	0.169	0.153	0.081
		AUC‡	0.930	0.937	0.979	0.930	0.937	0.981
	Gaussian	Miss	0.170	0.153	0.087	0.170	0.152	0.082
		AUC	0.903	0.922	0.972	0.902	0.928	0.978
X_{i2}	Linear	Miss	0.164	0.164	0.079	0.166	0.166	0.080
		AUC	0.931	0.930	0.984	0.930	0.929	0.985
	Gaussian	Miss	0.164	0.085	0.083	0.163	0.081	0.080
		AUC	0.897	0.980	0.978	0.900	0.983	0.982
Multiple	Linear	Miss	0.150	0.146	0.084	0.150	0.143	0.081
		AUC	0.935	0.940	0.977	0.936	0.944	0.978
	Gaussian	Miss	0.153	0.138	0.095	0.152	0.120	0.081
		AUC	0.920	0.944	0.972	0.921	0.957	0.979

Legends see Table 1.

Table 3.3: Overall performance over age for multiple markers models compared with various single marker models.

	Misclassification	AUC	Sensitivity	Specificity
All markers (KSVM)	0.190 (0.020)‡	0.878 (0.020)	0.700 (0.046)	0.864 (0.024)
All markers (SVM_1)	0.223 (0.023)	0.829 (0.024)	0.604 (0.062)	0.864 (0.032)
All markers (Penalized logistic regression†)	0.218 (0.029)	0.844 (0.038)	0.655 (0.048)	0.846 (0.032)
Total Motor Score	0.276 (0.021)	0.731 (0.034)	0.403 (0.082)	0.885 (0.038)
SDMT	0.307 (0.024)	0.668 (0.037)	0.258 (0.068)	0.910 (0.038)
BMI	0.322 (0.023)	0.657 (0.028)	0.296 (0.125)	0.870 (0.058)
Mini-Mental Exam	0.306 (0.022)	0.647 (0.033)	0.272 (0.061)	0.904 (0.033)
Verbal Fluency Test	0.314 (0.021)	0.651 (0.031)	0.242 (0.069)	0.907 (0.039)
Stroop score	0.326 (0.020)	0.639 (0.034)	0.226 (0.096)	0.898 (0.049)
Father affected by HD	0.293 (0.022)	–	0.366 (0.070)	0.880 (0.043)
Mother affected by HD	0.321 (0.023)	–	0.326 (0.121)	0.859 (0.070)
Currently see a Mental Health Professional	0.319 (0.024)	–	0.277 (0.104)	0.886 (0.052)
Significant history of depression	0.322 (0.023)	–	0.256 (0.099)	0.893 (0.053)
History of alcohol abuse	0.328 (0.025)	–	0.236 (0.112)	0.894 (0.061)
Significant history of suicidal ideation	0.332 (0.023)	–	0.239 (0.124)	0.887 (0.069)
History of tobacco abuse	0.330 (0.022)	–	0.240 (0.120)	0.890 (0.062)
Significant history of OCD	0.322 (0.024)	–	0.231 (0.097)	0.906 (0.050)
History of drug abuse	0.328 (0.022)	–	0.223 (0.103)	0.901 (0.058)
Current depression	0.326 (0.022)	–	0.180 (0.103)	0.925 (0.059)

†: Proposed in Paik and Hastie (2009).

‡: Mean and empirical standard deviation for 100 cross validations.

Figure 3.1: (Simulation 1) Age-specific misclassification rate (left) and AUC (right) for SVM_0, SVM_1 and KSVM. The corresponding analysis from the top to the bottom are: using X_{i1}, using X_{i2} and using multiple markers.

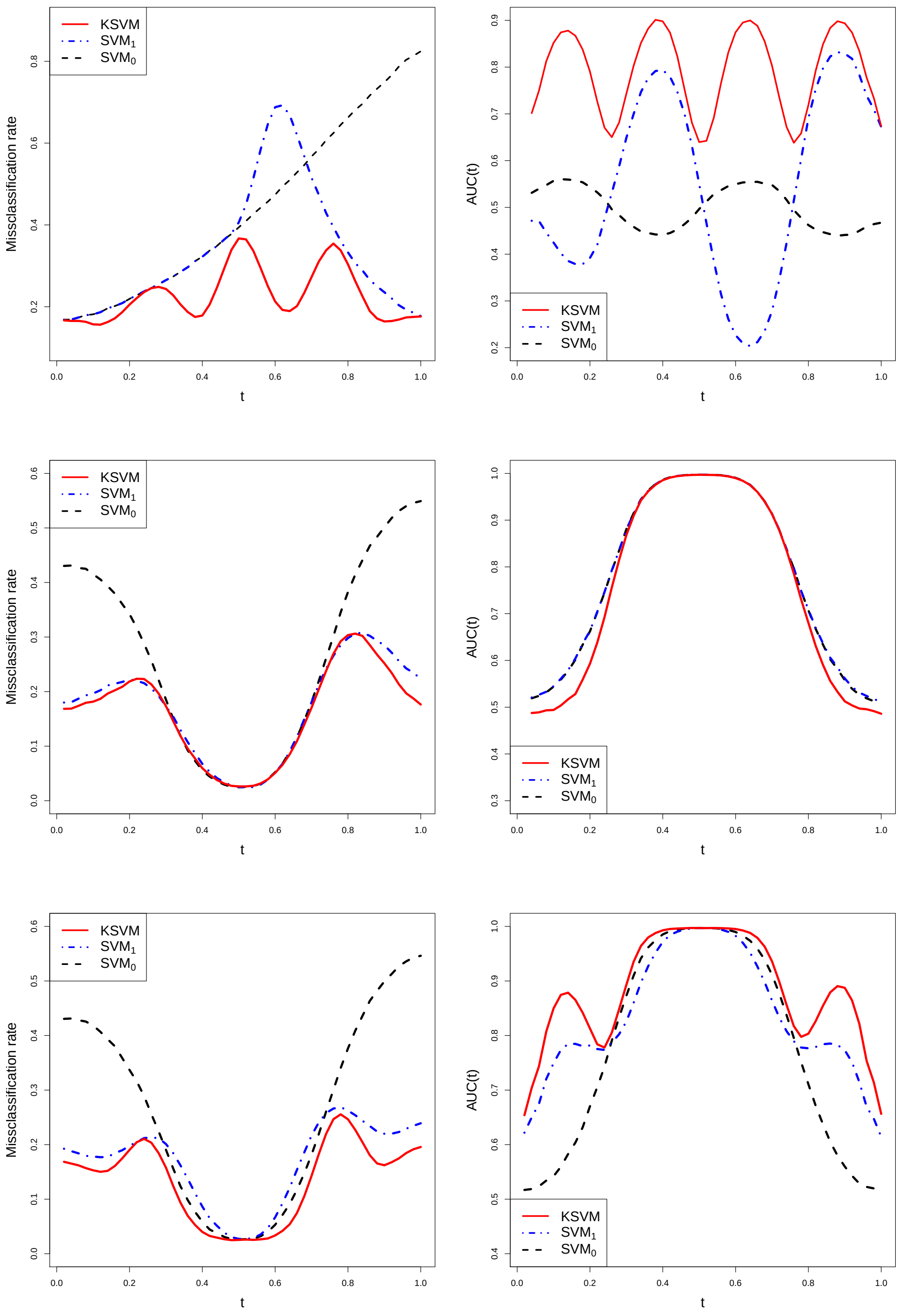


Figure 3.2: (Simulation 2) True classification boundary and a typical set of simulated data.

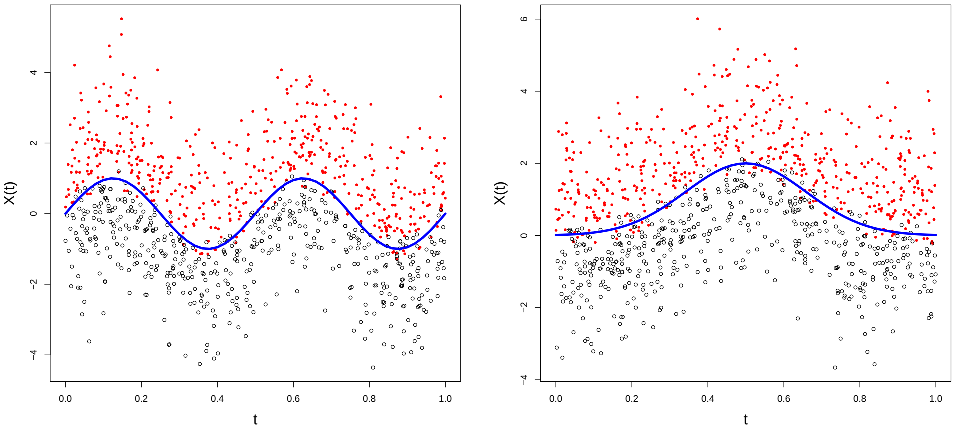


Figure 3.3: Descriptive scatter plots of several continuous markers and lowess smoothed mean curves in COHORT

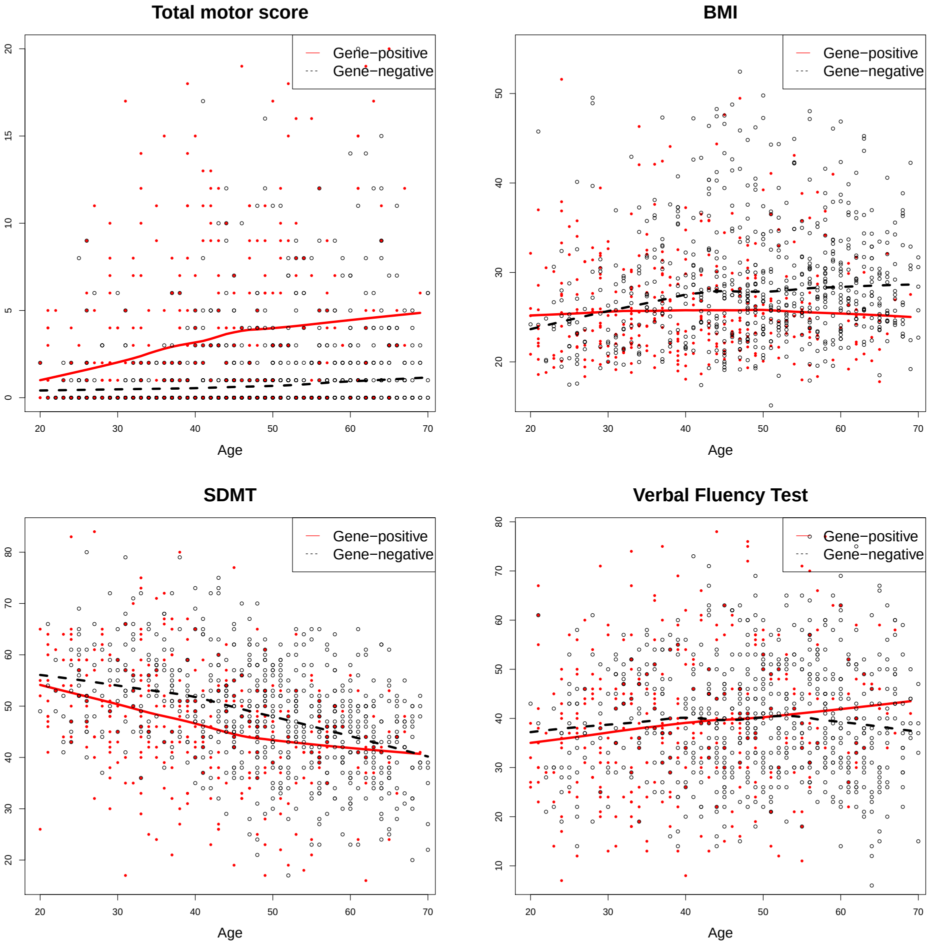


Figure 3.4: Comparison of KSVM, SVM$_1$ and penalized logistic using 19 markers in predicting at-risk status of Huntington's disease with COHORT premanifest subjects (overall 1-Misclassification Rate, AUC, Sensitivity, and Specificity).

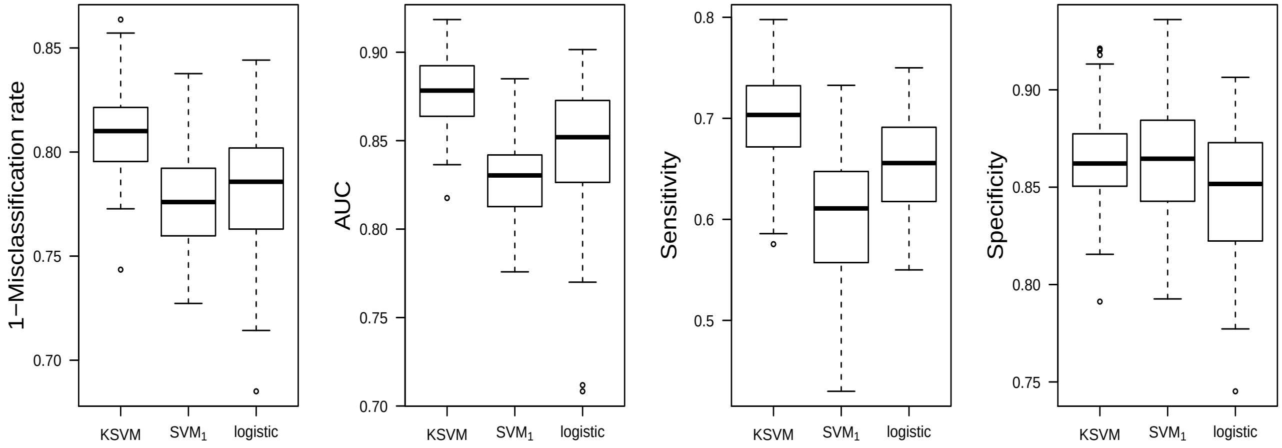


Figure 3.5: Comparison of 19-marker penalized logistic, 19-marker KSVM and single-marker KSVM in predicting at-risk status of Huntington's disease with COHORT premanifest subjects (age-specific sensitivity, specificity, AUC and misclassification rate).

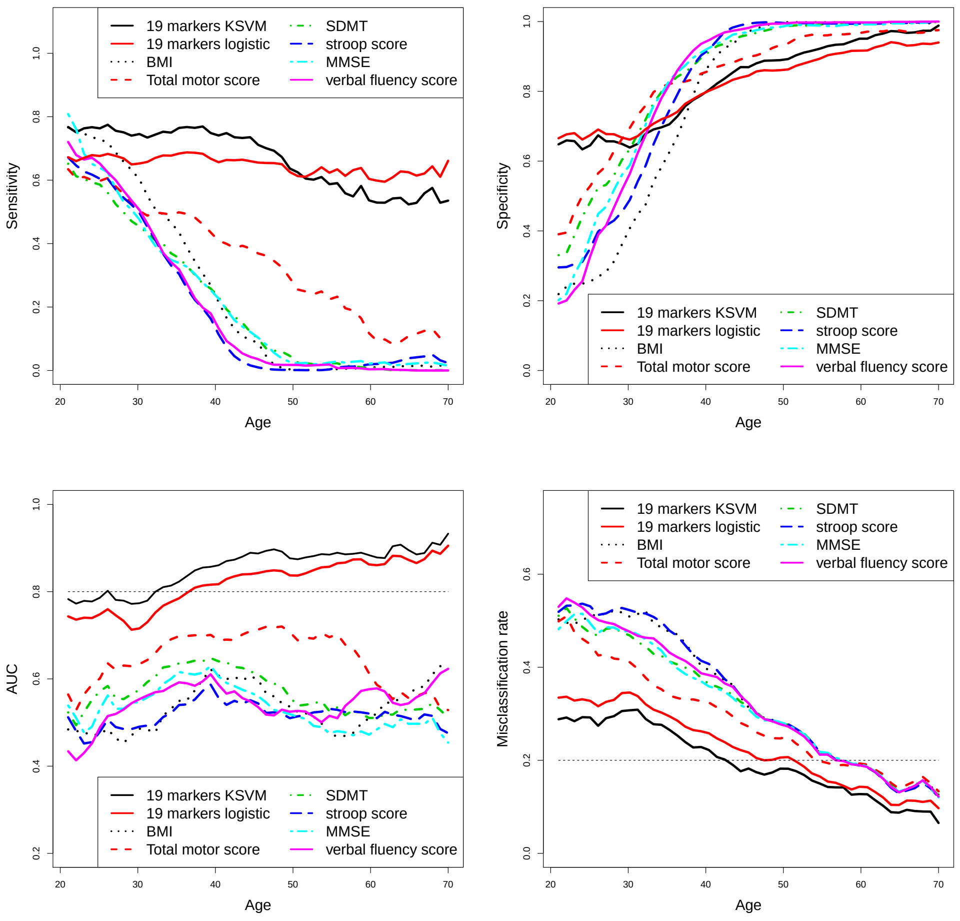


Figure 3.6: Standardized effect for key markers in COHORT study fitted by KSVM.

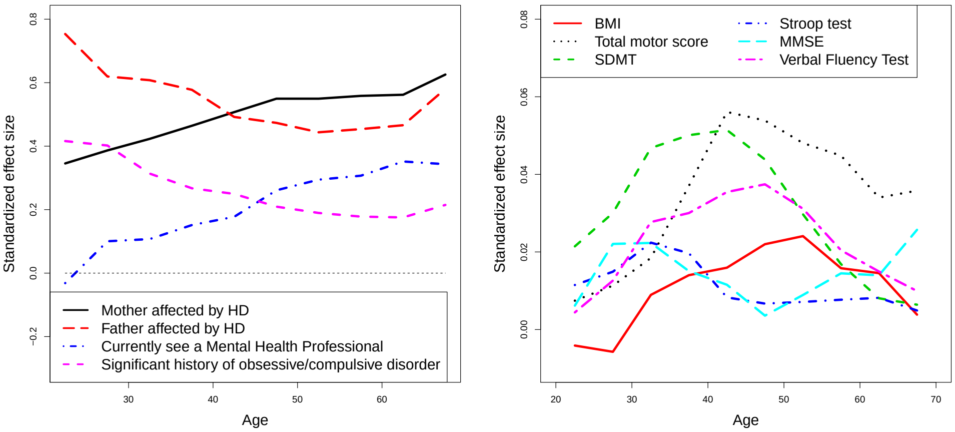
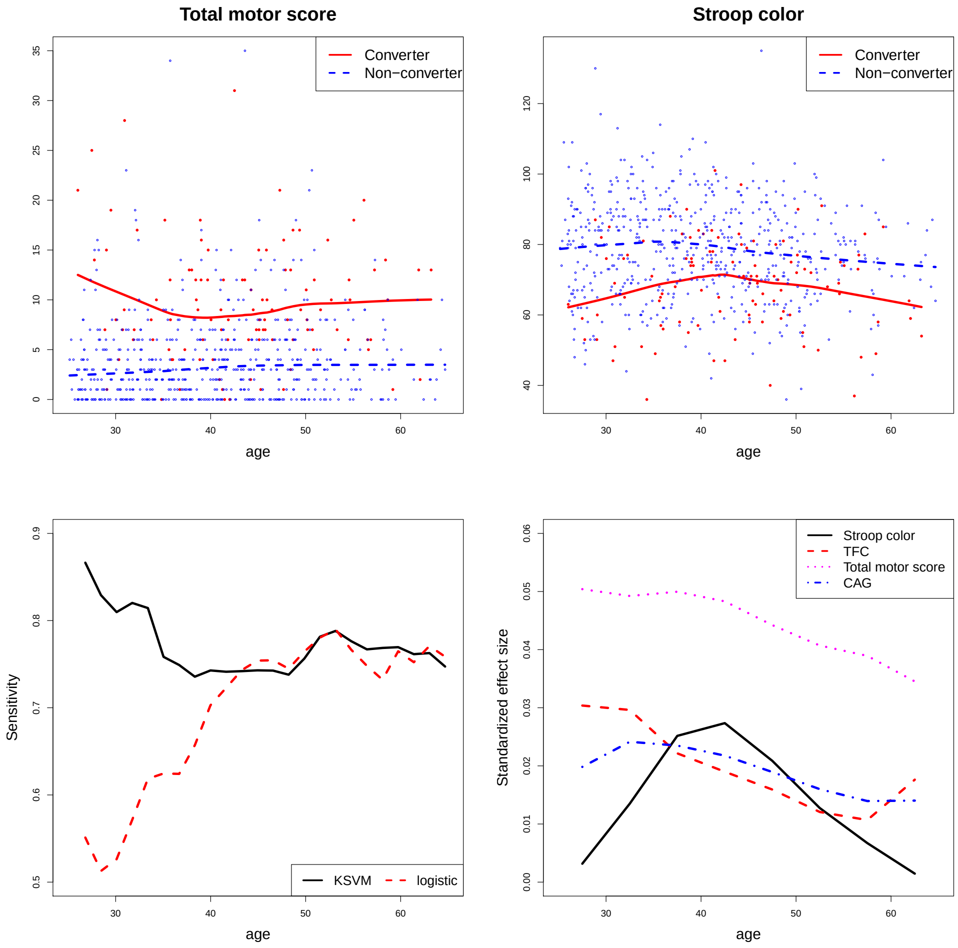


Figure 3.7: Age-specific descriptives, sensitivity, and standardized effect for predicting HD conversion status in PREDICT-HD premanifest subjects.



Chapter 4

Multiple kernel learning with latent effects for predicting longitudinal outcomes and data integration

4.1 Overview

In this chapter, we propose statistical learning methods for disease prediction for longitudinal binary data with heterogeneous data sources. In the first section, we propose a multiple kernel support vector machine with random effects for predicting longitudinal outcomes and data integration. In the second section, we conduct two simulation studies to investigate performance of the proposed methods. In the third section, we apply the method to two longitudinal study data sets, the PREDICT-HD data and the ADNI data. Finally, we summarize our findings and discuss possible extensions in the fourth section.

4.2 Multiple Kernel Fusion Learning for Longitudinal Data

We start by briefly introducing standard statistical learning through support vector machine with a single kernel, followed by incorporating longitudinal component to the learning through fusing two kernels, and lastly we discuss integration of multiple data sources through fusing multiple heterogeneous kernels.

4.2.1 Review of support vector machine

Let $\mathcal{X}$ denote a complete separable space for feature variables. The random feature variables $\mathbf{X}$ take values in $\mathcal{X}$, and the binary disease outcomes Y take values in $\mathbb{R}$. The goal of statistical learning is to train an optimal prediction function $f : \mathcal{X} \to \mathbb{R}$ to predict Y given $\mathbf{X}$ for any future subject, where the performance of prediction is quantified by the prediction error defined as $E[I(Yf(X) < 0]$. Due to the non-smoothness of $I(Yf(X) < 0)$, the optimal prediction function is usually obtained by minimizing the empirical version of some surrogate loss function. One such loss function most commonly used is the hinge loss, or the so called support vector machine (SVM, Vapnik, 1995), and it has been proven to be successful in a wide range of applications (Orru et al., 2012).

Assume that we have n independent observations $(\mathbf{x}_i, y_i), i = 1, ..., n$. With a linear prediction function $f(\mathbf{x}_i) = \langle \mathbf{x}_i, \mathbf{w} \rangle + b$, where the inner product here is defined as $\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a}^T \mathbf{b}$, the primal optimization problem of the SVM has the form (e.g., Hastie et al., 2009)

$$\min_{\mathbf{w} \in \mathcal{X}, b \in \mathbb{R}} \left\{ \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^{n} \xi_i \right\} \tag{4.1}$$

subject to the constraints with slack variables ξ_i

$$y_i(\langle \mathbf{x_i}, \mathbf{w} \rangle + b) \geq 1 - \xi_i \text{ and } \xi_i \geq 0, \text{ for all } i = 1, ..., n.$$

The corresponding dual form is

$$\begin{aligned} &\max_{\boldsymbol{\alpha}\in\mathbb{R}^n} \left\{ \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i\alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \right\}, \qquad (4.2) \\ \text{subject to} \quad & 0 \le \alpha_i \le C, \;\; i = 1, ...n, \\ \text{and} \quad & \sum_{i=1}^{n} \alpha_i y_i = 0. \end{aligned}$$

To accommodate nonlinear separating boundary, one defines a Mercer kernel $k(\cdot, \cdot)$ such that

$$k(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle,$$

where $\Phi(\cdot)$ is the mapping from the input space to a higher dimensional feature space, and $\langle \cdot, \cdot \rangle$ is the inner product defined in the reproducing kernel Hilbert space (RKHS, Wahba 1990). The corresponding dual form becomes

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^n} \left\{ \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i\alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \right\},$$

leading to the decision functions of the form

$$d(\mathbf{x}) = \text{sign } (\sum_{i=1}^{n} \alpha_i y_i k(\mathbf{x}, \mathbf{x}_i) + b).$$

Note that one advantage of solving the optimization from the dual form is that the explicit form of $\Phi(\cdot)$ does not need to be known as long as the kernel function $k(\cdot, \cdot)$ is well defined (Kimeldorf and Wahba, 1970).

4.2.2 Proposed multiple kernel learning for longitudinal data

The above formulation is designed for independent outcomes. For longitudinal biomedical data, outcome measures on the same subjects are correlated after accounting for the observed fixed effects feature variables. Taking advantage of such correlation is expected to lead to improved prediction. Classical longitudinal analysis divides into two camps: estimating the marginal population-average effect, and estimating the subject-specific effect given the random effects. For the former view, correlation among repeated measures is treated as nuisance parameter, while for the latter it is modeled through subject-specific random effects. In our setting, subject-specific classifications are of interest instead of population average effects, therefore we introduce random effects to the SVM framework to improve prediction in our proposed approach.

Assume that we have n independent subjects and the ith subject has n_i visits. Let y_{ij} denote the disease outcome for the ith subject at the jth visit coded as "1" for diseased subjects and "-1" for non-diseased subjects. Let $\mathbf{x}_{ij}$ denote a vector of feature variables collected at the same visit. We introduce two latent random effects for subject i, a time-invariant effect a_{ij}, which aims to capture the long-term latent effect across all the visits from the same subject, and a time-varying effect b_{ij}, which attempts to account for short-term latent effect or local influence from recent history that depends on the time interval between visits. Therefore, for a subject with feature variables $\mathbf{x}_{ij}$ at time t_{ij}, a prediction rule with subject-specific random effects can be expressed as

$$\text{sign}\{f(\mathbf{x}_{ij}, a_{ij}, b_{ij})\},$$

where the prediction function has the form

$$f(\mathbf{x}_{ij}, a_{ij}, b_{ij}) = \langle \Phi_x(\mathbf{x}_{ij}), \mathbf{w} \rangle + w_a \Phi_a(a_{ij}) + w_b \Phi_b(b_{ij}). \tag{4.3}$$

Here, $\Phi_x(\mathbf{x})$ consists of some mapping from the input space $\mathcal{X}$ to a higher-order feature space (for example, the basis function associated with some reproducing kernel Hilbert space) and both $\Phi_a(a)$ and $\Phi_b(b)$ are nonlinear transformation of the latent effects which will be induced by some kernel functions defined for a_{ij} and b_{ij}, respectively in Section 2.3. For identifiability, we also assume that a_{ij} and b_{ij} are standardized random variables with mean zero and variance one. Clearly, since $\mathbf{a}$ and $\mathbf{b}$ are unobserved random variables, conventional SVM techniques cannot be directly applied.

When including the random effects into the model, the single kernel SVM becomes a multi-kernel SVM with one kernel for fixed effects and two kernels for random effects. Following the multiple kernel learning framework, a weight parameter θ is then assigned to each kernel and a fused kernel is formed as a linear combination of kernels under an L_2-norm regularization constraint on the weight parameters. The weights are chosen in a data-driven way to minimize the loss function under the fused kernels. Thus, the primal form in the feature space becomes

$$\begin{aligned} &\min_{\mathbf{w}_x \in \mathcal{X}, b \in \mathbb{R}} \; \frac{1}{2}\left(\frac{\mathbf{w}_x^T \mathbf{w}_x}{\theta_x} + \frac{w_a^2}{\theta_a} + \frac{w_b^2}{\theta_b}\right) + C \sum_{i,j}^{n,n_i} \xi_{ij} \qquad (4.4) \\ \text{subject to} \quad & y_{ij}\left(\sqrt{\theta_x}\langle \Phi_x(\boldsymbol{x}_{ij}), \mathbf{w}_x\rangle + \sqrt{\theta_a} w_a \Phi_a(a_{ij}) + \sqrt{\theta_b} w_b \Phi_b(b_{ij})\right) \geq 1 - \xi_{ij} \\ & \xi_{ij} \geq 0, \; i = 1, ..., n, \text{ and } j = 1, ..., n_i, \\ & \theta_x^2 + \theta_a^2 + \theta_b^2 = 1, \quad \theta_x, \theta_a, \theta_b \geq 0. \end{aligned}$$

As a remark, comparing the optimization problem for longitudinal data (4.4) with the original standard SVM primal form (4.1), we observe that the objective function for the former is a conic combination of the separate objective functions for the latter with a quadratic constraint. Furthermore, the resemblance with multiple kernel learning allows easy generalization to accommodate data from heterogeneous sources by using separate kernels for observed feature variables from each source. Such method incorporates prior knowledge on each source while performing integration. Contrary to concatenating all variables in a single kernel, using separate ones reflects prior knowledge that the feature variables from the same source have stronger correlations than with variables from difference sources. For example, assuming there are P data sources for fixed effects $\mathbf{x}_{ij} = (\mathbf{x}_{ij1}, ..., \mathbf{x}_{ijP})$ and with two kernels for two types of random effects, the corresponding primal form is

$$\begin{aligned}
&\min_{\mathbf{w}\in\mathcal{X},b,\theta_p,\theta_a,\theta_b\in\mathbb{R}} \frac{1}{2}\left(\sum_{p=1}^{P}\frac{\mathbf{w}_p^T\mathbf{w}_p}{\theta_p}+\frac{w_a^2}{\theta_a}+\frac{w_b^2}{\theta_b}\right)+C\sum_{i,j}^{n,n_i}\xi_{ij}\\
\text{subject to}\quad & y_{ij}\left(\sum_{p=1}^{P}\sqrt{\theta_p}\langle\Phi_p(\boldsymbol{x}_{ijp}),\mathbf{w}_p\rangle+\sqrt{\theta_a}w_a\Phi_a(a_{ij})+\sqrt{\theta_b}w_b\Phi_b(b_{ij})\right)\geq 1-\xi_{ij}\\
&\xi_{ij}\geq 0,\ i=1,...,n,\ \text{and } j=1,...,n_i,\\
&\sum_{p=1}^{P}\theta_p^2+\theta_a^2+\theta_b^2=1,\quad \theta_p,\theta_a,\theta_b\geq 0,\quad p=1,\cdots,P.
\end{aligned}$$

The computation of the multiple kernel learning is essentially a quadratically-constrained quadratic programming (QCQP) problem (Lanckriet et al., 2004). Specifically, the dual form is

$$\begin{aligned}
\max_{\boldsymbol{\alpha}}\min_{\boldsymbol{\theta}}\quad & \sum_{ij}^{n,n_i}\alpha_{ij}-\frac{1}{2}\sum_{i,k}^{n}\sum_{j,l}^{n_i}\alpha_{ij}\alpha_{kl}y_{ij}y_{kl}\{\sum_{p=1}^{P}\theta_p k_p(\mathbf{x}_{ijp},\mathbf{x}_{klp})+\\
& \theta_a k_a(a_{ij},a_{kl})+\theta_b k_b(b_{ij},b_{kl})\}\\
\text{subject to}\quad & 0\leq\alpha_{ij}\leq C,\ i,k=1,...,n,\ j,l=1,...,n_i,\ \sum_{i,j}^{n,n_i}\alpha_{ij}y_{ij}=0,\\
& \sum_{p}^{P}\theta_p^2+\theta_a^2+\theta_b^2=1,\ \theta_p,\theta_a,\theta_b\geq 0,\ p=1,\cdots,P.
\end{aligned}$$

where $k_p(\mathbf{x}_{ijp},\mathbf{x}_{klp}) = \langle\Phi_p(\mathbf{x}_{ijp}),\Phi_p(\mathbf{x}_{klp})\rangle$ is the kernel for the reproducing kernel Hilbert space for $\mathbf{x}_{ijp}$, and $k_a(a_{ij},a_{kl})=\langle\Phi_a(a_{ij}),\Phi_a(a_{kl})\rangle$, $k_b(b_{ij},b_{kl})=\langle\Phi_b(b_{ij}),\Phi_b(b_{kl})\rangle$ are kernel functions for some inner products defined for latent effects we discuss next.

4.2.3 Choice of kernel functions for latent effects

Here we introduce kernels to model the two random effects a_{ij} and b_{ij}, respectively. Recall kernel matrix measures similarity between two observations, a natural choice of kernel function is the covariance structure of the random effects which can also be considered as the inner product with respect to its distribution function. Thus, we assume that the similarity between the latent effects from independent subjects is zero, the similarity between the long term random effects on the same subjects is a constant ρ, and the similarity between local short term random effects depends on the time interval between the two measurements.

Specifically, to account for the long-term latent effects, we can consider a_{ij} to represent the common random effect shared across visits plus an independent random error component, and therefore the commonly shared random effect will contribute to prediction at each visit. Equivalently, construct elements in a kernel matrix as $k_a(a_{ij}, a_{kl}) = 1$ if $i = k, j = l$; $k_a(a_{ij}, a_{kl}) = \rho$ if $i = k, j \neq l$; and $k_a(a_{ij}, a_{kl}) = 0$ if $i \neq k$. That is, the kernel function for n_i long-term random effects $\mathbf{a_i} = (a_{i1}, \cdots, a_{in_i})^T$ is

$$\mathbf{K}_{\mathbf{a}_i} = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \cdots & \cdots & \cdots & \cdots \\ \rho & \cdots & \rho & 1 \end{pmatrix}_{n_i \times n_i},$$

and the kernel matrix for all $N = (n_1 + n_2 + ...)$ observations from all the subjects is

$$\mathbf{K}_a = \begin{pmatrix} \mathbf{K}_{a_1} & 0 & \cdots & 0 \\ 0 & \mathbf{K}_{a_2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \mathbf{K}_{a_n} \end{pmatrix}_{N \times N}.$$

Next, in order to account for short term latent random effects, we assume an exponential covariance structure for $\mathbf{b}_i$. Thus, $k_b(b_{ij}, b_{kl}) = \exp\{-\alpha|t_{ij} - t_{il}|\}$ if $i = k$; and $k_b(b_{ij}, b_{kl}) = 0$ if $i \neq k$. The kernel matrix for the short term random effects $\mathbf{b}_i = (b_{i1}, \cdots, b_{in_i})^T$ with measurement time points $(t_{i1}, \cdots, t_{in_i})^T$ is defined as

$$\mathbf{K}_{\mathbf{b}_i} = \begin{pmatrix} 1 & e^{-\alpha|t_{i1}-t_{i2}|} & e^{-\alpha|t_{i1}-t_{i3}|} & \ldots & e^{-\alpha|t_{i1}-t_{in_i}|} \\ e^{-\alpha|t_{i1}-t_{i2}|)} & 1 & e^{-\alpha|t_{i2}-t_{i3}|} & \ldots & e^{-\alpha|t_{i2}-t_{in_i}|} \\ \vdots & \vdots & \vdots & \ddots & \vdots \end{pmatrix}_{n_i \times n_i},$$

where α is a pre-specified scale parameter, and the kernel matrix for all time-varying short-term random effects on all subjects is

$$\mathbf{K_b} = \begin{pmatrix} \mathbf{K_{b_1}} & 0 & \cdots & 0 \\ 0 & \mathbf{K_{b_2}} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \mathbf{K_{b_n}} \end{pmatrix}_{N\times N}.$$

Under the above choice of kernels, we can optimize the dual form (4.5) using the quadratic programming. Earlier work suggests exhaustive search at given values of $\boldsymbol{\theta}$ and treating the fused kernels as a new kernel in a standard SVM optimization problem. However, the computational burden is high. A computationally efficient algorithm for solving the optimization problem (4.5) was proposed in Yu et al. (2010) to solve for weights $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ simultaneously. Specifically, the dual form (4.5) is solved under the Cauchy-Schwarz inequality as

$$\begin{aligned} \min_{t,\boldsymbol{\alpha}} \quad & \frac{1}{2}t - \sum_{i,j}^{n,n_i} \alpha_{ij} \\ \text{subject to} \quad & \sum_{i,j}^{n,n_i} \alpha_{ij} y_{ij} = 0,\ 0 \leq \alpha_{ij} \leq C, \\ & t \geq \|\boldsymbol{\gamma}\|_2, \end{aligned}$$

where $\boldsymbol{\gamma} = \left\{\boldsymbol{\alpha}^T \boldsymbol{Y}\boldsymbol{K}_1\boldsymbol{Y}\boldsymbol{\alpha}, ..., \boldsymbol{\alpha}^T\boldsymbol{Y}\boldsymbol{K}_P\boldsymbol{Y}\boldsymbol{\alpha}, \boldsymbol{\alpha}^T\boldsymbol{Y}\boldsymbol{K}_a\boldsymbol{Y}\boldsymbol{\alpha}, \boldsymbol{\alpha}^T\boldsymbol{Y}\boldsymbol{K}_b\boldsymbol{Y}\boldsymbol{\alpha}\right\}^T$, and the optimal weight parameters for the pth kernel is $\theta_p^* = \boldsymbol{\alpha}^T\boldsymbol{Y}\boldsymbol{K}_p\boldsymbol{Y}\boldsymbol{\alpha}/\|\boldsymbol{\gamma}\|_2$.

4.2.4 Prediction of future observations

For a longitudinal study, we distinguish two types of prediction of interest. We refer type A prediction as predicting outcome for a new subject with the observed feature variables $\mathbf{x}$ only and no prior history information, for example, prediction for a new subject at the baseline visit. We refer type B prediction as predicting outcomes at future follow-up time points for an existing subject with observed prior visit outcomes and feature variables $\mathbf{x}$. One of the main components of our proposed learning is to extract information from exisiting correlated outcomes to improve future prediction. For each type of the prediction, we discuss a different strategy in predicting the outcomes.

For type A prediction on a new subject with feature variables $\mathbf{x}_i$, directly using designed kernel functions and the fitted prediction function (4.3) is equivalent to using fixed effects only to predict the outcome and set the random effects at their mean level, zero. This is because the designed kernel functions $\mathbf{k}_a$ and $\mathbf{k}_b$ for random effects have non-zero values only between two visits on the same subject. In type A problem, the existing subjects and the new subject are independent, and therefore the fitted score from solving the dual form (4.5) do not involve random effects, which corresponds to using the population mean value for all subjects with fixed effects $\mathbf{x}_i$ to perform prediction.

We suggest an alternative to use random effects for type A prediction. We repeatedly draw independent random effects a_i and b_i from a working Gaussian distribution. For each random draw, we computed the predictive function as in (4.3) and classify the outcome using the sign of $f(\mathbf{x}_i, a_i, b_i)$. The final predicted outcome is based on a majority vote: if more than 50% of random draws lead to positive predicted outcomes, the final predicted outcome would be positive, and otherwise negative.

For type B prediction, we use an existing subject's predictors and outcomes at prior visits to predict their future follow up outcomes. We can then directly compute the random effects for the same subject at a future time t^* using the designed kernel matrices $\mathbf{K}_a$ and $\mathbf{K}_b$, and the fitted predictive function is obtained from the solutions to (4.5).

4.3 Simulation Studies

In this section, we conducted simulation studies to compare the empirical performance of multi-kernel SVM with several standard alternatives for analyzing longitudinal data.

4.3.1 Setting 1: single data source

In the first simulation setting, we generated the dichotomous outcomes from the following model:

$$Y_{ij} = \text{sign}\{\beta W_{ij}^{*} + a_{ij} + b_{ij} + \epsilon_{ij}\},$$

where W^* is the radius of the two spheres in Figure 4.1. First we generated $\boldsymbol{W}_{ij}$, a 3-dimensional vector randomly located either on the outer sphere with a radius equal to 2 (with a small random error) or on the inner sphere with a radius equal to 1 (with a small random error) at each visit for a subject. We used the radius W^* in the score function for generating the binary outcome. The radius changes at each visit (with equal probability to be 1 or 2). A single radial kernel SVM can generate a sphere-shaped boundary and perfectly separate the two groups of $\boldsymbol{W}$'s. $\boldsymbol{a_i}$ and $\boldsymbol{b_i}$ are subject-specific random effects. Specifically, $\boldsymbol{a_i}$ is generated from $MVN(\mathbf{0}, \Sigma_a)$, where Σ_a is a correlation matrix with compound-symmetric structure ($\rho = 0.5$), and $\boldsymbol{b_i}$ is generated from $MVN(\mathbf{0}, \Sigma_b)$, where Σ_b is a correlation matrix with exponential correlation structure, e.g., $\rho_{j,k} = \exp(-\alpha|t_j - t_k|)$ with $\alpha = 1$. Here ϵ_{ij} are normally distributed random errors of the i^{th} subject at the j^{th} visit. We performed 100 simulation runs and compared various performance indices of the proposed method under a single linear or radial kernel with and without random effects to logistic regression ignoring correlation and generalized mixed effects regression with subject-specific random intercepts. For logistic regression and generalized mixed effects regression, we included all the feature variables, their squared terms and pairwise interactions.

In type A prediction, we predicted outcomes for a new subject based on his/her observed feature variables alone and the trained model. We generated longitudinal data from the single source W plus latent random effects with a sample size of $n = 250$ subjects, each having 4 visits. Two-thirds of the subjects are included in the training set and the rest one-third as the testing set. The results are summarized in the top panel of table 4.1 and Figure 4.2. The performance of the linear kernel SVM is poor so only the radial kernel SVM results are shown in the figure. On average, the two radial kernel SVMs with and without random effects have better accuracy (1-misclassification rate), sensitivity and negative predictive value (NPV). The specificity and positive predictive value (PPV) is slightly lower. Including random effects in the prediction improves accuracy and leads to smaller variability over repeated simulations. Similar phenomenon holds for other indices.

In type B prediction, we predicted the future follow-up outcomes for the same subject based on his/her observed features variables and prior visits' outcomes and the trained model. In this case each subject was generated to have 6 visits. The first 3 visits of each subject are used as the model-building set and the rest 3 visits as the testing set. In this case we can compute the fitted random effects for each subject using the designed kernel functions, and the subject-specific outcomes for the last 3 visits can be predicted for each subject incorporating fitted random effects. The results are summarized in the bottom panel of table 4.1 and Figure 4.2. Here we see more improvement for SVM-based approach compared to the generalized mixed effects regression or logistic regression, and again extracting information from the distribution of random effects leads to smaller variability for each of the performance index.

4.3.2 Setting 2: multiple data sources

In order to mimic the real data application where the data are complex and from heterogeneous sources, we generated the dichotomous outcomes from the following model:

$$Y_{ij} = \text{sign}\{\beta_0 T_{ij} + \boldsymbol{\beta}_1^T \boldsymbol{Z}_i + \boldsymbol{\beta}_2^T \boldsymbol{X}_{ij}^* + \beta_3 W_i^* + a_{ij} + b_{ij} + \epsilon_{ij}\},$$

where T_{ij} is the age of the i^{th} subject at the j^{th} visit. The age ranges from 10 years old to 70 years old uniformly, and two subsequent visits of a subject have a distance of around 3 years in age. Here $\boldsymbol{Z}_i$ is a vector of time-invariant binary markers of the i^{th} subject which remain the same at each visit; $\boldsymbol{X}_{1i}$ is a vector of time-invariant continuous markers of i^{th} subject uniformly ranging from -2 to 2; and $\boldsymbol{X}_{2ij}$ is a vector of time-varying continuous markers with a correlation $\rho(X_{2ij}, X_{2ik}) = \exp(-\alpha|t_{ij} - t_{ik}|)$ with $\alpha = 1$ between the j^{th} and k^{th} visits of the i^{th} subject. Vector $\boldsymbol{X} = (\boldsymbol{X}_1^*, \boldsymbol{X}_2^*)$ are the mapping of $(\boldsymbol{X}_1, \boldsymbol{X}_2)$ in the new feature space corresponding to a polynomial kernel with degree 2, e.g., the inner product $< u^*, v^* >$ in the feature space equals $K(u, v)$ in the original space, where K is a polynomial kernel with degree 2. In Figure 4.3 we demonstrated a typical set of $\boldsymbol{X}$ when its dimension is 2. The boundary for the two groups is nonlinear in the original space (top panel), while in the new 3-dimensional feature space the boundary becomes a separating plane which is linear (bottom panel). Markers $\boldsymbol{W}_i$ is a 3-dimensional vector generated in the same way as in the single source simulation (Figure 4.1), except that in this setting $\boldsymbol{W}$ is time-invariant, which means that it varies between subjects but not among visits from the same subject. Therefore the corresponding oracle kernels to use for the fixed effects in this setting are a linear kernel for T, a linear kernel for $\boldsymbol{Z}$, a polynomial kernel with degree 2 for $\boldsymbol{X}$, and a radial kernel for $\boldsymbol{W}$. Subject-specific random effects $\boldsymbol{a_i}$ and $\boldsymbol{b_i}$ were generated in the same way as in the single source simulation.

We also conducted two types of prediction for different purposes. In type A prediction we generated samples with a size of $n = 500$ subjects, each having 4 visits. Two-thirds of the subjects are included in the training set and the rest one-third as the testing set. We present the results in Figure 4.4. In the top panel we compared a single radial kernel SVM (concatenate all feature variables in a single radial kernel), a multiple radial kernel SVM (one separate radial kernel for each group of variables) and a multiple fused kernel SVM (combination of linear, polynomial and radial kernels) with and without accounting for random effects. In this case, the logistic regression without random effects and the generalized mixed effects regression perform substantially worse than the SVM based methods in terms of all fit indices (accuracy, sensitivity, specificity, PPV and NPV). In addition, the variability of the former two approaches are much larger than the latter, indicating that the SVM based methods provide more stable predictions.

Comparing four SVM-based approaches, the single radial kernel SVM performs the worst (results for the single linear or polynomial kernel are even worse than using radial kernel, so they are not shown here), indicating the advantage for using separate kernels for fixed effects when data are heterogenous. Using multiple fused kernels (different types of kernels, oracle) greatly improves the performance comparing to using multiple radial kernels (same type of kernels), which confirms the importance of using appropriate kernels for data from different sources. When comparing the performance of multiple fused kernel SVM with and without random effects, we see that including kernels for random effects reduces variability for all fit indices and improves or maintains their mean values.

In type B prediction we generated samples with a size of $n = 500$ subjects, each having 6 visits. The first 3 visits of each subject are used as the training set and the rest 3 visits as the testing set. We predicted the subject-specific outcomes for the last 3 visits for each subject. The bottom panel of Figure 4.4 compares the performance of multiple fused kernel SVM with or without random effects to logistic regression and generalized mixed effects regression. The improvement of including random effects is greater than that in type A prediction, suggesting that the developed method is more powerful when predicting subject-specific outcomes when some outcomes on the prior visits of the same subject are available.

4.4 Application to two epidemiological studies

4.4.1 PREDICT-HD study

We applied the developed method to PREDICT-HD (Paulsen et al., 2008), a multi-center observational study on Huntington's disease (HD). HD is an autosomal dominant disease caused by an expansion of CAG trinucleotide repeats in ITI5 gene on chromosome 4 (Huntington's Disease Collaborative Research Group, 1993). Majority of subjects with an expansion of CAG repeats in IT15 gene (CAG repeats ≥ 36) on one allele will develop HD if not censored by death (Kieburtz and Huntington Study Group, 1996b). It is well established that the risk of HD diagnosis increases with age and CAG repeats length (Zhang et al., 2011a). The diagnosis of HD is based on the diagnostic confidence level (DCL), a measure ranging from 0 to 4 based on the UHDRS assessment. A DCL of 0 means no abnormalities and 4 means motor abnormalities that are unequivocal signs of HD with 99% confidence. Subjects with a DCL of 2 or higher can be considered as showing motor abnormalities that may be signs of HD with more than 50% confidence. There are 941 CAG-expanded participants in the data set who have complete data for analysis. The median age is 40 years old and the range is from 18 to 75. 195 participants have a DCL of 2 or higher at the baseline and totally 126 subjects reached a DCL of 4 during the study.

The goal of PREDICT-HD analysis is to distinguish among those who showed noticeable motor signs of HD from those who did not. The analysis sample included 449 participants who had 4 or more visits, and the outcome of interest is whether a subject had a DCL$\geq$ 2 versus DCL$<$ 2 at each visit. The data sources include demographic data (age, gender and education level), genetic marker (CAG repeat length), motor and functional measures (total motor score and total functional capacity), cognitive function measures (stroop color, digital and word, and symbol digit modalities tests) and psychiatric assessment scores obtained through FRSBE (Frontal Systems Behavior Scale), SCL90 (Global Severity Index, Positive Symptom Total and Distress

Index) and UHDRS (Unified Huntington's disease rating scale). For the multiple fused kernel SVM, we used 5 separate kernels for the feature variables: a linear kernel for age at visit since age appears to be an important biomarker for HD (Chen et al., 2014), a radial kernel for all the continuous clinical measures and cognitive scores, another radial kernel for their interaction with age, a linear kernel for genetic marker and other demographic variables, and another linear kernel for their interaction with age.

We first assessed the type A prediction by treating one third of subjects as new subjects and predicting their outcomes at several visits based on the model trained from the rest of two-thirds subjects. For the standard approaches, we only reported results from logistic regression without random effects since generalized mixed effects model failed to converge due to large number of feature variables included in the model. We compared the performance of five methods: logistic regression, single radial kernel SVM, multiple radial kernel SVM, multiple fused kernel SVM and multiple fused kernel SVM with random effects. The tuning parameter C for cost was selected by five-fold cross-validation. The performance of methods using multiple kernels are much better than using the single kernel in all the measures except for specificity (due to very low sensitivity). For example, the accuracy for the single-kernel SVM is only 0.66 and the sensitivity is 0.15, while for all other methods the accuracy is around 0.85 and the sensitivity is between 0.75 and 0.80. In the top panel of Figure 4.5, we show the performance of the other four methods. We can see that the kernel-based methods perform better the logistic regression in four out of five fit indices (similar sensitivity). Including random effects into multiple fused kernels improves accuracy, specificity and NPV.

Next, we assessed the type B prediction of future observations on existing subjects. We used the first two visits of each subject as the training set to predict the subject-specific outcomes at the rest of follow-up visits. Since the division of training set and testing set is fixed in this setting, we repeated the process n times, taking one subject out each time. From the bottom panel of Figure 4.5, we can see that the accuracy is higher when including random effects, and its standard deviation is much smaller, indicating stability of the results. Although the sensitivity and NPV are slightly lower, the specificity and PPV are much higher.

4.4.2 Application to ADNI data

The Alzheimer's Disease Neuroimaging Initiative (ADNI) was a large joint initiative by National Institute of Health, Food and Drug Administration, private pharmaceutical companies, and nonprofit organizations. It is a naturalistic, non-randomized, non-treatment study in which a total of 800 subjects including 200 normal controls, 400 individuals with mild cognitive impairment (MCI), and 200 subjects with mild Alzheimer's Disease (AD) recruited at approximately 50 sites in the United States and Canada for longitudinal follow-up. MCI is a transition state between the age-related decline in cognitive functions and clinically diagnosed features of AD (Petersen, 2007). The goal of the study is to test whether a combination of MRI, positron emission tomography (PET), other biological markers, genetic markers, and clinical and neuropsychological assessments can be used to track the progression of MCI and early AD. There are three phases of ADNI study: ADNI1, ADNI GO and ADNI2. Further study design information is provided at http://www.adni-info.org/, and detailed clinical characteristics of the ADNI sample are in Mueller et al. (2005) and Petersen et al. (2010).

According to the ADNI protocol, all the subjects had clinical and cognitive assessments and 1.5 T structural MRI at specified intervals (6 or 12 month) for 2-3 years. Approximately 50% of the subjects also had PET scans at the same time intervals and 25% of the subjects (who have not been scanned using PET) would have MRI at 3 Tesla. MCI subjects at high risk for conversion to AD would be studied at 0, 6, 12, 18, 24 and 36 months. Age matched controls would be studied at the same assessment points. Detailed information regarding MRI and imaging protocol are presented in

Clifford R. Jack et al. (2008). To the best of our knowledge, very few of the existing analyses of ADNI data has optimally taken advantage of the longitudinal MCI status assessments in a statistical learning framework.

In 2009, efforts to integrate genetic research related to ADNI biomarkers were planned and carried out to assess genes beyond ApoE, the largest known genetic risk factor for AD (Ashford, 2004). Since then, genetic and imaging data are available to contribute to the understanding of biological etiology of AD and MCI. The proposed multiple kernel framework exploits this unique opportunity to combine imaging and genetic data to predict the progression of MCI and early AD. Previous studies showed that some imaging biomarkers are important in predicting conversion from MCI to AD and early AD progression (Devanand et al., 2008; Hampel et al., 2008; Nestor et al., 2008). It is conceivable that imaging variables are more correlated with each other than with genetic markers. If both types of data are concatenated in a single kernel, for instance, a polynomial kernel, unnecessary polynomial correlation will be imposed between imaging and genetic markers. In a multiple kernel learning with separate kernels, however, such correlation is reduced, avoiding overfitting and unwanted complexity. In our framework, one could use existing kernels designed for imaging data and genetic data separately. Such analyses has not been reported in ADNI literature before.

Our analysis goal is to distinguish the subjects who have MCI and the subjects who have dementia using demographic, clinical, imaging, and genetic markers. The key data were merged from various case report forms and biomarker lab measures across the ADNI protocols by ADNI investigators and posted to ADNI website (http://www.adni-info.org/). Our further inclusion criteria of samples were: subject's disease status being MCI or dementia, having 4 or more follow-up records, and having complete imaging and genetic data. The sample used in our analysis contains 213 participants from all 3 phases with 1055 longitudinal follow-up records.

The feature variables we used include demographic variables (age, gender, and education level), clinical variables (clinical dementia rating sum of boxes scores (CDR-SB), the Alzheimer's Disease Assessment Scale (11 and 13), mini-mental state examination (MMSE), Rey auditory verbal learning test (RAVLT) (forgetting and immediate) and functional assessment questionnaire (FAQ)), imaging markers (volume measures of ventricles, hippocampus, entorhinal cortex, and intra-cranial volume (ICV)), and genetic markers (ApoE4 and 16 SNPs on the PICALM gene). The PICALM gene was reported to be a causal gene for AD (Harold et al., 2009), and therefore the SNPs in this gene were included in our analyses. We used four separate kernels for each source of variables in the multiple fused kernel SVM: a polynomial kernel with degree two for age at each visit, a radial kernel for demographic variables and clinical variables, a linear kernel for imaging variables, and an identity-by-state (IBS) kernel for genetic markers. The IBS kernel is specially designed to measure the similarity between two subjects' SNPs based on their identity by state information and has been proven to be useful in genome-wide association studies (Wu et al., 2010).

The top panel of Figure 4.6 summarized the results of logistic regression, single radial kernel SVM, multiple fused kernel SVM with and without random effects for type A prediction. The performance of multiple kernel SVMs improve upon the logistic regression in terms of all the fit indices, and upon the single radial kernel SVM in terms of accuracy, specificity, and PPV. Sensitivity of the single kernel SVM is slightly better than multiple kernel SVMs. The inclusion of random latent effects to a multiple fused kernel SVM makes little difference in terms of type A prediction. The bottom panel of Figure 4.6 compares the multiple fused kernel SVM with and without random effects for type B prediction. In this case, accounting for random effects in the multiple fused kernel leads to a substantial gain in accuracy, sensitivity and NPV, which reflects the ability of using the latent random effects kernel matrix to extract correlated similarity information of the outcomes on the same subject (within-subject outcomes are often similar to some extent). In this example, the fixed effects feature variables explained some proportion of variability while the latent effects improve prediction by extracting information from the unexplained variability in type B prediction. Specificity and PPV for the multiple SVM incorporating random effects is slightly lower, however, to a much lesser extent.

4.5 Discussion

In this work, we present new methods for statistical learning with random effects for longitudinal data. For analyzing longitudinal data, conventional approaches such as generalized mixed effects regression may fail to converge especially when a large number of predictors are included. Marginal approaches are alternatives, however, they aim at population average effects and may lead to inferior results. Our proposed statistical learning method offers an effective alternative especially when the number of predictors is large. A key feature is to embed correlation of longitudinal observations into kernel matrices and take advantage of multiple kernel learning methodologies. With a single data source, the classical methods perform adequately. However, when there are multiple heterogeneous data sources, the improvement of the proposed method is more evident. Making connections to multiple kernel learning allows proposed method to enjoy easy integration of heterogeneous data sources to boost information while accounting for longitudinal data structure. We have shown through our simulations and real data analyses that when prior scientific knowledge suggests distinct distribution of feature variables, treating each component with a separate kernel and then combine in an optimal way allows substantial information gain.

We discuss two types of prediction problems here. We show that by extracting information on the distributions of the random effects, we improve prediction both for future subjects and for future outcomes on the same subject given feature variables and past outcomes. However, for longitudinal studies, the type B problems are more commonly encountered in applications where the outcome at a follow-up visit for the same subject is desirable, and our learning method is more effective than ignoring correlation among observations. When the interest is on predicting outcomes for a new subject at the baseline, conventional approaches may work as well. The choice of covariance structure and the choice of appropriate kernel functions is related to the choice of the best representation of the kernel space. There is no consensus on these issues in the current literature which warrants future study on these matters.

We adopt the use of L_2-norm kernel fusion which leads to a non-sparse integration of multiple data sources, which may be more appealing in biomedical applications where it is believed there is no clear “winner” and each data modality contributes partial information to the prediction. Besides the L_2-norm on weights θ_p, other regularization, such as L_1-norm and L_∞-norm, can also be imposed in the kernel fusion. L_1-norm generates a sparse integration, which can be used for data source selection when the number of data sources is large and no prior information on which source is more predictive is available. L_∞-norm assigns the dominantly weight parameter to only one kernel, which can be used when there is the need for a unique data source competition.

Lastly, for the proposed method, the decision function takes an additive structure of the feature variables and the latent effects. A natural extension will be to include the interactions between them in the prediction rule. The proposed algorithm can be easily modified to handle this issue through tensor products of kernel matrices. Here we do not assume a distribution for random effects, but uses kernel functions to capture correlation. The kernel matrices for $\mathbf{a}_i$ and $\mathbf{b}_i$ may be misspecified so that it will be interesting to study the robustness of the prediction rule to the specification of these matrices.

Table 4.1: Simulation: single data source)

	type A prediction				
	misclassification	sensitivity	specificity	PPV	NPV
logistic regression	0.0609 (0.0163)	0.9339 (0.0254)	0.9443 (0.0206)	0.9433 (0.0208)	0.9352 (0.0242)
generalized mixed effects regression	0.0613 (0.0156)	0.9316 (0.0249)	0.9458 (0.0202)	0.9446 (0.0203)	0.9332 (0.0237)
single linear kernel	0.4482 (0.0331)	0.3950 (0.0652)	0.7085 (0.0645)	0.5768 (0.0612)	0.5398 (0.0354)
single linear kernel with random effects	0.4867 (0.0431)	0.4960 (0.1678)	0.5310 (0.2136)	0.5337 (0.0889)	0.4985 (0.0706)
single radial kernel	0.0591 (0.0163)	0.9450 (0.0228)	0.9370 (0.0206)	0.9373 (0.0208)	0.9445 (0.0230)
single radial kernel with random effects	0.0572 (0.0155)	0.9469 (0.0219)	0.9387 (0.0208)	0.9391 (0.0207)	0.9465 (0.0220)

	type B prediction				
	misclassification	Sensitivity	Specificity	PPV	NPV
logistic regression	0.0590 (0.0092)	0.9334 (0.0141)	0.9485 (0.0122)	0.9476 (0.0126)	0.9347 (0.0130)
generalized mixed effects regression	0.0618 (0.0097)	0.9322 (0.0136)	0.9440 (0.0149)	0.9434 (0.0145)	0.9332 (0.0123)
single linear kernel	0.4497 (0.0223)	0.4093 (0.0397)	0.6911 (0.0549)	0.5720 (0.0371)	0.5388 (0.0251)
single linear kernel with random effects	0.4676 (0.0237)	0.5400 (0.0912)	0.5236 (0.0821)	0.5312 (0.0310)	0.5342 (0.0302)
single radial kernel	0.0553 (0.0091)	0.9482 (0.0120)	0.9414 (0.0122)	0.9416 (0.0131)	0.9478 (0.0122)
single radial kernel with random effects	0.0547 (0.0088)	0.9487 (0.0121)	0.9420 (0.0118)	0.9423 (0.0122)	0.9484 (0.0121)

Figure 4.1: A typical set of simulated data (3-dimensional vector W).

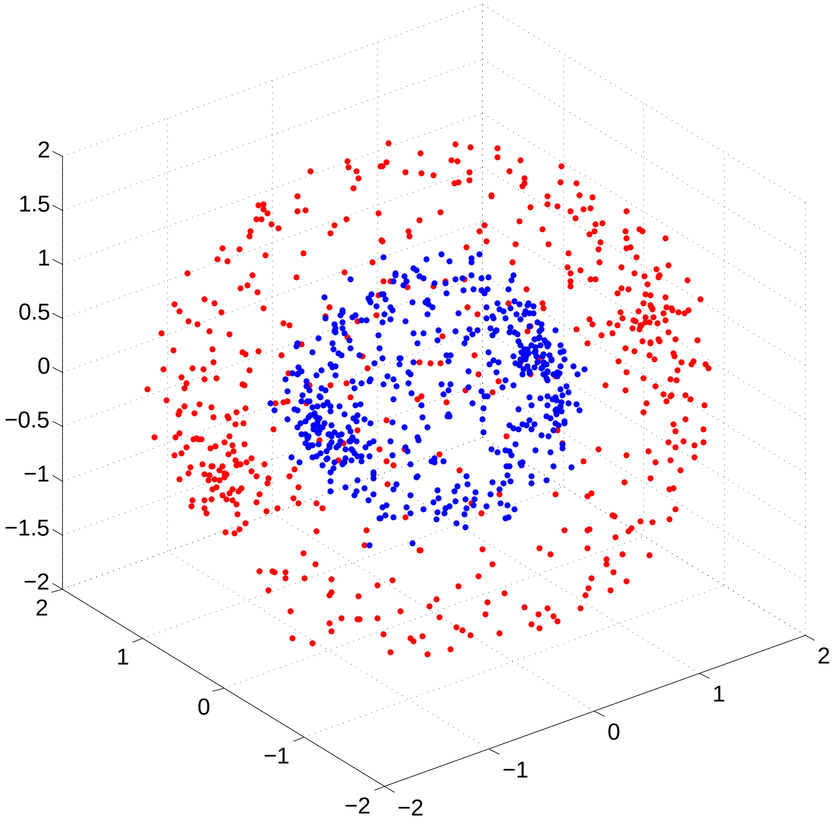


Figure 4.2: Simulation setting 1 (single data source). Top panel presents type A prediction of new subjects (left to right): 1-logistic regression, 2-generalized mixed effects regression, 3-single radial kernel SVM, 4-single radial kernel SVM with random effects. Bottom panel presents type B prediction of outcomes at future visits on the same subjects (labels same as top the panel).

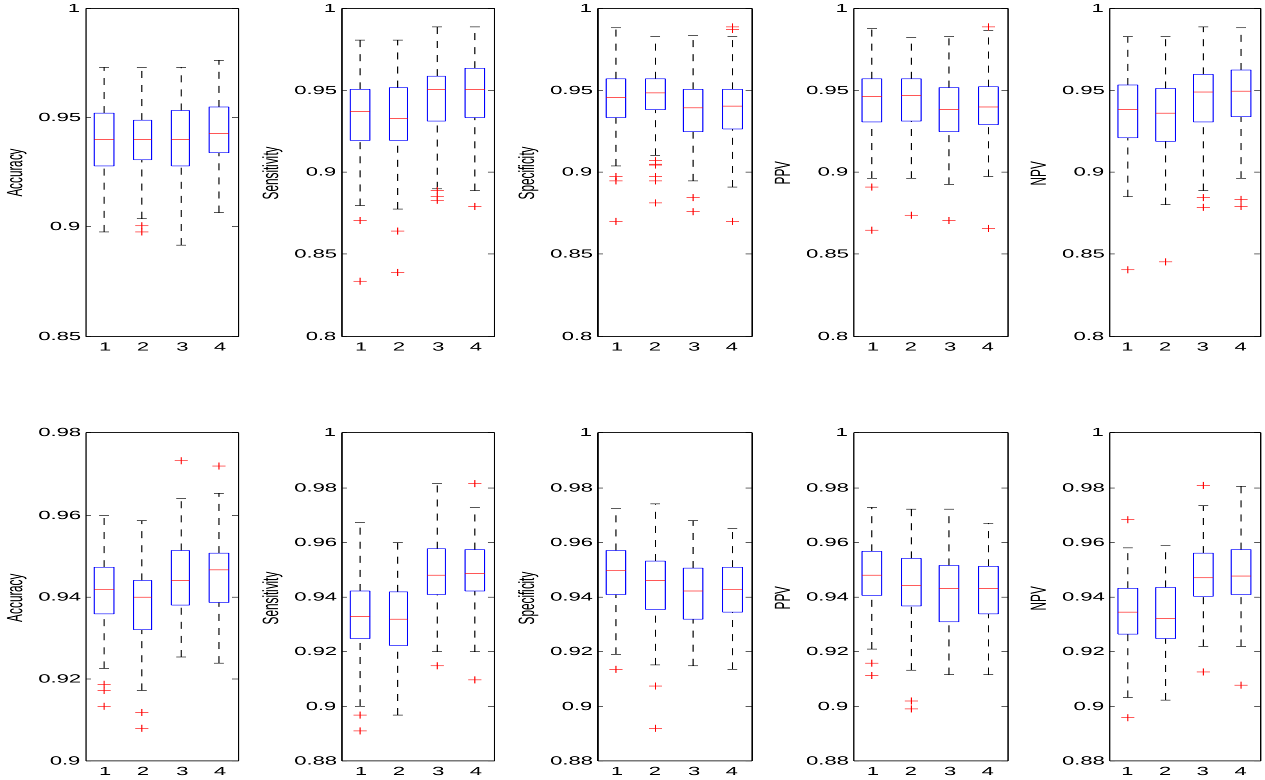


Figure 4.3: A typical set of simulated data (2-dimensional vector X). Top panel: nonlinear boundary in original space. Bottom panel: linear boundary (separating plane) in new 3-dimensional space.

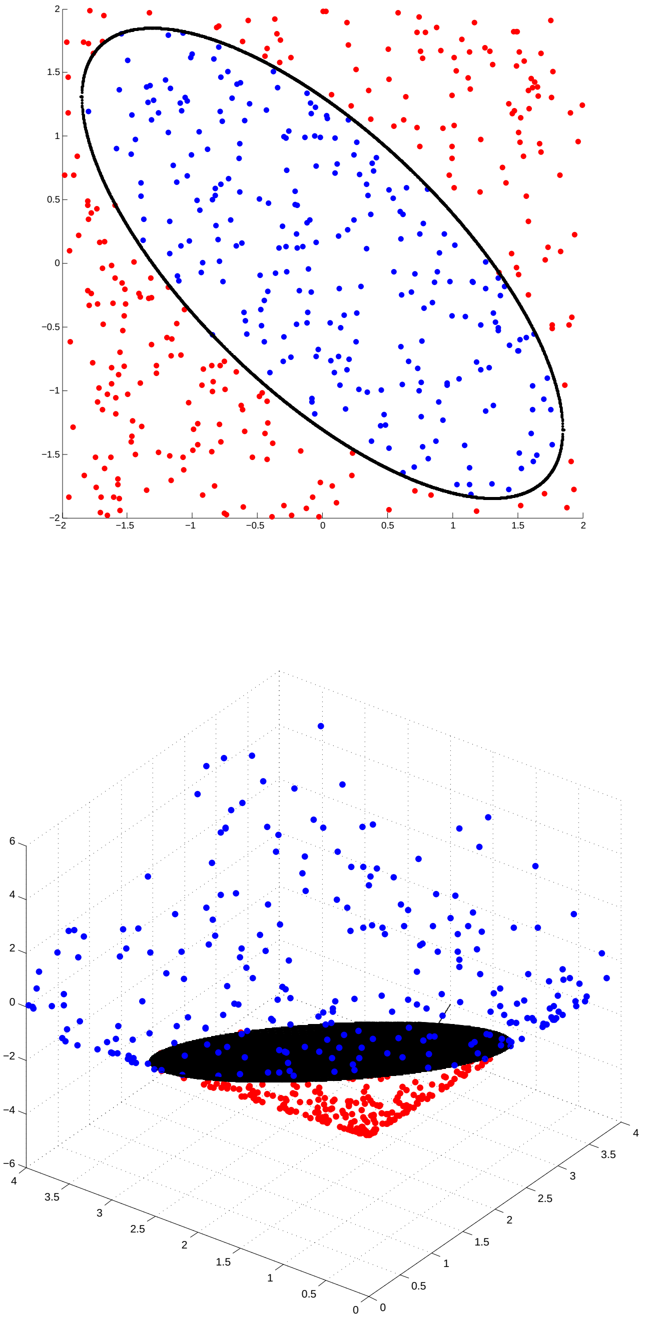


Figure 4.4: Simulation setting 2 (multiple data sources). Top panel presents type A prediction of new subjects (left to right): 1-logistic regression, 2-generalized mixed effects regression, 3-single radial kernel SVM, 4-multiple radial kernel SVM, 5-multiple fused kernel SVM, 6-multiple fused kernel SVMwith random effects. Bottom panel presents type B prediction of outcomes at future visits on the same subjects (left to right): 1-logistic regression, 2-generalized mixed effects regression, 3-multiple fused kernel SVM, 4-multiple fused kernel SVM with random effects.

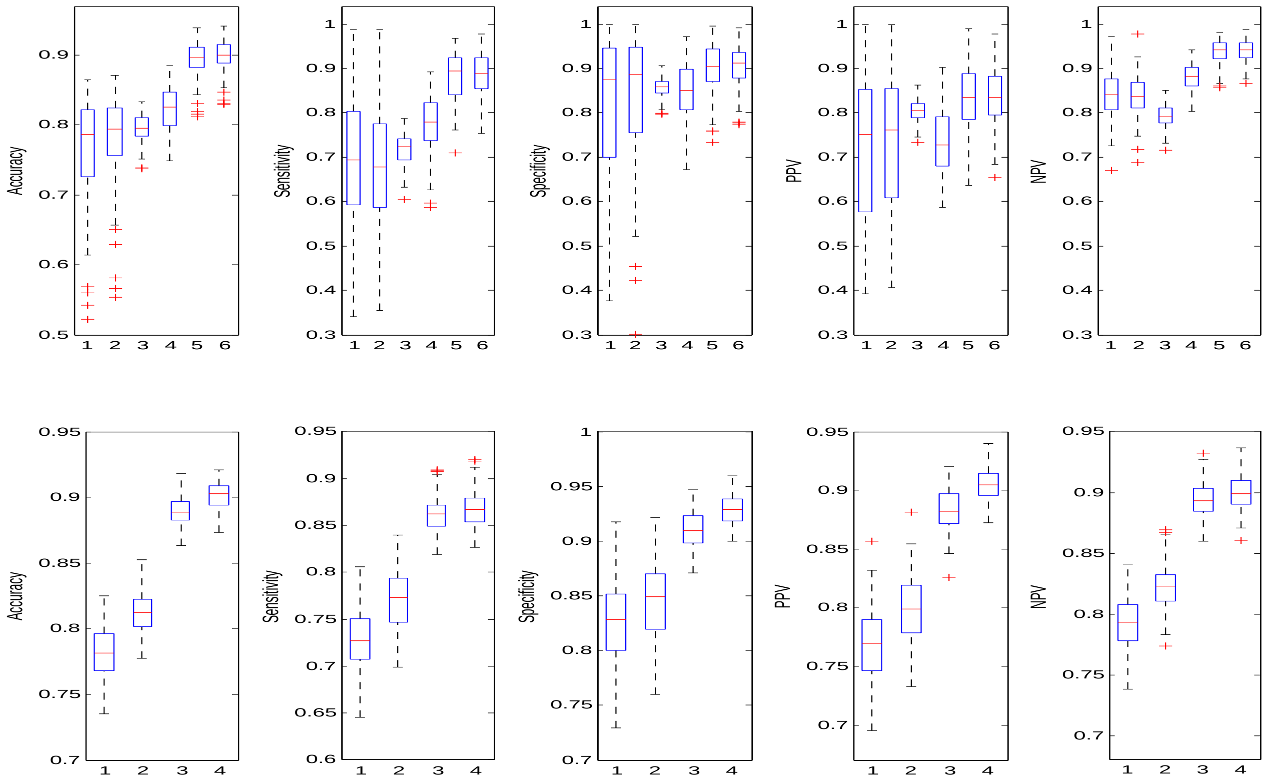


Figure 4.5: PREDICT-HD study. Top panel presents type A prediction of new subjects (left to right): 1-logistic regression, 2-multiple radial kernel SVM, 3-multiple fused kernel SVM, 4-multiple fused kernel SVM with random effects. Bottom panel presents type B prediction of outcomes at future visits on the same subjects (left to right): 1-multiple fused kernel SVM, 2-multiple fused kernel SVM with random effects.

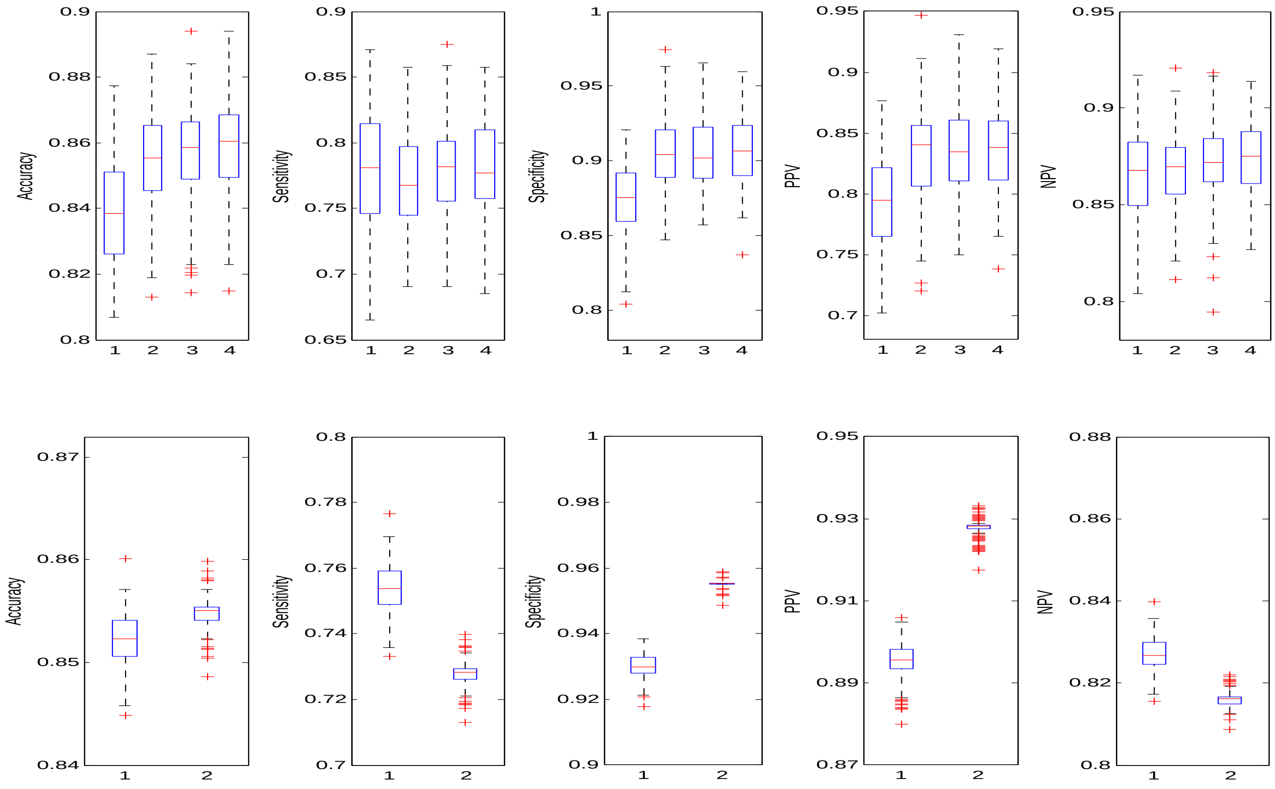
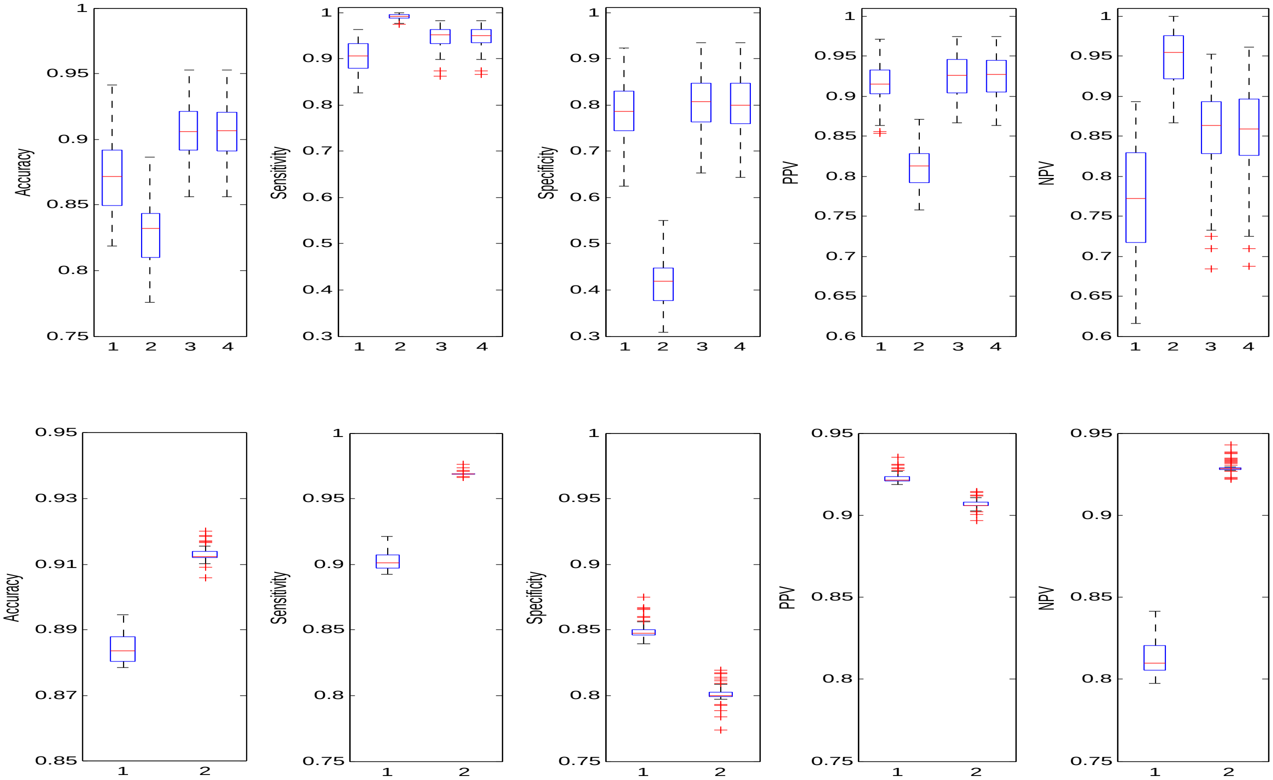


Figure 4.6: ADNI study. Top panel presents type A prediction of new subjects (left to right): 1-logistic regression, 2-single radial kernel SVM, 3-multiple fused kernel SVM, 4-multiple fused kernel SVM with random effects. Bottom panel presents type B prediction of outcomes at future visits on the same subjects (left to right): 1-multiple fused kernel SVM, 2-multiple fused kernel SVM with random effects.



Part I

Bibliography

Bibliography

Ashford, J. W. (2004). Apoe genotype effects on alzheimers disease onset and epidemiology. *Journal of Molecular Neuroscience*, 23(3):157–165.

Bach, F. R. and Lanckriet, G. R. G. (2004). Multiple kernel learning, conic duality, and the smo algorithm. In *In Proceedings of the 21st International Conference on Machine Learning (ICML)*.

Bang, H. and Tsiatis, A. A. (2000). Estimating medical costs with censored data. *Biometrika*, 87(2):329–343.

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2006). Convexity, classification and risk bounds. *Journal of the American Statistical Association*, 101:138–156.

Brinkman, R. R., Mezei, M. M., Theilmann, J., Almqvist, E., and Hayden, M. R. (1997). The likelihood of being affected with huntington disease by a particular age, for a specific cag size. *American Journal of Human Genetics*, 60:1202–1210.

Cai, Z., Fan, J., and Li, R. (2000). Efficient estimation and inferences for varying coefficient models. *Journal of the American Statistical Association*, 95:888–902.

Celsis, P. (2000). Age-related cognitive decline, mild cognitive impairment or preclinical alzheimer's disease? *Annals of Medicine*, 32:6–14.

Chen, T., Wang, Y., Chen, H., Marder, K., and Zeng, D. (2014). Targeted local support vector machine for age-dependent classification. *Journal of the American Statistical Association*, In press.

Chen, X., Linton, O., and Keilegom, I. V. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*, 71(5):1591–1608.

Chen, Y. Q., Hu, N., Cheng, S.-C., Musoke, P., and Zhao, L. P. (2012). Estimating regression parameters in an extended proportional odds model. *Journal of the American Statistical Association*, 107:318–330.

Clifford R. Jack, J., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., et al. (2008). The alzheimer's disease neuroimaging initiative (adni): Mri methods. *J Magn Reson Imaging*, 27(4):685–691.

de Leeuw, J., Hornik, K., and Mair, P. (2009). Isotone optimization in r: Pool-adjacent-violators algorithm (pava) and active set methods. *Journal of Statistical Software*, 32:1–24.

Devanand, D. P., Liu, X., Tabert, M. H., Pradhaban, G., Cuasay, K., Bell, K., de Leon, M. J., Doty, R. L., Stern, Y., and Pelton, G. H. (2008). Combining early markers strongly predicts conversion from mild cognitive impairment to alzheimer's disease. *Biological Psychiatry*, 64:871–879.

Diggle, P., Heagerty, P., Liang, K.-Y., and Zeger, S. (2002). *Analysis of Longitudinal Data (second edition)*. Oxford: Oxford University Press.

Dorsey, E. R., Beck, C., Adams, M., and Huntington Study Group (2008). Trend-hd communicating clinical trial results to research participants. *Archives of Neurology*, 65(12):1590–1595.

Dorsey, E. R. and Huntington Study Group COHORT Investigators (2012). Characterization of a large group of individuals with huntington disease and their relatives enrolled in the cohort study. *PLoS ONE*, 7(2, Article ID e29522).

Efron, B. (1967). The two sample problem with censored data. *Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 4.

Foroud, T., Gray, J., Ivashina, J., and Conneally, P. M. (1999). Differences in duration of huntington's disease based on age at onset. *Journal of Neurology, Neurosurgery & Psychiatry*, 66:52–56.

Gutierrez, C. and MacDonald, A. (2004). Huntington's disease, critical illness insurance and life insurance. *Scandinavian Actuarial Journal*, 4:279–313.

Ha, A. D., Beck, C. A., and Jankovic, J. (2012). Intermediate cag repeats in huntingtons disease: Analysis of cohort. *Tremor and Other Hyperkinetic Movements*, tre-02-64-287-4.

Hampel, H., Brgerb, K., Teipelb, S. J., Bokdea, A. L., Zetterbergc, H., and Blennow, K. (2008). Core candidate neurochemical and imaging biomarkers of alzheimers disease. *Dementia*, 4:38–48.

Harold, D., Abraham, R., Hollingworth, P., Sims, R., et al. (2009). Genome-wide association study identifies variants at clu and picalm associated with alzheimer's disease. *Nature Genetics*, 41:1088–1093.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer.

Huntington's Disease Collaborative Research Group (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on huntingtons disease chromosomes. *Cell*, 72:971–983.

James, G. M. and Hastie, T. J. (2001). Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(3):533–550.

Jung, S.-H. (1996). Regression analysis for long-term survival rate. *Biometrika*, 83:227–232.

Kanga, J., Chobc, J., and Zhao, H. (2010). Practical issues in building risk-predicting models for complex diseases. *Journal of Biopharmaceutical Statistics*, 20(2):415–440.

Kieburtz, K. and Huntington Study Group (1996a). The unified huntington's disease rating scale: Reliability and consistency. *Movement Disorder*, 11:136–142.

Kieburtz, K. and Huntington Study Group (1996b). The unified huntington's disease rating scale: reliability and consistency. *Movement Disorders*, 11:136–142.

Kimeldorf, G. and Wahba, G. (1970). A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41:495C502.

Ladicky, L. and Torr, P. H. S. (2011). Locally linear support vector machines. In *ICML2011*, pages 985–992.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974.

Lanckriet, G. R., Cristianini, N., Bartlett, P., Ghaoui, L. E., and Jordan, M. I. (2004). Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72.

Langbehn, D. R., Brinkman, R. R., Falush, D., Paulsen, J. S., and Hayden, M. R. (2004). A new model for prediction of the age of onset and penetrance for huntington's disease based on cag length. *Clinical Genetics*, 65:267–277.

Langbehn, D. R., Hayden, M. R., Paulsen, J. S., and the PREDICT-HD Investigators of the Huntington Study Group (2010a). Cag-repeat length and the age of onset in huntington disease (hd): A review and validation study of statistical approaches. *American Journal of Medical Genetics Part B*, 153B:397–408.

Langbehn, D. R., Hayden, M. R., Paulsen, J. S., and the PREDICT-HD Investigators of the Huntington Study Group (2010b). Cag-repeat length and the age of onset in huntington disease (hd): A review and validation study of statistical approaches. *American Journal of Medical Genetics Part B*, 153B:397–408.

Langbehn, D. R., Paulsen, J. S., and Huntington Study Group (2007). Predictors of diagnosis in huntington disease. *Neurology*, 68(20):1710–1717.

Lin, Y. (2002). Support vector machines and the bayes rule in classification. *Data Mining and Knowledge Discovery*, 6:259–275.

Luts, J., Molenberghs, G., Verbeke, G., Van Huffel, S., and Suykens, J. A. (2012). A mixed effects least squares support vector machine model for classification of longitudinal data. *Computational Statistics & Data Analysis*, 56(3):611–628.

Ma, Y. and Wei, Y. (2012). Analysis on censored quantile residual life model via spline smoothing. *Statistica Sinica*, 22:47–68.

Marshall, G. and Baron, A. E. (2000). Linear discriminant models for unbalanced longitudinal data. *Statistics in medicine*, 19(15):1969–1981.

McNeil, S. M., Novelletto, A., Srinidhi, J., Barnes5, G., Kornbluth, I., Altherr, M. R., Wasmuth, J. J., Gusella, J. F., MacDonald, M. E., and Myers, R. H. (1997). Reduced penetrance of the huntington's disease mutation. *Human Molecular Genetics*, 6(5):775–779.

Mitrushina, M., Boone, K. B., Razani, J., and D'Elia, L. F. (2005). *Handbook of normative data for neuropsychological assessment*. New York: Oxford University Press.

Moguerza, J. M. and Munoz, A. (2006). Support vector machines with applications. *Statistical Science*, 21(3):322–336.

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C., Jagust, W., Trojanowski, J. Q., Toga, A. W., and Beckett, L. (2005). The alzheimers disease neuroimaging initiative. *Neuroimaging Clinics of North America*, 15(4):869–877.

Nestor, S. M., Rupsingh, R., Borrie, M., Smith, M., Accomazzi, V., L.Wells, J., Fogarty, J., Bartha, R., and the Alzheimers Disease Neuroimaging Initiative (2008). Ventricular enlargement as a possible measure of alzheimers disease progression validated using the alzheimers disease neuroimaging initiative database. *Brain*, 131:2443–2454.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica*, 62:1349–1382.

Oquendo, M. A., Baca-Garcia, E., Artes-Rodriguez, A., Perez-Cruz, F., Galfalvy, H. C., Blasco-Fontecilla, H., , and Duan, N. (2012). Machine learning and data mining: strategies for hypothesis generation. *Molecular Psychiatry*, 17(10):956–959.

Orru, G., Pettersson-Yeoa, W., Marquand, A. F., Sartori, G., and Mechelli, A. (2012). Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: A critical review. *Neuroscience and Biobehavioral Reviews*, 36(4):1140–1152.

Paulsen, J., Hayden, M., Stout, J. C., Langbehn, D. R., Aylward, E., Ross, C. A., Guttman, M., Nance, M., Kieburtz, K., Oakes, D., Shoulson, I., Kayson, E., Johnson, S., Penziner, E., and Predict-HD Investigators of the Huntington Study Group (2006). Preparing for preventive clinical trials: the predict-hd study. *Archives of Neurology*, 65(6):883–890.

Paulsen, J. S., Langbehn, D. R., Stout, J. C., Aylward, E., Ross, C. A., Nance, M., Guttman, M., Johnson, S., MacDonald, M., Beglinger, L. J., Duff, K., Kayson, E., Biglan, K., Shoulson, I., Oakes, D., and Hayden, M. (2008). Detection of huntington's disease decades before diagnosis: the predict-hd study. *Journal of Neurology, Neurosurgery & Psychiatry*, 79:874–880.

Pavlidis, P., Cai, J., Weston, J., and Noble, W. S. (2002). Learning gene functional classifications from multiple data types. *Journal of Computational Biology*, 9(2):401–411.

Pearce, N. and Wand, M. (2009). Explicit connections between longitudinal data analysis and kernel machines. *Electronic Journal of Statistics*, 3:797–823.

Peng, L. and Huang, Y. (2007). Survival analysis with temporal covariate effects. *Biometrika*, 94:719–733.

Pepe, M. S., Cai, T., and Longton, G. (2006). Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics*, 62:221–229.

Pepe, M. S., Janes, H., Longton, G., Leisenring, W., and Newcomb, P. (2004). Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *American Journal of Epidemiology*, 159:882–890.

Petersen, R. C. (2007). Mild cognitive impairment: current research and clinical implications. *Semin Neurol*, 27(1):22–31.

Petersen, R. C., Aisen, P. S., Beckett, L. A., Donohue, M. C., Gamst, A. C., Harvey, D. J., Jr, C. R. J., Jagust, W. J., Shaw, L. M., Toga, A. W., Trojanowski, J. Q., and Weiner, M. W. (2010). Alzheimers disease neuroimaging initiative (adni): clinical characterization. *Neurology*, 74(3):201–209.

Portnoy, S. (2003). Censored regression quantiles. *Journal of the American Statistical Association*, 98(464):1001–1012.

Ross, C. A. (1995). When more is less: pathogenesis of glutamine repeat neurodegenerative diseases. *Neuron*, 15(3):493–496.

Ross, C. A. and Tabrizi, S. J. (2011). Huntington's disease: from molecular pathogenesis to clinical treatment. *The Lancet Neurology*, 10:83–98.

Rubinsztein, D. C., Leggo, J., Coles, R., Almqvist, E., and et al (1996). Phenotypic characterization of individuals with 30-40 cag repeats in the huntington disease (hd) gene reveals hd cases with 36 repeats and apparently normal elderly individuals with 36-39 repeats. *American Journal of Human Genetics*, 59(1):16–22.

Shen, X., Tseng, G. C., Zhang, X., and Wong, W. H. (2003). On ψ-learning. *Journal of the American Statistical Association*, 98:724–734.

Smith, A. (1982). *Symbol digit modalities test: Manual.* Los Angeles: Western Psychological Services.

Steinwart, I., Hush, D., and Scovel, C. (2006). An explicit description of the reproducing kernel hilbert spaces of gaussian rbf kernels. *IEEE Transactions on Information Theory*, 52:4635–4643.

Steinwart, I. and Scovel, C. (2007). Fast rates for support vector machines using gaussian kernels. *Annals of Statistics*, 35:575–607.

Stine, O. C., Pleasant, N., Franz, M. L., Abbott, M. H., Folstein, S. E., and Ross, C. A. (1993). Correlation between the onset age of huntingtons disease and length of the trinucleotide repeat in it-15. *Human Molecular Genetics*, 2(10):1547–1549.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology: General*, 18:643–662.

Van der Vaart, A. W. and Weller, J. (1996). *Weak Convergence and Empirical Processes.* New York: Springer-Verlag.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. New York: Springer-Verlag.

Verbeke, G. and Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, 91(433):217–221.

Wahba, G. (1990). *Spline Models for Observational Data*. SIAM.

Wang, H. J. and Wang, L. (2009). Locally weighted censored quantile regression. *Journal of the American Statistical Association*, 104(487):1117–1128.

Wang, J., Shen, X., and Pan, W. (2009a). On efficient large margin semisupervised learning: Method and theory. *Journal of Machine Learning Research*, 10:719–742.

Wang, L., Kai, B., and Li, R. (2009b). Local rank inference for varying coefficient models. *Journal of the American Statistical Association*, 104(488):1631C1645.

Wang, Y., Clark, L. N., Louis, E. D., Mejia-Santana, H., Harris, J., Cote, L. J., Waters, C., Andrews, D., Ford, B., Frucht, S., Fahn, S., Ottman, R., Rabinowitz, D., and Marder, K. (2008). Risk of parkinson's disease in carriers of parkin mutations: estimation using the kin-cohort method. *Archives of Neurology*, 65(4):467–474.

Wang, Y., Garcia, T. P., and Ma, Y. (2012). Nonparametric estimation for censored mixture data with application to the cooperative huntingtons observational research trial. *Journal of the American Statistical Association*, 107(500):1324–1338.

Ware, J. H. (2006). The limitations of risk factors as prognostic tools. *The New England Journal of Medicine*, 355(4):2615–2617.

Wei, Z., Wang, K., Qu, H.-Q., Zhang, H., Bradfield, J., Kim, C., Frackleton, E., Hou, C., Glessner, J. T., Chiavacci, R., Stanley, C., Monos, D., Grant, S. F. A., Polychronakos, C., and Hakonarson, H. (2009). From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. *PLoS Genetics*, 5(10):e1000678.

Williams, J. K., Erwin, C., Juhl, A., Mills, J., Brossman, B., Paulsen, J. S., and the I-RESPOND-HD Investigators of the Huntington Study Group (2010). Personal factors associated with reported benefits of huntington disease family history or genetic testing. *Genetic Testing and Molecular Biomarkers*, 14(5):629–636.

Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanock, S. J., Hunter, D. J., and Lin, X. (2010). Powerful snp-set analysis for case-control genome-wide association studies. *The American Journal of Human Genetics*, 86:929–942.

Wu, Y. and Liu, Y. (2013). Functional robust support vector machines for sparse and irregular longitudinal data. *Journal of Computational and Graphical Statistics*, 22(2):379–395.

Yu, S., Falck, T., Daemen, A., Tranchevent, L.-C., Suykens, J. A., Moor, B. D., and Moreau, Y. (2010). L2-norm multiple kernel learning and its application to biomedical data fusion. *BMC Bioinformatics*, 11:309.

Zeng, D. and Lin, D. Y. (2007). Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of the Royal Statistical Society, Series B*, 69:1–30.

Zhang, D. and Shen, D. (2012). Predicting future clinical changes of mci patients using longitudinal and multimodal biomarkers. *PloS one*, 7(3):e33182.

Zhang, H. H., Ahn, J., Lin, X., and Park, C. (2006). Gene selection using support vector machine with non-convex penalty. *Bioinformatics*, 22(1):88–95.

Zhang, Y., Long, J. D., Mills, J. A., Warner, J. H., Lu, W., Paulsen, J. S., and the PREDICT-HD Investigators and Coordinators of the Huntington Study Group (2011a). Indexing disease progression at study entry with individuals at-risk for huntington disease. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 156B(7):751–763.

Zhang, Z., Ladicky, L., Torr, P. H. S., and Saffari, A. (2011b). Learning anchor planes for classification. In *NIPS*.

Zhu, J., Rosset, S., Hastie, T., and Tibshirani, R. (2003). 1-norm support vector machines. In *Neural Information Processing Systems*, page 16. MIT Press.

Zou, H. and Yuan, M. (2008). The f_∞-norm support vector machine. *Statistica Sinica*, 18(1):379–398.

Part II

Appendices

Appendix A

Proofs of theorems in Chapter 2

A.1 Proof of Theorem 1

We show consistency by Lemma 5.2 in Newey (1994). We need to show uniform consistency of the initial IPW estimator, i.e., $\sup_{t\in[a,b]}|\widehat{\beta}(t)-\beta(t)|=o_p(1)$, and verify assumption 5.4 and 5.5 in Newey (1994). First show uniform consistency of the initial IPW estimator. Wang et al. (2012) showed that the IPW estimator can be expanded as

$$\widehat{\beta}(t)-\beta(t)=\frac{1}{n}\sum_{i=1}^{n}\psi\{X_i,T_i;t,\beta(t)\}+o_p(n^{-1/2}), \tag{A.1}$$

where

$$\psi\{X_i,T_i;t,\beta(t)\}=\phi\{X_i,T_i;t,\beta(t)\}-\sum_{i=1}^{n}\int\frac{[\phi\{X_i,T_i;t,\beta(t)\}-\mathcal{B}(\phi,u)]dM_i^c(u)}{G(u)},$$

$\mathcal{B}(\phi,u)=E[\phi\{X_i,T_i;t,\beta(t)\}|T_i\geq u,X_i]$, and $dM_i^c(u)$ is the martingale of the censoring process. To show uniform consistency, we need to show that the set $\big\{\psi\{X_i,T_i;t,\beta(t)\}:t\in[a,b]\big\}$ is a Glivenko-Cantelli class. Note that $\phi\{X_i,T_i;t,\beta(t)\}=A\{X_i;\beta(t)\}[I(T_i\leq t)-\mu\{X_i;\beta(t)\}]Z_i$. Indicator functions are cadlag processes which are bounded in total variation and belong to the Vapnic-Červonencis class. Thus they are bounded in uniform entropy integral with square-integrable envelope. It follows that they belongs to a Donsker class, and hence Glivenko-Cantelli.

In addition, $\mu\{X_i;\beta(t)\}$ is Lipschitz continuous. By assumption A1, $\beta(t)$ belongs to a cadlag processes therefore are also bounded in uniform entropy integral. Since Lipschitz continuous functions of classes bounded in uniform entropy integral and pointwise measurable are also bounded in uniform entropy integral and pointwise measurable, $\big\{\mu\{X_i;\beta(t)\}, t\in[a,b]\big\}$, is Glivenko-Cantelli. From $A\{X_i;\beta(t)\} = \Big\{E(\mu\{X_i;\beta(t)\}[1-\mu\{X_i;\beta(t)\}]Z_iZ_i^T)\Big\}^{-1}$, under assumption A5, $A\{X_i;\beta(t)\}$ is bounded from below and above by positive constants component-wise and bounded in uniform entropy integral, therefore is Glivenko-Cantelli. Lastly, since X_i is bounded and products of classes with bounded uniform entropy integral also have bounded uniform entropy integral, we have $\Big\{\phi\{X_i,T_i;t,\beta(t)\}: t\in[a,b]\Big\}$ is Glivenko-Cantelli.

Now we check the second term in $\psi\{X_i,T_i;t,\beta(t)\}$. Note that

$$\begin{aligned}
\mathcal{B}(\phi,u) &= E[\phi\{X_i,T_i;t,\beta(t)\}|X_i,T_i\ge u]\\
&= \frac{E[\phi\{X_i,T_i;t,\beta(t)\}I(T_i\ge u)|X_i]}{E(T_i\ge u|X_i)}\\
&= \frac{\int_u^\infty A\{X_i;\beta(t)\}[I(s\le t)-\mu\{X_i;\beta(t)\}]Z_i dF(s|X_i)}{1-F(u|X_i)}\\
&= \frac{A\{X_i;\beta(t)\}[F(t|X_i)-F(u|X_i)-\mu\{X_i;\beta(t)\}\{1-F(u|X_i)\}]Z_i}{1-F(u|X_i)}.
\end{aligned}$$

Therefore

$$\begin{aligned}
&\sum_{i=1}^n \int \frac{[\phi\{X_i,T_i;t,\beta(t)\}-\mathcal{B}(\phi,u)]dM_i^c(u)}{G(u)}\\
=\ &\sum_{i=1}^n \frac{(1-\delta_i)}{G(C_i)}\Big\{\phi\{X_i,T_i;t,\beta(t)\}\\
&-\frac{A\{X_i;\beta(t)\}[F(t|X_i)-F(C_i|X_i)-\mu\{X_i;\beta(t)\}\{1-F(C_i|X_i)\}]Z_i}{1-F(C_i|X_i)}\Big\}. \quad \text{(A.2)}
\end{aligned}$$

Under condition A4, $G(C_i)>0$. Under model (**??**) and conditions A1, A2, the above term indexed by t is also Glivenko-Cantelli. This proves that $\Big\{\psi\{X_i,T_i;t,\beta(t)\}: t\in[a,b]\Big\}$ is Glivenko-Cantelli. It follows that

$$\sup_{t\in[a,b]}\left|n^{-1}\sum_{i=1}^n \psi\{X_i,T_i;t,\beta(t)\}-E[\psi\{X_i,T_i;t,\beta(t)\}]\right|\to 0.$$

Since $E[\psi\{X_i, T_i; t, \beta(t)\}] = 0$, we have shown the uniform consistency of the IPW estimator,

$$\sup_{t \in [a,b]} |\widehat{\beta}(t) - \beta(t)| = 0.$$

Now we verify assumptions 5.4 and 5.5 in Newey (1994). In what follows, we use θ and $\beta(\cdot)$ to denote true parameter values and use $\widetilde{\theta}$ and $\widetilde{\beta}(\cdot)$ to denote other values different from the truth. For assumption 5.4 (i), it is straightforward to see that $s(O_i; t_0, \widetilde{\theta}, \beta)$ is continuous in $\widetilde{\theta}$ and is bounded under the assumptions A1, A3, and A4. For the assumption 5.4 (ii), note

$$\begin{aligned}
& s(O_i; t_0, \widetilde{\theta}, \widetilde{\beta}) - s(O_i, ; t_0, \widetilde{\theta}, \beta) \\
= \; & I(T_i > C_i) I(C_i < t_0) \{w(O_i; t_0, \widetilde{\beta}) - w(O_i; t_0, \beta)\} Z_i \\
= \; & I(T_i > C_i) I(C_i < t_0) Z_i \left[\frac{\mu\{X_i; \widetilde{\beta}(t_0)\} - \mu\{X_i; \widetilde{\beta}(C_i)\}}{1 - \mu\{X_i; \widetilde{\beta}(C_i)\}} - \frac{\mu\{X_i; \beta(t_0)\} - \mu\{X_i; \beta(C_i)\}}{1 - \mu\{X_i; \beta(C_i)\}} \right] \\
= \; & I(T_i > C_i) I(C_i < t_0) Z_i Z_i^T \Bigg(\frac{\mu\{X_i; \check{\beta}(t_0)\}[1 - \mu\{X_i; \check{\beta}(t_0)\}]}{1 - \mu\{X_i; \check{\beta}(C_i)\}} \{\widetilde{\beta}(t_0) - \beta(t_0)\} \\
& - \frac{\mu\{X_i; \check{\beta}(C_i)\}[1 - \mu\{X_i; \check{\beta}(t_0)\}]}{1 - \mu\{X_i; \check{\beta}(C_i)\}} \{\widetilde{\beta}(C_i) - \beta(C_i)\} \Bigg),
\end{aligned} \tag{A.3}$$

where $\check{\beta}(u)$ is on the line segment between $\widetilde{\beta}(u)$ and $\beta(u)$. Here the last equality is obtained by taking pathwise derivative with respect to β. See also (A.4). Since $0 < \mu\{x; \check{\beta}(u)\} < 1$ for $u \in [a, b]$, it follows that there exists $b(O_i)$ such that componentwise we have

$$||s(O_i; t_0, \theta, \widetilde{\beta}) - s(O_i, ; t_0, \theta, \beta)|| \leq b(O_i) ||\widetilde{\beta} - \beta||.$$

By condition A5, the assumption 5.5 in Newey (1994) is satisfied. Finally, by Lemma 5.2 of Newey (1994), we have $\widehat{\theta}_n = \theta + o_p(1)$.

A.2 Proof of Theorem 2

We show the asymptotic normality of $\widehat{\theta}_n$ by Lemma 5.3 of Newey (1994). For assumption 5.1(i), note again

$$s(O_i; t_0, \theta, \widetilde{\beta}) - s(O_i; t_0, \theta, \beta) = I(T_i > C_i) I(C_i < t_0)\{w(O_i; t_0, \widetilde{\beta}) - w(O_i; t_0, \beta)\} Z_i.$$

We now compute a pathwise derivative of $w(O_i; t_0, \beta)$ w.r.t. β evaluated at the true β in the direction $[\widetilde{\beta} - \beta]$. Let $\beta_\epsilon(u) = \beta(u) + \epsilon\{\widetilde{\beta}(u) - \beta(u)\}$. We can verify that

$$\begin{aligned} & \lim_{\epsilon \to 0} \frac{1}{\epsilon} \left\{ \frac{F(t_0|X_i; \beta_\epsilon) - F(C_i|X_i; \beta_\epsilon)}{1 - F(C_i|X_i; \beta_\epsilon)} - \frac{F(t_0|X_i) - F(C_i|X_i)}{1 - F(C_i|X_i)} \right\} \\ = & \frac{F(t_0|X_i)\{1 - F(t_0|X_i)\} Z_i^T}{1 - F(C_i|X_i)} \{\widetilde{\beta}(t_0) - \beta(t_0)\} \\ & - \frac{F(C_i|X_i)\{1 - F(t_0|X_i)\} Z_i^T}{1 - F(C_i|X_i)} \{\widetilde{\beta}(C_i) - \beta(C_i)\}. \qquad \text{(A.4)} \end{aligned}$$

Let

$$\begin{aligned} D(O_i; \widetilde{\beta} - \beta) &= I(T_i > C_i) I(C_i < t_0) Z_i Z_i^T \Big[\frac{F(t_0|X_i)\{1 - F(t_0|X_i)\}}{1 - F(C_i|X_i)} \{\widetilde{\beta}(t_0) - \beta(t_0)\} \\ & \quad - \frac{F(C_i|X_i)\{1 - F(t_0|X_i)\}}{1 - F(C_i|X_i)} \{\widetilde{\beta}(C_i) - \beta(C_i)\} \Big]. \qquad \text{(A.5)} \end{aligned}$$

From (A.3), we can verify

$$\begin{aligned} & s(O_i; t_0, \theta, \widetilde{\beta}) - s(O_i; t_0, \theta, \beta) - D(O_i; \widetilde{\beta} - \beta) \\ = \; & I(T_i > C_i) I(C_i < t_0) Z_i Z_i^T \Bigg(\frac{\mu\{X_i; \check{\beta}(t_0)\}[1 - \mu\{X_i; \check{\beta}(t_0)\}]}{1 - \mu\{X_i; \check{\beta}(C_i)\}} \{\widetilde{\beta}(t_0) - \beta(t_0)\} \\ & - \frac{\mu\{X_i; \check{\beta}(C_i)\}[1 - \mu\{X_i; \check{\beta}(t_0)\}]}{1 - \mu\{X_i; \check{\beta}(C_i)\}} \{\widetilde{\beta}(C_i) - \beta(C_i)\} \Bigg) - D(O_i; \widetilde{\beta} - \beta) \\ = \; & I(T_i > C_i) I(C_i < t_0) Z_i Z_i^T \\ & \times \Bigg(\Big[\frac{\mu\{X_i; \check{\beta}(t_0)\}[1 - \mu\{X_i; \check{\beta}(t_0)\}]}{1 - \mu\{X_i; \check{\beta}(C_i)\}} - \frac{F(t_0|X_i)\{1 - F(t_0|X_i)\}}{1 - F(C_i|X_i)} \Big] \{\widetilde{\beta}(t_0) - \beta(t_0)\} \\ & - \Big[\frac{\mu\{X_i; \check{\beta}(C_i)\}[1 - \mu\{X_i; \check{\beta}(t_0)\}]}{1 - \mu\{X_i; \check{\beta}(C_i)\}} - \frac{F(C_i|X_i)\{1 - F(t_0|X_i)\}}{1 - F(C_i|X_i)} \{\widetilde{\beta}(C_i) - \beta(C_i)\} \Big] \Bigg), \end{aligned}$$

where again $\check{\beta}(u)$ is on the line segment of $\widetilde{\beta}(u)$ and $\beta(u)$. It is now easy to see that

$$||s(O_i;t_0,\theta,\widetilde{\beta}) - s(O_i;t_0,\theta,\beta) - D(O_i;\widetilde{\beta}-\beta)|| \leq b(O_i)||\widetilde{\beta}-\beta||^2.$$

For (ii) in assumption 5.1, we need to show that the convergence rate of the IPW estimator $\widehat{\beta}$ is at least $n^{1/4}$. Let $\mathcal{F}$ denote all cadlag functions uniformly bounded on $[a,b]$. By adapting the proof in the previous item, we know that $\{\psi\{X_i,T_i;\beta(t)\} : t \in [a,b], \beta \in \mathcal{F}\}$ belongs to a Donsker class. Therefore $\sqrt{n}\{\widehat{\beta}(\cdot)-\beta(\cdot)\}$ converges weakly to a Gaussian process. Therefore this assumption is satisfied.

We now prove assumption 5.2 (stochastic equicontinuity). Note

$$\begin{aligned}
& \int D(o;\widetilde{\beta}-\beta)dGdH \\
= & \int_0^{t_0} g(u)\int h(x)\Big[\frac{F(t_0|x)\{1-F(t_0|x)\}zz^T\{\widetilde{\beta}(t_0)-\beta(t_0)\}}{1-F(u|x)}\{1-F(u|x)\} \\
& -\frac{F(u|x)\{1-F(t_0|x)\}zz^T\{\widetilde{\beta}(u)-\beta(u)\}}{1-F(u|x)}\{1-F(u|x)\}\Big]dxdu \\
= & \int_0^{t_0} g(u)\int h(x)zz^T\Big[F(t_0|x)\{1-F(t_0|x)\}\{\widetilde{\beta}(t_0)-\beta(t_0)\} \\
& -F(u|x)\{1-F(t_0|x)\}\{\widetilde{\beta}(u)-\beta(u)\}\Big]dxdu.
\end{aligned}$$

A sufficient condition for stochastic equicontinuity is provided in Chen et al. (2003), Remark 2. To be specific, we need to show for $\delta_n = o_p(1)$,

$$\sup_{||\widetilde{\beta}-\beta||\leq\delta_n} ||\frac{1}{n}\sum_{i=1}^n D(O_i,\widetilde{\beta}-\beta) - \int D(o,\widetilde{\beta}-\beta)dGdH|| = o_p(n^{-1/2}).$$

This can be proved by showing the process $\{D(O_i,\widetilde{\beta}-\beta) : t \in [a,b], \widetilde{\beta}-\beta \in \mathcal{F}\}$ belongs to a Donsker class. Note the form of $D(O_i,\widetilde{\beta}-\beta)$ in (A.5), again by adapting proof in item 4 this holds under the conditions A1-A5.

A sufficient condition for assumption 5.3 in Newey (1994) is

$$\sqrt{n}\int D(o;\widehat{\beta}-\beta)dGdH - \sum_{i=1}^n \alpha(O_i)/\sqrt{n} \to 0,$$

for some $\alpha(\cdot)$ (p.1366, Newey, 1994). Using the expansion (A.1) for $\widehat{\beta}(t)$, we obtain

$$\int D(o; \widehat{\beta} - \beta) dG dH \;=\; \frac{1}{n} \sum_{i=1}^{n} \xi(T_i; t_0, \theta, \beta) + o_p(n^{-1/2}),$$

where

$$\begin{aligned} \xi(T_i; t_0, \theta, \beta) \;=\; & \int_0^{t_0} g(u) \int h(x) z z^T \Big[F(t_0|x)\{1 - F(t_0|x)\}\psi(x, T_i; t_0, \theta) \\ & - F(u|x)\{1 - F(t_0|x)\}\psi\{x, T_i; u, \beta(u)\} \Big] dx du. \end{aligned}$$

Therefore assumption 5.3 holds.

For assumption 5.6, it is straightforward that (i) and (ii) are satisfied. We have

$$A \;=\; E\left\{ \frac{\partial s(O_i; t_0, \theta, \beta)}{\partial \theta} \right\} = E[\mu(X_i; \theta)\{1 - \mu(X_i; \theta)\} Z_i Z_i^T],$$

which is nonsingular under the assumption A5. It is easy to see that (iv) holds. For (v), since $\frac{\partial s(O_i; t_0, \theta, \beta)}{\partial \theta}$ is continuous in θ, assumption 5.4 (i) holds for $\frac{\partial s(O_i; t_0, \theta, \beta)}{\partial \theta}$. The assumption 5.4 (ii) holds for $\frac{\partial s(O_i; t_0, \theta, \beta)}{\partial \theta}$ since it does not depend on β.

By Lemma 5.3 of Newey (1994), we obtain

$$\sqrt{n}(\widehat{\theta}_n - \theta) \rightarrow N(0, A^{-1} V A^{-1}).$$

Appendix B

Proofs of Theorem 1 in Chapter 3

Let $\mathbf{P}_n$ denote the empirical measure and $\mathbf{P}$ be the probability measure. Let $\mathcal{H}_n$ be the reproducing kernel Hilbert space with the Gaussian kernel function with variance $1/\sigma_n^2$. Then the estimated decision function $\widehat{f}(\boldsymbol{x}; w)$ minimizes

$$l_n(f; w) + \lambda_n \|f\|_{\mathcal{H}_n}^2,$$

where $l_n(f; w) = \mathbf{P}_n \left[K_{h_n}(W - w)\phi(Df(\boldsymbol{X}))\right]$, $\phi(x) = (1 - x)_+$, and $\|\cdot\|_{\mathcal{H}_n}$ is the norm in $\mathcal{H}_n$. In our following derivations, we use c_d to denote any constant only depending on d.

First, we find a function f_{λ_n} in $\mathcal{H}_n$ which has the prediction error close to the true Bayes error. Let $\widetilde{l}(f; w) = E[\phi(Df(\boldsymbol{X}; w))|W = w]$. Note that since f_0 minimizes $E[\phi(Df(\boldsymbol{X}; w))|\boldsymbol{X} = \boldsymbol{x}, W = w]$, f_0 also minimizes $\widetilde{l}(f; w)$. Under assumption (C.2), we obtain from Theorem 2.7 in Steinwart and Scovel (2007) that there exists a constant c_d such that for any w

$$\inf_{f \in \mathcal{H}_n} \left(\widetilde{l}(f; w) + \lambda_n \|f\|_{\mathcal{H}_n} - \widetilde{l}(f_0; w)\right) \leq c_d \left(\sigma_n^d \lambda_n + C(2d)^{\alpha d/2} \sigma_n^{-\alpha d}\right)$$

Since $\sigma_n = \lambda_n^{-1/[(\alpha+1)d]}$ from condition (C.3), it gives

$$\inf_{f \in \mathcal{H}_n} \left(\widetilde{l}(f; w) + \lambda_n \|f\|_{\mathcal{H}_n} - \widetilde{l}(f_0; w)\right) \leq c_d \lambda_n^{\alpha/(\alpha+1)}.$$

Therefore, if let $f_{\lambda_n}(\cdot; w)$ be the unique function in $\mathcal{H}_n$ (the uniqueness is due to the strictly convexity) minimizing the left-hand side of the above inequality, then it holds that uniformly in $w \in \mathcal{W}$, where $\mathcal{W}$ is the support of W,

$$\widetilde{l}(f_{\lambda_n}; w) + \lambda_n \|f_{\lambda_n}\|_{\mathcal{H}_n} - \widetilde{l}(f_0; w) \leq c_d \lambda_n^{\alpha/(\alpha+1)}. \qquad (A.1)$$

From (A.1), we have

$$\|f_{\lambda_n}\|_{\mathcal{H}_n} \leq \widetilde{l}(f_0; w)/\lambda_n + c_d \lambda_n^{\alpha/(\alpha+1)}/\lambda_n.$$

Since $\widetilde{l}(f_0, w)$ is uniformly bounded, by redefining constant c_d, we obtain $\sup_{w \in \mathcal{W}} \|f_{\lambda_n}\|_{\mathcal{H}_n} \leq c_d/\lambda_n$. According to Lemma 3.1 in Steinwart et al. (2006), $\sup_{w,x} |f_{\lambda_n}| \leq c_d/\lambda_n$. Moreover, Steinwart and Scovel (2007) shows that for any $m > d/2$, $\|f_{\lambda_n}\|_{W^{2,m}} \leq \lambda_n^{-1} c_d \sigma_n^{2m-d}$ where $\| \cdot \|_{W^{p,k}}$ is the Sobolev norm. Thus, from the Sobolev embedding theorem, we conclude $\|f_{\lambda_n}\|_{W^{\infty, m-(d+1)/2}} \leq \lambda_n^{-1} c_d \sigma_n^{2m-d}$.

Second, to establish the risk bound for $\widehat{f}(\boldsymbol{x}; w)$, we need an upper bound for $\sup_w \|\widehat{f}(\boldsymbol{x}; w_0)\|_{\mathcal{H}_n}$. From the fact that

$$l_n(\widehat{f}; w) + \lambda_n \|\widehat{f}\|^2_{\mathcal{H}_n} \leq l_n(f_{\lambda_n}; w) + \lambda_n \|f_{\lambda_n}\|^2_{\mathcal{H}_n}, \qquad (A.2)$$

we have

$$\lambda_n \|\widehat{f}\|^2_{\mathcal{H}_n} \leq c_d/\lambda_n \mathbf{P}_n K_{h_n}(W - w) + \lambda_n \|f_{\lambda_n}\|^2_{\mathcal{H}_n}.$$

Thus, using the uniform consistency of the kernel density estimator, we have

$$\lambda_n \|\widehat{f}\|^2_{\mathcal{H}_n} \leq c_d \lambda_n^{-1} + \lambda_n \|f_{\lambda_n}\|^2_{\mathcal{H}_n}. \qquad (A.3)$$

This implies $\|\widehat{f}\|_{\mathcal{H}_n}$ is bounded by $O(\lambda_n^{-1})$ with probability one.

We consider probability sample in the event

$$\mathcal{A} = \left\{ \sup_{w \in \mathcal{W}} \|\widehat{f}\|_{\mathcal{H}_n} \leq c_d(\lambda_n^{-1} + t) \right\}.$$

From (A.2), we obtain

$$l(\widehat{f}; w) - l(f_{\lambda_n}; w) - \lambda_n \|f_{\lambda_n}\|^2_{\mathcal{H}_n} + \lambda_n \|\widehat{f}\|^2_{\mathcal{H}_n}$$

$$\leq (\mathbf{P}_n - \mathbf{P})\left[K_{h_n}(W-w)\phi(Df_{\lambda_n}(\boldsymbol{X};w))\right] - (\mathbf{P}_n - \mathbf{P})\left[K_{h_n}(W-w)\phi(D\widehat{f}(\boldsymbol{X};w))\right], \tag{A.4}$$

where $l(f;w) = \mathbf{P}[K_{h_n}(W-w)\phi(Df(\boldsymbol{X};w))]$. By the continuous differentiability of the conditional density of $(D,\boldsymbol{X})$ given W in condition (C.1), it is easy to show $\sup_{w\in\mathcal{W}}|l(f;w)-\widetilde{l}(f;w)|f_W(w) \leq c_d h_n^2/\lambda_n$ if $|f| \leq c_d/\lambda_n$. Furthermore, using (A.1) , the left-hand side of (A.4) is bounded from below by

$$c_d[\widetilde{l}(\widehat{f};w) - \widetilde{l}(f_0;w) - \tilde{c}_d\lambda_n^{\alpha/(\alpha+1)}] + \lambda_n\|\widehat{f}\|^2_{\mathcal{H}_n}.$$

On the other hand, if we define

$$\mathcal{F}_1 = \left\{K_{h_n}(W-w)\phi(Df(\boldsymbol{X})) : w\in\mathcal{W}, \|f\|_{W^{\infty,m-(d+1)/2}} \leq c_d(\lambda_n^{-1}+t)\sigma_n^{2m-d}\right\},$$

then following the embedding arguments from the reproducing kernel Hilbert space $\mathcal{H}_n$ to the Sobolev space, we conclude that both f_{λ_n} and $\widehat{f}$ belong to $\mathcal{F}_2$. Hence, we obtain from equation (A.4) that

$$\sup_{w\in\mathcal{W}}\left\{\widetilde{l}(\widehat{f};w) - \widetilde{l}(f_0;w) + \lambda_n c_d\|\widehat{f}\|^2_{\mathcal{H}_n}\right\} \leq c_d\left\{h_n^2\lambda_n^{-1} + \lambda_n^{\alpha/(\alpha+1)}\right\} + 2\|\mathbf{P}_n - \mathbf{P}\|_{\mathcal{F}_1}. \tag{A.5}$$

From Theorem 2.7.1 in Van der Vaart and Weller (1996), the bracket number of

$$\mathcal{F}_2 = \left\{f(\boldsymbol{X}) : \|f\|_{W^{\infty,m-(d+1)/2}} \leq \lambda_n^{-1}c_d\sigma_n^{2m-d}\right\}$$

satisfies $\log N(\epsilon, \mathcal{F}_2\lambda_n/\sigma_n^{2m-d}, \|\cdot\|_\infty) \leq c_d\epsilon^{-d/(m-(d+1)/2)}$. Since for any w_1, w_2 and f_1, f_2,

$$\left|K_{h_n}(W-w_1)\phi(Df_1(\boldsymbol{X})) - K_{h_n}(W-w_2)\phi(Df_2(\boldsymbol{X}))\right| \leq c_d h_n^{-2}/\lambda_n|w_1-w_2| + h_n^{-1}|f_1(\boldsymbol{X}) - f_2(\boldsymbol{X})|.$$

Therefore, the covering number for $\mathcal{F}_1$ satisfies

$$\log N(\epsilon, \mathcal{F}_1(\lambda_n h_n)^2/\sigma_n^{2m-d}, \|\cdot\|_\infty) \leq c_d\epsilon^{-d/(m-(d+1)/2)}\left\{1+\log\epsilon^{-1}\right\}.$$

From the large deviation results for the empirical process (Theorem 2.14.10, Van der Vaart and Weller, 1996), we obtain

$$P\left(\sqrt{n}\|\mathbf{P}_n - \mathbf{P}\|_{\mathcal{F}_1} > t\sigma_n^{2m-d}c_d(\lambda_n^{-1}+t)/(h_n)^2\right) \leq c_d e^{-t^2}$$

when $t > t_0(d)$ for some constant $t_0(d)$ if we choose m so that $m > d + 1/2$. We conclude

$$P\left(\|\mathbf{P}_n - \mathbf{P}\|_{\mathcal{F}_1} > r_n t\right) \leq c_d e^{-t}.$$

Hence,

$$P\left(\sup_{w \in \mathcal{W}} \left\{\widetilde{l}(\widehat{f}; w) - \widetilde{l}(f_0; w) + c_d \lambda_n \|\widehat{f}\|^2_{\mathcal{H}_n}\right\} > c_d \left\{h_n^2 \lambda_n^{-1} + \lambda_n^{\alpha/(\alpha+1)} + r_n t\right\}\right) \leq e^{-t},$$

where $r_n = n^{-1/2} \sigma_n^{2m-d} \lambda_n^{-1}/(h_n)^2$.

Finally, using the relationship between the hinge loss and the zero-one loss (cf. Bartlett et al., 2006), we conclude

$$P\left(\sup_{w \in \mathcal{W}} \left\{|Err(\widehat{f}; w) - Err(f_0; w)| + c_d \lambda_n \|\widehat{f}\|^2_{\mathcal{H}_n}\right\}\right.$$
$$\left. > c_d(h_n^2/\lambda_n + \lambda_n^{\alpha/(\alpha+1)}) + r_n t\right) \leq e^{-t}.$$

The theorem follows if we choose $m = d + 1$.

Remarks. Under the linear rules, we can follow exactly the same arguments as before but the Hilbert space $\mathcal{H}_n$ is replaced by the Euclidean space R^d. Thus, we can set $\sigma_n = 1$. Furthermore, we can use $f_0(\boldsymbol{x}; w)$ as $f_{\lambda_n}(\boldsymbol{x}; w)$ in the proof of Theorem 1. Then from (A.2), we obtain

$$\|\hat{\boldsymbol{\beta}}(w)\|^2 \leq c_d \lambda_n^{-1} + \|\boldsymbol{\beta}_0(w)\|^2.$$

Thus, using the large deviation result for the first term in the right-hand side, we obtain

$$P(\sup_w \|\hat{\boldsymbol{\beta}}(w)\|^2 > c_d(\lambda_n^{-1} + t)) < e^{-t^2}. \tag{A.6}$$

Thus, we restrict to the probability set of $\sup_w \|\hat{\boldsymbol{\beta}}\|^2 \leq c_d(\lambda_n^{-1} + t))$. Then using the inequality (A.4), we obtain

$$\sup_{w \in \mathcal{W}} \left\{l(\hat{f}; w) - l(f_0, w)\right\} \leq 2\|\mathbf{P}_n - \mathbf{P}\|_{\mathcal{F}_4},$$

where

$$\mathcal{F}_4 = \left\{ K_{h_n}(W-w)\phi(Df) : f = \boldsymbol{\beta}^T \boldsymbol{x}, \|\boldsymbol{\beta}\|^2 \leq c_d(\lambda_n^{-1}+t) \right\}.$$

From the large deviation result of empirical process, we know

$$P(\sqrt{n}\|\mathbf{P}_n - \mathbf{P}\|_{\mathcal{F}_4} > h_n^{-2}\lambda_n^{-1/2}c_d t) \leq e^{-t^2}. \qquad (A.7)$$

Therefore, using $l(f;w) = E[\phi(Df(\boldsymbol{X};w))|W = w] + O(h_n^2\lambda_n^{-1/2}\sqrt{1+t})$ and the relationship between the hinge loss and zero-one loss, it gives that with at least probability $1-e^{-t}$,

$$\sup_{w\in\mathcal{W}} |Err(\widehat{f};w) - Err(f_0;w)| \leq h_n^2\lambda_n^{-1/2}\sqrt{1+t} + n^{-1/2}h_n^{-2}\lambda_n^{-1/2}c_d t \leq n^{-1/4} + n^{-1/4}t.$$

That is, $\sup_{w\in\mathcal{W}} |Err(\widehat{f};w) - Err(f_0;w)| = O_p(n^{-1/4})$.