Ordinal Outcome Prediction and Treatment Selection in Personalized Medicine

Ordinal Outcome Prediction and Treatment Selection in Personalized Medicine

A dissertation presented

by

Yuanyuan Shen

to

The Department of Biostatistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biostatistics

Harvard University
Cambridge, Massachusetts

April 2015

Dissertation Advisor: Professor Tianxi Cai Yuanyuan Shen

Ordinal Outcome Prediction and Treatment Selection in Personalized Medicine

Abstract

In personalized medicine, two important tasks are predicting disease risk and selecting appropriate treatments for individuals based on their baseline information. The dissertation focuses on providing improved risk prediction for ordinal outcome data and proposing score-based test to identify informative markers for treatment selection. In Chapter 1, we take up the first problem and propose a disease risk prediction model for ordinal outcomes. Traditional ordinal outcome models leave out intermediate models which may lead to suboptimal prediction performance; they also don't allow for non-linear covariate effects. To overcome these, a continuation ratio kernel machine (CRKM) model is proposed both to let the data reveal the underlying model and to capture potential non-linearity effect among predictors, so that the prediction accuracy is maximized. In Chapter 2, we seek to develop a kernel machine (KM) score test that can efficiently identify markers that are predictive of treatment difference. This new approach overcomes the shortcomings of the standard Wald test, which is scale-dependent and only take into account linear effect among predictors. To do this, we propose a model-free score test statistics and implement the KM framework. Simulations and real data applications demonstrated the advantage of our methods over the Wald test. In Chapter 3, based on the procedure proposed in Chapter 2, we further add sparsity assumption on the predictors to take into account the real world problem of sparse signal. We incorporate the generalized higher criticism (GHC) to threshold the signals in a group and maintain a high detecting power. A comprehensive comparison of the procedures in Chapter 2 and Chapter 3

demonstrated the advantages and disadvantages of difference procedures under different scenarios.

Contents

Title page i

Abstract iii

Table of Contents v

Contents **v**

Acknowledgments vii

1 Sparse Kernel Machine Regression for Ordinal Outcomes **1**

1.1 Introduction 2

1.2 Continuation Ratio Kernel Machine Regression 4

1.2.1 Inference under the full model 5

1.2.2 Estimation Under the sCR_{KM} Assumption 8

1.2.3 Model Evaluation 10

1.2.4 Data Driven Rule for Kernel Selection 11

1.3 Numerical Studies 12

1.3.1 Simulation Study 12

1.3.2 Data Example: Genetic Risk Prediction of Shared Autoimmunity . . 17

1.4 Discussion 19

1.5 Appendix A: Algorithm details 20

1.6 Appendix B: Parameter tuning 21

1.7 Appendix C: Asymptotic Properties of $\widehat{h}^{(c)}(\cdot)$ 21

2 Identifying Predictive Markers for Personalized Treatment Selection **24**

2.1 Introduction 25

2.2 Treatment Selection Model 27
2.2.1 Score Statistic for Identifying Important Baseline Predictors for Treatment Selection 27
2.2.2 Approximating the Null Distribution by Resampling Procedure . . . 30
2.2.3 Additional Consideration: Scale Parameters and Kernel PCA 31
2.3 Numerical Studies 33
2.3.1 Simulation Study 33
2.3.2 Example: Predictors Useful for Individualized Treatment of HIV Infected Patients 37
2.3.3 Example: Predictors Useful for Treatment of Patients with Advanced Chronic Heart Failure 38
2.4 Discussion 39
2.5 Appendix: Convergence of the Proposed Test Statistic 40

3 Identifying Sparse Predictive Markers for Personalized Treatment Selection 42
3.1 Introduction 43
3.2 Identifying Informative Baseline Predictors for Treatment selection 46
3.2.1 Wald Test for Identifying Informative Baseline Predictors for Treatment Selection 46
3.2.2 Score Test for Identifying Important Baseline Predictors for Treatment Selection 47
3.2.3 Sparsity Assumption and Incorporating Generalized Higher Criticism 49
3.2.4 The Omnibus Test 51
3.3 Simulation Study 52
3.4 Data Example: Detecting Informative Baseline Predictors for HIV Treatment Selection 58
3.5 Discussion 60

References 62

Acknowledgments

I would like to thank my advisor, Tianxi Cai, for being an excellent mentor, teacher and friend. She has the greatest kindness, patience and understanding, and she has been giving me guidance well beyond academia, for my career and for my life. I feel extremely lucky to have her as my advisor, I am so grateful to her for all the help.

Thank you to my committee members Xihong Lin and Robert Gray for their support and guidance. I am grateful to them for making the time and proposing helpful suggestions in every meeting with me. I am also grateful for the kind encouragement.

I am grateful to have made some very close friends while at Harvard. All the great classmates from my cohort gave me the warmest treatment and made my transition to US life so much easier. I would like to thank Shelley Liu, Emma Schwager, Godwin Yung, and Heather Mattie. Also, I couldn't have done this without friends who share my emotions. I want to thank Cheng Peng and Yuan Qiao for sharing your life with me. And of course, a big thank you to my boyfriend Matey Neykov. Thank you for all the support, patience, and dealing with my emotional moments.

This work is completely dedicated to my parents, Hua Wang and Zaixian Shen. Thank you for your love and support. All I have done is to make you proud of me. Thank you for understanding when I am too busy to video chat and thank you for the great meal and entertainment every time I went back home.

Sparse Kernel Machine Regression for Ordinal Outcomes

Yuanyuan Shen, Katherine P. Liao and Tianxi Cai

Department of Biostatistics

Harvard School of Public Health

1.1 Introduction

Ordinal outcome data, such as pain scales, disease severity, and quality of life scales, arise frequently in medical research. To derive classification rules for an ordinal outcome y with a $p \times 1$ predictor vector $\mathbf{x}$, one may employ regression models relating $\mathbf{x}$ to y and classify future subjects into different categories based on their predicted $P(y = c \mid \mathbf{x})$. Naive analysis strategies, such as dichotomizing y into a binary variable and fitting multinomial regression models, are not efficient as they do not take into account the ordinal property of the outcome. Commonly used traditional methods for modeling ordinal response data include the cumulative proportional odds model, the forward and backward continuation ratio (CR) models and the corresponding proportional odds version of the CR (pCR) model (Ananth and Kleinbaum, 1997). The forward full CR (fCR) model assumes that

$$\text{logit}\, P(y = c \mid y \geq c, \mathbf{x}) = \boldsymbol{\gamma}_0^{(c)} + \mathbf{x}^{\mathsf{T}}\boldsymbol{\beta}^{(c)}, c = 1, \ldots, C-1 \tag{1.1}$$

where y is assumed to take C ordered categories, $\{1, ..., C\}$, $\boldsymbol{\gamma}_0^{(c)}$ and $\boldsymbol{\beta}^{(c)}$ are unknown regression parameters that are allowed to vary across continuation ratios. When the covariate effects $\boldsymbol{\beta}^{(c)}$ are assumed to be constant across c, (1.1) reduces to the pCR model

$$\text{logit}\, P(y = c \mid y \geq c, \mathbf{x}) = \boldsymbol{\gamma}_0^{(c)} + \mathbf{x}^{\mathsf{T}}\boldsymbol{\beta}, c = 1, \ldots, C-1 \tag{1.2}$$

When choosing between these two models, we come across the trade-off between the model complexity and the efficiency in estimating the model parameters. With the fCR model, we might suffer from loss of efficiency due to estimating too many parameters if the true model is a sub-model of (1.1), especially when the dimension of $\mathbf{x}$ is not small. On the other hand, the pCR model might lead to poor prediction performance if the true covariate effects do vary across continuation ratios. However, for many applications, it is reasonable to expect that a compromise between the fCR model and the pCR model might

be optimal. That is, $\boldsymbol{\beta}^{(c)} = \boldsymbol{\beta}^{(c+1)}$ for some c but not all and thus it is possible to improve the estimation by leveraging the sparsity on $\boldsymbol{\beta}^{(c)} - \boldsymbol{\beta}^{(c-1)}$. Regularization methods can be easily adapted to incorporate sparsity for CR models. Under the pCR model, Archer and Williams (2012) imposed L_1 penalty to incorporate the sparsity of the elements of $\boldsymbol{\beta}$. For the fCR model, to leverage the additional sparsity on $\boldsymbol{\beta}^{(c)} - \boldsymbol{\beta}^{(c-1)}$, one may impose a "fused lasso" type of penalty (Tibshirani et al., 2005), which penalizes the L_1-norm of both $\boldsymbol{\beta}^{(c)}$ and $\boldsymbol{\beta}^{(c)} - \boldsymbol{\beta}^{(c-1)}$.

In the presence of non-linear covariate effects, these existing methods based on linearity assumptions may lead to classification rules with unsatisfactory performance. On the other hand, fully non-parametric methods are often not feasible due to the curse of dimensionality. Alternatively, one may account for non-linearity by including interaction or non-linear basis functions. However, in practice, there is typically no prior information on which non-linear basis should be used and including a large number of non-informative basis could result in significant overfitting. In recent years, kernel machine (KM) regression has been advocated as a powerful tool to incorporate complex covariate effects (Bishop et al., 2006; Schölkopf and Smola, 2001). The KM regression methods flexibly account for linear/non-linear effects, without necessitating explicit specification of the non-linear basis. For ordinal outcomes, some KM based algorithms have also been proposed. For example, in Cardoso and Da Costa (2007), the problem of classifying ordered classes is reduced to two-class problems and mapped into support vector machines (SVMs) and neural networks. In Chu and Keerthi (2005), SVM is used to optimize multiple thresholds to define parallel discriminant hyperplanes for the ordinal scales. Kernel Discriminant Analysis was extended using a rank constraint to solve the ordinal regression problem in Sun et al. (2010). However, none of these existing methods provide a good solution to leverage the potential similarity between sequential logits or to select optimal kernels.

In this paper, we propose a sparse CR KM (sCR_{KM}) regression method for ordinal out-

comes where we use the KM framework to incorporate non-linear effects and impose sparsity on the differences in the covariate effects between sequential categories to control for overfitting. To improve estimation and computational efficiency, we propose the use of the kernel principal component analysis (PCA) (Mika et al., 1999; Schölkopf et al., 1998) to transform the dual representation of the optimization problem back to the primal form with basis functions estimated from the PCA and reduce the number of parameters by thresholding the estimated eigenvalues. One key challenge in KM regression is the selection of appropriate kernel functions. Here, we propose a data driven rule for selecting an optimal kernel by minimizing a cross-validated prediction error measure. Simulation results suggest that the proposed procedures work well with relatively little price paid for the additional kernel selection. The rest of the paper is organized as the following. We introduce the logistic CR KM model in section 1.2.1 and detail the estimation procedures under the sparsity assumption in section 1.2.2. We also describe the model evaluation criteria in section 1.2.3 and propose a data driven rule for selecting an optimal kernel in section 1.2.4. Simulation and real data analysis results are given in section 3.3 and 3.4.

1.2 Continuation Ratio Kernel Machine Regression

Suppose data for analysis consist of n independent and identically distributed random vectors, $\{(y_i, \mathbf{x}_i^\intercal)^\intercal, i = 1, \ldots, n\}$. The forward fCR KM ($\text{fCR}_{\text{KM}}$) model assumes that

$$P(y_i = c \mid y_i \geq c, \mathbf{x}_i) = g\left\{\gamma_0^{(c)} + h^{(c)}(\mathbf{x}_i)\right\}, \quad \text{for } c = 1, \ldots, C-1, \tag{1.3}$$

where $g(x) = e^x/(1+e^x)$, $h^{(c)}(\cdot)$ is an unknown centered smooth function that belongs to a Reproducible Kernel Hilbert Space (RKHS) $\mathcal{H}_k$, with the Hilbert space generated by a given positive definite kernel function $k(\cdot, \cdot; \rho)$, and ρ is some tuning parameter associated with the kernel function (Cristianini and Shawe-Taylor, 2000). The kernel function $k(\mathbf{x}_1, \mathbf{x}_2; \rho)$ measures the similarity between $\mathbf{x}_1$ and $\mathbf{x}_2$ and different choices of k

lead to different RKHS. Some of the popular kernel functions include the gaussian kernel $k(\mathbf{x}_1, \mathbf{x}_2; \rho) = \exp\{-\|\mathbf{x}_1 - \mathbf{x}_2\|_2^2/2\rho^2\}$, which can be used to capture complex smooth non-linear effects; the linear kernel $k(\mathbf{x}_1, \mathbf{x}_2; \rho) = \rho + \mathbf{x}_1^{\intercal}\mathbf{x}_2$, which corresponds to $h(\mathbf{x})$ being linear in $\mathbf{x}$; and the quadratic kernel $k(\mathbf{x}_1, \mathbf{x}_2; \rho) = (\mathbf{x}_1^{\intercal}\mathbf{x}_2 + \rho)^2$, which allows for 2-way interactive effects. Here, we use $\|\cdot\|_p$ to denote the L_p norm and $\|\cdot\|_F$ to denote Fubini's norm for matrices. From here onward, for notational ease, we suppress ρ from the kernel function k.

By Mercer's Theorem (Cristianini and Shawe-Taylor, 2000), any $h(\mathbf{x}) \in \mathcal{H}_k$ has a $primal\ representation$ with respect to the eigensystem of k. Specifically, under the probability measure of $\mathbf{x}$, k has eigenvalues $\{\lambda_l, l = 1, ..., \mathcal{J}\}$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{\mathcal{J}}$ and the corresponding eigenfunctions $\{\phi_l, l = 1, ..., \mathcal{J}\}$ such that $k(\mathbf{x}_1, \mathbf{x}_2) = \sum_{l=1}^{\mathcal{J}} \lambda_l \phi_l(\mathbf{x}_1)\phi_l(\mathbf{x}_2)$, where $\mathcal{J}$ could be infinity and $\lambda_l > 0$ for any $l < \infty$. The basis functions, $\{\psi_l(\mathbf{x}) = \sqrt{\lambda_l}\phi_l(\mathbf{x}), l = 1, ..., \mathcal{J}\}$, span the RKHS $\mathcal{H}_k$. Hence all $h^{(c)} \in \mathcal{H}_k$ has a primal representation, $h^{(c)}(\mathbf{x}) = \sum_{l=1}^{\mathcal{J}} \beta_l^{(c)}\psi_l(\mathbf{x})$, and (1.3) is equivalent to

$$P(y_i = c \mid y_i \geq c, \mathbf{x}_i) = g\left\{\gamma_0^{(c)} + \sum_{l=1}^{\mathcal{J}} \beta_l^{(c)}\psi_l(\mathbf{x}_i)\right\}, \quad \text{for } c = 1, \ldots, C-1. \tag{1.4}$$

Assuming $\beta_l^{(c)} = \beta_l^{(c')}$ for all l, c, c' leads to a pCR KM (pCR_{KM}) model. Throughout, we assume that $\mathbf{x}$ is bounded and k is smooth, leading to bounded $\{\psi_l(\mathbf{x})\}$ on the support of $\mathbf{x}$.

1.2.1 Inference under the full model

To make inference about the model parameters with observed data, we may maximize the following penalized likelihood with respect to $\{(\gamma_0^{(c)}, h^{(c)}), c = 1, ..., C-1\}$ with penalty

accounting for the smoothness of $h^{(c)}$:

$$\sum_{i=1}^{n}\sum_{c=1}^{C-1}\ell_c\left\{y_i,\gamma_0^{(c)},h^{(c)}(\mathbf{x}_i)\right\}-\tau_2\sum_{c=1}^{C-1}\|h^{(c)}\|_{\mathcal{H}_k}^2 \tag{1.5}$$

where $\tau_2 \geq 0$ is a tuning parameter controlling the amount of penalty,

$$\ell_c(y_i,\gamma_0^{(c)},h_i^{(c)})=I(y_i\geq c)\left[D_i^{(c)}\log\{g(\gamma_0^{(c)}+h_i^{(c)})\}+(1-D_i^{(c)})\log\{1-g(\gamma_0^{(c)}+h_i^{(c)})\}\right]$$

and $D_i^{(c)}=I(y_i=c)$. From the primal representation of $h^{(c)}$, maximizing (1.5) is equivalent to maximizing the following penalized likelihood with respect to $\{(\gamma_0^{(c)},\boldsymbol{\beta}^{(c)}),c=1,...,C-1\}$:

$$\sum_{c=1}^{C-1}\left[\sum_{i=1}^{n}\ell_c\left\{y_i,\gamma_0^{(c)},\psi_i^{\mathsf{T}}\boldsymbol{\beta}^{(c)}\right\}-\tau_2\|\boldsymbol{\beta}^{(c)}\|_2^2\right] \tag{1.6}$$

Thus, if the basis functions $\{\psi_l\}$ were known, we can directly estimate $h^{(c)}$ in the primal form. Unfortunately, in practice the true basis are typically unknown as they involve the unknown distribution of $\mathbf{x}$. On the other hand, by the representer theorem (Kimeldorf and Wahba, 1970), it is not difficult to show that the maximizer in (1.5) always takes the dual representation with $h^{(c)}(\mathbf{x}_i)=\mathbf{k}_i^{\mathsf{T}}\boldsymbol{\alpha}^{(c)}$, where $\mathbf{k}_i=[k(\mathbf{x}_i,\mathbf{x}_1),\ldots,k(\mathbf{x}_i,\mathbf{x}_n)]^{\mathsf{T}}$ and $\boldsymbol{\alpha}^{(c)}$ is an $n\times 1$ vector of unknown weights to be estimated as model parameters. This representation reduces (1.6) to an explicit optimization problem in the dual form:

$$\sum_{c=1}^{C-1}\left[\sum_{i=1}^{n}\ell_c\left\{y_i,\gamma_0^{(c)},\mathbf{k}_i^{\mathsf{T}}\boldsymbol{\alpha}^{(c)}\right\}-\tau_2\boldsymbol{\alpha}^{(c)\mathsf{T}}\mathbb{K}_n\boldsymbol{\alpha}^{(c)}\right] \tag{1.7}$$

where $\mathbb{K}_n=n^{-1}[k(\mathbf{x}_i,\mathbf{x}_j)]_{n\times n}$. Note that unlike the hinge loss in SVM, the logistic loss function is smooth, and consequently the resulting estimate of $\alpha^{(c)}$ based on (1.7) is not sparse.

Maximization of (1.7), however, could be both numerically and statistically unstable due to the large number of $(n+1)(C-1)$ parameters to be estimated, especially when the sample size n is not small. On the other hand, if the eigenvalues of k decay quickly, then we may reduce the complexity by approximating k by a truncated kernel $k^{(r)}(\mathbf{x}_1, \mathbf{x}_2) = \sum_{l=1}^{r} \lambda_l \phi_l(\mathbf{x}_1)\phi_l(\mathbf{x}_2)$, for some r such that $\sum_{l=r+1}^{\mathcal{J}} \lambda_l = o(\sum_{l=1}^{\mathcal{J}} \lambda_l)$. The error $\mathcal{E}_n = \|\mathbb{K}_n - \mathbb{K}_n^{(r)}\|$ can be bounded by $O\{\lambda_r + \sum_{l=r+1}^{\infty} \lambda_l\}$, where $\mathbb{K}_n^{(r)}$ is the kernel matrix constructed from kernel $k^{(r)}$ (Braun et al., 2005, Theorem 3.7). In many practical situations with fast decaying eigenvalues for k, r is typically fairly small and we can effectively approximate $\mathcal{H}_k$ by a finite dimensional space $\mathcal{H}_{k^{(r)}}$. Although $k^{(r)}$ is generally not attainable directly in practice, we may approximate the space spanned by $k^{(r)}$ through kernel PCA by applying a singular value decomposition to $\mathbb{K}_n$: $\mathbb{K}_n = \widetilde{\mathbf{\Phi}}\hat{\Lambda}\widetilde{\mathbf{\Phi}}^{\intercal}$, where $\widetilde{\mathbf{\Phi}} = (u_1, \ldots, u_n)$, $\hat{\Lambda} = \text{diag}\{a_1, \ldots, a_n\}$, $a_1 \geq \ldots \geq a_n \geq 0$ are the eigenvalues of $\mathbb{K}_n$ and $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ are the corresponding eigenvectors. By the relative-absolute bound for the estimated principal values in kernel PCA, the principal values a_l converge to the eigenvalues λ_l and the projection error $\{\sum_{l=r_n+1}^{n} a_l\}^2$ can be bounded by $O[\{\sum_{l=r_n+1}^{\mathcal{J}} \lambda_l\}^2 + \mathcal{E}_n]$. Thus, with properly chosen r_n and sufficiently fast decay rate for $\{\lambda_l\}$, $\mathbb{K}_n$ can be approximated well by $\widetilde{\mathbb{K}}_n^{(r_n)} = \widetilde{\mathbf{\Phi}}_{(r_n)}\widetilde{\Lambda}_{(r_n)}\widetilde{\mathbf{\Phi}}_{(r_n)}^{\intercal} = \widetilde{\mathbf{\Psi}}_{(r_n)}\widetilde{\mathbf{\Psi}}_{(r_n)}^{\intercal}$, where $\widetilde{\mathbf{\Phi}}_{(r_n)} = [\mathbf{u}_1, ..., \mathbf{u}_{r_n}]$, $\widetilde{\Lambda}_{(r_n)} = \text{diag}\{a_1, ..., a_{r_n}\}$, and $\widetilde{\mathbf{\Psi}}_{(r_n)} = \widetilde{\mathbf{\Phi}}_{(r_n)}\text{diag}\{a_1^{1/2}, ..., a_{r_n}^{1/2}\}$. Replacing $\mathbb{K}_n$ with $\widetilde{\mathbb{K}}_n^{(r_n)}$ and applying a variable transformation $\widetilde{\boldsymbol{\beta}}_{(r_n)}^{(c)} = \widetilde{\mathbf{\Psi}}_{(r_n)}^{\intercal}\boldsymbol{\alpha}^{(c)}$, the maximization of (1.7) can be approximately solved by maximizing

$$\widehat{\mathcal{L}}_0(\boldsymbol{\theta}; \tau_2) = \sum_{c=1}^{C-1}\left[\sum_{i=1}^{n} \ell_c\left\{y_i, \gamma_0^{(c)}, \widetilde{\boldsymbol{\psi}}_i^{\intercal}\boldsymbol{\beta}_{(r_n)}^{(c)}\right\} - \tau_2\|\boldsymbol{\beta}_{(r_n)}^{(c)}\|_2^2\right] \tag{1.8}$$

with respect to $\boldsymbol{\theta} = \{\gamma_0^{(c)}, \boldsymbol{\beta}_{(r_n)}^{(c)}, c = 1, ..., C-1\}$, where $\widetilde{\boldsymbol{\psi}}_i$ is the ith row of $\widetilde{\mathbf{\Psi}}_{(r_n)}$. In practice, we may choose $r_n = \min\{r : \sum_{l=1}^{r} a_l / \sum_{l=1}^{n} a_l \geq \eta\}$ for some η close to 1.

Let $\widetilde{\boldsymbol{\theta}} = \{\widetilde{\gamma}_0^{(c)}, \widetilde{\boldsymbol{\beta}}_{(r_n)}^{(c)}, c = 1, ..., C-1\}$ denote the estimator from the maximization of (1.8). Then for a future subject with $\mathbf{x}$, the probability $\pi_+(c \mid \mathbf{x}) = P(y = c \mid y \geq c, \mathbf{x})$ can be

estimated as

$$\widetilde{\pi}_+(c \mid \mathbf{x}) = g\left\{\widetilde{\gamma}_0^{(c)} + \widetilde{\mathbf{\Psi}}_{(r_n)}(\mathbf{x})^\intercal \widetilde{\boldsymbol{\beta}}_{(r_n)}^{(c)}\right\}$$

$\widetilde{\mathbf{\Psi}}_{(r_n)}(\mathbf{x}) = n^{-1}\text{diag}(a_1^{-1/2}, ..., a_{r_n}^{-1/2})\widetilde{\mathbf{\Phi}}_{(r_n)}^\intercal [k(\mathbf{x}, \mathbf{x}_1), ..., k(\mathbf{x}, \mathbf{x}_n)]^\intercal$, by the Nystrom method (Rasmussen, 2004). Subsequently, $\pi(c \mid \mathbf{x}) = P(y = c \mid \mathbf{x})$ can be estimated as

$$\widetilde{\pi}(c \mid \mathbf{x}) = \widetilde{\pi}_+(1 \mid \mathbf{x})^{I(c=1)} \left\{\widetilde{\pi}_+(c \mid \mathbf{x}) \prod_{c'=1}^{c-1} \{1 - \widetilde{\pi}_+(c' \mid \mathbf{x})\}\right\}^{I(c\geq 2)}$$

A future subject with $\mathbf{x}$ can then be classified as $\widetilde{y}(\mathbf{x}) = \text{argmax}_c\, \widetilde{\pi}(c \mid \mathbf{x})$.

1.2.2 Estimation Under the sCR$_{\text{KM}}$ Assumption

When the effects of $\mathbf{x}$ may differ across some continuation ratios but not all, one may improve the efficiency in estimating $h^{(c)}$ by imposing sparsity on $\{h^{(c+1)} - h^{(c)}, c = 1, ..., C-2\}$ in (1.3), or equivalently on $\{\boldsymbol{\beta}^{(c+1)} - \boldsymbol{\beta}^{(c)}, c = 1, ..., C-2\}$ in (1.4). To leverage the sparsity in estimation under the sCR$_{\text{KM}}$ assumption, we propose to modify (1.8) and instead maximize the penalized likelihood,

$$\widehat{\mathcal{L}}(\boldsymbol{\theta}; \tau_1, \tau_2) = \widehat{\mathcal{L}}_0(\boldsymbol{\theta}; \tau_2) - \tau_1 \sum_{c=1}^{C-2} \frac{\|\boldsymbol{\beta}_{(r_n)}^{(c+1)} - \boldsymbol{\beta}_{(r_n)}^{(c)}\|_2}{\|\widetilde{\boldsymbol{\beta}}_{(r_n)}^{(c+1)} - \widetilde{\boldsymbol{\beta}}_{(r_n)}^{(c)}\|_2} \tag{1.9}$$

where τ_1 is another tuning parameter controlling the amount of penalty for the differences between adjacent $h^{(c)}$'s. The adaptive penalty is imposed here to ensure the consistency in identifying the set of unique $h^{(c)}$'s.

To carry out the maximization of (1.9) in practice, we first obtain a quadratic expansion of the log-likelihood $\widehat{\mathcal{L}}_0(\boldsymbol{\theta}; \tau_2)$ around the initial estimator $\widetilde{\boldsymbol{\theta}}$:

$$n^{-1}\widehat{\mathcal{L}}_0(\boldsymbol{\theta}; \tau_2) \approx n^{-1}\widehat{\mathcal{L}}_0(\widetilde{\boldsymbol{\theta}}, \tau_2) - \frac{1}{2}(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}})^\intercal \widetilde{\mathbb{A}}(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}),$$

where $\widetilde{\mathbb{A}} = n^{-1}\partial\widehat{\mathcal{L}}_0(\boldsymbol{\theta}; \tau_2)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\intercal|_{\boldsymbol{\theta}=\widetilde{\boldsymbol{\theta}}}$. Subsequently, we approximate the maximizer of

(1.9) by

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left[\frac{1}{2}(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}})^{\intercal} \widetilde{\mathbb{A}} (\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}) + n^{-1} \tau_1 \sum_{c=1}^{C-2} \frac{\|\boldsymbol{\beta}_{(r_n)}^{(c+1)} - \boldsymbol{\beta}_{(r_n)}^{(c)}\|_2}{\|\widetilde{\boldsymbol{\beta}}_{(r_n)}^{(c+1)} - \widetilde{\boldsymbol{\beta}}_{(r_n)}^{(c)}\|_2} \right] \tag{1.10}$$

Such a quadratic approximation has been previously proposed to ease the computation for maximizing LASSO penalized likelihood functions and was shown to perform well in general (Wang and Leng, 2007). Followed by a sequence of variable transformations, we reformulate our optimization problem into a standard group lasso penalized maximization problem (Yuan and Lin, 2006; Wang and Leng, 2008). The detailed algorithm is in Web Appendix A. We tune the 3-tuple parameters (ρ, τ_1 and τ_2) by varying ρ within a range of values. For any given ρ, we first get $\tau_2(\rho)$ by GCV criterion in the ridge regression. Subsequently, we select ρ and τ_1 by optimizing the AIC. The detailed tuning procedure is in Web Appendix B.

With $\widehat{\boldsymbol{\theta}} = \{\widehat{\gamma}_0^{(c)}, \widehat{\boldsymbol{\beta}}_{(r_n)}^{(c)}, c = 1, ..., C-1\}$ obtained from (1.10), $h^{(c)}(\mathbf{x})$ can be estimated as

$$\widehat{h}^{(c)}(\mathbf{x}) = \widetilde{\boldsymbol{\Psi}}_{(r_n)}(\mathbf{x})^{\intercal} \widehat{\boldsymbol{\beta}}_{(r_n)}^{(c)}$$

We also obtain the corresponding $\widehat{\pi}_+(c \mid \mathbf{x})$ and $\widehat{\pi}(c \mid \mathbf{x})$ by replacing $\widetilde{\boldsymbol{\theta}}$ in $\widetilde{\pi}_+(c \mid \mathbf{x})$ and $\widetilde{\pi}(c \mid \mathbf{x})$ respectively. Then subjects with $\mathbf{x}$ can be classified as $\widehat{y}(\mathbf{x}) = \text{argmax}_c\, \widehat{\pi}(c \mid \mathbf{x})$. We expect that the proposed sparse estimator $\widehat{\boldsymbol{\theta}}$ and the resulting classification $\widehat{y}$ will outperform the corresponding estimators and classifications derived from the fCR_{KM} model based on $\widetilde{\boldsymbol{\theta}}$ and the reduced pCR_{KM} model when the underlying model has $h^{(c)} = h^{(c+1)}$ for some c but not all. When $\mathcal{J} = \infty$, the convergence rate of $\widehat{h}^{(c)}$ would depend on the decay rate of the eigenvalues $\{\lambda_l\}$. On the other hand, for many practical settings, $\mathcal{H}_k$ with the optimal ρ can be approximated well with a finite dimensional space $\mathcal{H}_{k_r}$ with a fixed r. In Web Appendix C, we show that when $\mathcal{H}_k$ is finite dimensional, $\|\widehat{h}^{(c)} - h^{(c)}\| = O_p(n^{-\frac{1}{2}})$. Furthermore, we establish the model selection consistency in the sense that $P(\widehat{h}^{(c)} = \widehat{h}^{(c+1)}) \to 1$ if $h^{(c)} = h^{(c+1)}$.

1.2.3 Model Evaluation

To evaluate the prediction performances of different methods for future observations $(y^0, \mathbf{x}^0)$, we consider three prediction error measures: the overall mis-classification error (OME) $P(\widehat{y}(\mathbf{x}^0) \neq y^0)$, the absolute prediction error (APE) $E|\widehat{y}(\mathbf{x}^0) - y^0|$ and the average size of prediction sets ($\mathscr{L}_{\text{PS}}$) to be defined below. The OME puts equal weights to any error as long as $\widehat{y}(\mathbf{x}^0) \neq y^0$ and APE weights the error by the absolute distance between $\widehat{y}(\mathbf{x}^0)$ and y^0. When comparing classification rules, one often sees a trade-off between these two error measures as they are capturing slightly different aspects of the prediction performance. In addition to these measures of accuracy based on $\widehat{y}(\mathbf{x}^0)$, we also examine the performance by taking the uncertainty in the classification into account. Specifically, for a given $\mathbf{x}^0$ with predicted probabilities $\{\widehat{\pi}(c \mid \mathbf{x}^0), c = 1, ..., C\}$, instead of classifying the subject according to $\widehat{y}(\mathbf{x}^0) = \text{argmax}_c\, \widehat{\pi}(c \mid \mathbf{x}^0)$, we construct a prediction set (PS) $\widehat{\mathscr{P}}_{\alpha_0}(\mathbf{x}^0) = \{c : \widehat{\pi}(c \mid \mathbf{x}^0) \geq \wp\}$, consisting of all categories whose predicted probabilities exceed $\wp$, where $\wp$ is chosen as the largest value such that these prediction sets of all samples achieve a desired coverage level $100(1 - \alpha_0)\%$, i.e. $P\{y^0 \in \widehat{\mathscr{P}}_{\alpha_0}(\mathbf{x}^0)\} \geq 1 - \alpha_0$(Faulkenberry, 1973; Jeske and Harvallie, 1988; Lawless and Fredette, 2005; Cai et al., 2008). The average size of the PS,

$$\mathscr{L}_{\text{PS}} = E\|\widehat{\mathscr{P}}_{\alpha_0}(\mathbf{x}^0)\| = \sum_{c=1}^{C} P\left\{c \in \widehat{\mathscr{P}}_{\alpha_0}(\mathbf{x}^0)\right\},$$

can be also used to quantify the prediction performance based on those estimated $\widehat{\pi}(c \mid \mathbf{x}^0)$ derived from each model. The $\wp$ and $\mathscr{L}_{\text{PS}}$ can be calculated from the constructed PS' for all the samples in the testing set. Using the same argument as given in Cai et al. (2008), we may show that $\mathscr{L}_{\text{PS}}$ is minimized by the true model and hence is a useful measure as a basis for model selection. The use of $\mathscr{L}_{\text{PS}}$ allows us to achieve two goals: (i) to obtain a set of potential classifications $\widehat{\mathscr{P}}_{\alpha_0}(\mathbf{x})$ rather than $\widehat{y}(\mathbf{x}^0)$ to account for the uncertainty in the classification; and (ii) to provide a more comprehensive evaluation for the prediction

performance of $\widehat{\pi}(c \mid \mathbf{x}^0)$ that accounts for the uncertainty.

1.2.4 Data Driven Rule for Kernel Selection

The choice of kernels is critical in the prediction performance, since different kernels have different features in terms of accounting for the non-linearity properties of the data. For example, the fCR$_{\text{KM}}$ with linear kernel is equivalent to the linear CR model in (1.1). The quadratic kernel is useful for capturing two-way interactions among predictors; while the gaussian kernel performs well in capturing smooth and complex effects. Unfortunately, in practice, with a given dataset, it is typically unclear which kernel would be the most appropriate. Here we propose to select an optimal kernel via K-fold cross-validation to minimize the $\mathscr{L}_{\text{PS}}$. To carry out the K-fold cross-validation for kernel selection, we randomly split the training data into K disjoint subset of about equal sizes and label them as $S_k, k = 1, \cdots, K$. For each k, we use all observations which are *not* in S_k to fit our proposed procedures with several candidate kernels and obtain the corresponding estimate $\widehat{\boldsymbol{\theta}}$. Then we use samples in S_k to calculate their predicted probabilities $\widehat{\pi}(c \mid \mathbf{x}), c = 1, \cdots, C$. After obtaining the predicted probabilities for all the samples from cross-validation, $\widehat{\mathscr{P}}_{\alpha_0}(\mathbf{x})$ and $\mathscr{L}_{\text{PS}}$ can be computed for each of the kernels. The kernel with the smallest $\mathscr{L}_{\text{PS}}$ will be selected as the optimal kernel and the corresponding estimate $\widehat{\boldsymbol{\theta}}$ would then be used for prediction in the validation set. In regards to the choice of K, it is imperative that the size of the training set is large enough to accurately estimate the sCR$_{\text{KM}}$ model parameters. We recommend $K = 10$ as previously suggested in Breiman and Spector (1992).

1.3 Numerical Studies

1.3.1 Simulation Study

We conducted extensive simulations to evaluate the finite sample performance of our proposed methods and compared with three existing methods: the "one-against-one" SVM method (Hsu and Lin, 2002), the L_1 penalized pCR method (Archer and Williams, 2012), denoted by pCR_{L_1} and the classification tree for ordinal outcomes (CART) (Galimberti et al., 2012). For the "one-against-one" SVM, $C(C-1)/2$ binary classifiers are trained by SVM, and the appropriate class is found by a voting scheme.

We simulated 5 category ordinal outcome Y with continuous covariates under the CR_{KM} model in (1.3). The 20×1 predictor vector $\mathbf{X}$ was generated from MVN(0, 3.6+6.4$\mathbb{I}_{20\times 20}$). We generate $Y \mid \mathbf{X}$ based on two types of $h^{(c)}(\mathbf{X})$: (i) linear in $\mathbf{X}$; and (ii) linear effects plus two-way interactions between X_j and X_{j+1}. The regression coefficients $\boldsymbol{\beta}^{(c)}, c = 1, \cdots, C-1$ were set to be between 0 and 0.4 and the intercept parameters $\{\gamma_0^{(c)}, c = 1, \cdots, C-1\}$ were selected such that there were approximately the same number of observations in each of the five classes. For each setting, we considered three types of $h^{(c)}(\mathbf{X})$: (a) $h^{(1)} \neq h^{(2)} \neq h^{(3)} \neq h^{(4)}$ representing a fCR_{KM} model; (b) $h^{(1)} = h^{(2)} \neq h^{(3)} = h^{(4)}$ representing a model between a fCR_{KM} model and a pCR_{KM} model; and (c) $h^{(1)} = h^{(2)} = h^{(3)} = h^{(4)}$ representing a pCR_{KM} model. For each simulated data, we let $n = 500$ in the training set to estimate all model parameters including kernel selection. Then to evaluate the performances of different procedures, we generate independent test sets of sample size 5000 to approximate the expected accuracy of the trained models.

For each scenario, we generate 50 datasets to compare the performance of SVM, pCR_{L_1}, CART, and the three models: sCR_{KM} , fCR_{KM} and pCR_{KM} under the different choices of kernels. Three kernels, including linear, quadratic and gaussian, are considered as candidates for kernel selection. Our recommended procedure would be sCR_{KM} with adaptive kernel selection, denoted by $\text{sCR}_{\text{KM}}^{\mathbb{S}}$. Parameter tuning is performed based on

the whole training set, and the selected parameters are fixed and used for the CV as well as building and evaluating the prediction model. We also compare the performance of our proposed procedures with data driven selection of the kernel versus those obtained under the true optimal kernel (linear for setting i and quadratic for setting ii) in each setting to examine the price paid for selecting the kernel.

In Table 2.1, we present results comparing different procedures when the data are generated from setting i with linear effects. Results for setting ii with interactive effects are given in Table 1.2. Table 1.3 shows the percentage of times different kernels being selected as the optimal kernel based on proposed data driven rule for kernel selection. Under both settings, applying our sCR_{KM} method always results in similar performance as the true model when the underlying model is either (a) the fCR_{KM} model; or (c) the pCR_{KM} model. This indicates that little penalty is paid for letting the data determine the underlying sparsity of $h^{(c+1)} - h^{(c)}$. When the underlying model is in between the two, sCR_{KM} performs the best. When the effects are linear, our proposed procedures with adaptive kernel selection perform similarly to those based on linear kernel. In the settings with interaction effects, $\text{sCR}^{\mathbb{S}}_{\text{KM}}$ outperforms sCR_{KM} with linear kernel by capturing non-linear effects. For example, when $h^{(1)} = h^{(2)} \neq h^{(3)} = h^{(4)}$, the average prediction set size $\mathscr{L}_{\text{PS}}$ was 2.9 for $\text{sCR}^{\mathbb{S}}_{\text{KM}}$ and, 4.49 for sCR_{KM} with linear kernel. In this setting, $\text{sCR}^{\mathbb{S}}_{\text{KM}}$ also outperforms both the fCR_{KM} and the pCR_{KM} models regardless how kernel was selected for these models. The prediction accuracy from $\text{sCR}^{\mathbb{S}}_{\text{KM}}$ was also similar to sCR_{KM} with quadratic kernel, which is the optimal kernel in this setting, indicating little loss of accuracy for the additional adaptive kernel selection. In general, the kernel selection procedure makes sensible choices of the kernels. When the underlying effects are linear, the linear kernel is selected 100% of the times; when the underlying model involves interactions, either quadratic or gaussian kernels are selected but not the linear kernel. This suggests that the use of cross-validation can overcome the overfitting issue. Under setting ii with inter-

Table 1.1: Average prediction performances with respect to average size of the prediction set ($\mathscr{L}_{\text{PS}}$), the overall mis-classification error (OME), and the absolute prediction error (APE), for setting i with linear effects.

	(a) $h^{(1)} \neq h^{(2)} \neq h^{(3)} \neq h^{(4)}$			
kernel choice		$\mathscr{L}_{\text{PS}}$	OME	APE
Linear	fCR_{KM}	1.82(0.04)	0.33(0.01)	0.44(0.01)
	sCR_{KM}	1.83(0.04)	0.33(0.01)	0.44(0.01)
	pCR_{KM}	2.29(0.03)	0.46(0.01)	0.53(0.01)
Data Driven	fCR_{KM}	1.82(0.04)	0.33(0.01)	0.44(0.01)
	$\text{sCR}^{\mathbb{S}}_{\text{KM}}$	1.83(0.04)	0.33(0.01)	0.44(0.01)
	pCR_{KM}	2.29(0.03)	0.46(0.01)	0.53(0.01)
-	SVM	2.00(0.04)	0.34(0.01)	0.46(0.01)
	CART	-	0.56(0.02)	0.80(0.02)
	pCR_{L_1}	2.33(0.04)	0.47(0.01)	0.55(0.01)
	(b) $h^{(1)} = h^{(2)} \neq h^{(3)} = h^{(4)}$			
Linear	fCR_{KM}	1.90(0.05)	0.36(0.01)	0.44(0.02)
	sCR_{KM}	1.85(0.04)	0.35(0.01)	0.42(0.02)
	pCR_{KM}	2.21(0.03)	0.44(0.01)	0.51(0.01)
Data Driven	fCR_{KM}	1.90(0.05)	0.36(0.01)	0.44(0.02)
	$\text{sCR}^{\mathbb{S}}_{\text{KM}}$	1.85(0.04)	0.35(0.01)	0.42(0.02)
	pCR_{KM}	2.21(0.03)	0.44(0.01)	0.51(0.01)
-	SVM	2.04(0.04)	0.38(0.01)	0.47(0.01)
	CART	-	0.55(0.02)	0.76(0.03)
	pCR_{L_1}	2.19(0.03)	0.44(0.01)	0.51(0.01)
	(c) $h^{(1)} = h^{(2)} = h^{(3)} = h^{(4)}$			
Linear	fCR_{KM}	1.93(0.03)	0.39(0.01)	0.43(0.01)
	sCR_{KM}	1.86(0.03)	0.37(0.01)	0.39(0.01)
	pCR_{KM}	1.85(0.03)	0.36(0.01)	0.39(0.01)
Data Driven	fCR_{KM}	1.93(0.03)	0.39(0.01)	0.43(0.01)
	$\text{sCR}^{\mathbb{S}}_{\text{KM}}$	1.86(0.03)	0.37(0.01)	0.39(0.01)
	pCR_{KM}	1.85(0.03)	0.36(0.01)	0.39(0.01)
-	SVM	2.03(0.04)	0.42(0.01)	0.46(0.01)
	CART	-	0.55(0.01)	0.68(0.02)
	pCR_{L_1}	1.82(0.03)	0.36(0.01)	0.38(0.01)

Table 1.2: Average prediction performances with respect to average size of the prediction set ($\mathscr{L}_{\text{PS}}$), the overall mis-classification error (OME), and the absolute prediction error (APE), for setting ii with interactive effects.

	(a) $h^{(1)} \neq h^{(2)} \neq h^{(3)} \neq h^{(4)}$			
kernel choice		$\mathscr{L}_{\text{PS}}$	OME	APE
Linear	fCR_{KM}	4.47(0.10)	0.75(0.03)	1.53(0.15)
	sCR_{KM}	4.45(0.11)	0.75(0.03)	1.55(0.16)
	pCR_{KM}	4.44(0.05)	0.76(0.02)	1.65(0.17)
Data Driven	fCR_{KM}	1.99(0.07)	0.36(0.02)	0.52(0.02)
	$\text{sCR}^{\mathbb{S}}_{\text{KM}}$	2.03(0.09)	0.37(0.02)	0.52(0.02)
	pCR_{KM}	2.70(0.14)	0.53(0.02)	0.67(0.05)
Quadratic	fCR_{KM}	1.93(0.05)	0.35(0.01)	0.51(0.02)
	sCR_{KM}	1.95(0.05)	0.35(0.01)	0.51(0.02)
	pCR_{KM}	2.84(0.06)	0.55(0.01)	0.72(0.02)
-	SVM	2.32(0.06)	0.44(0.01)	0.63(0.02)
	CART	-	0.66(0.03)	1.02(0.04)
	pCR_{L_1}	4.05(0.14)	0.79(0.02)	1.72(0.30)
	(b) $h^{(1)} = h^{(2)} \neq h^{(3)} = h^{(4)}$			
Linear	fCR_{KM}	4.50(0.10)	0.75(0.03)	1.42(0.16)
	sCR_{KM}	4.49(0.10)	0.75(0.03)	1.44(0.18)
	pCR_{KM}	4.46(0.04)	0.75(0.03)	1.53(0.19)
Data Driven	fCR_{KM}	2.19(0.07)	0.43(0.02)	0.55(0.02)
	$\text{sCR}^{\mathbb{S}}_{\text{KM}}$	2.09(0.10)	0.41(0.02)	0.52(0.02)
	pCR_{KM}	2.44(0.07)	0.50(0.01)	0.59(0.03)
Quadratic	fCR_{KM}	2.14(0.04)	0.42(0.01)	0.54(0.02)
	sCR_{KM}	2.00(0.05)	0.39(0.01)	0.50(0.02)
	pCR_{KM}	2.50(0.06)	0.51(0.01)	0.61(0.02)
-	SVM	2.48(0.05)	0.51(0.01)	0.68(0.02)
	CART	-	0.64(0.02)	0.92(0.04)
	pCR_{L_1}	4.06(0.18)	0.79(0.02)	1.54(0.34)
	(c) $h^{(1)} = h^{(2)} = h^{(3)} = h^{(4)}$			
Linear	fCR_{KM}	4.45(0.09)	0.77(0.02)	1.52(0.13)
	sCR_{KM}	4.44(0.09)	0.77(0.02)	1.55(0.15)
	pCR_{KM}	4.44(0.05)	0.77(0.02)	1.63(0.18)
Data Driven	fCR_{KM}	2.53(0.07)	0.49(0.01)	0.62(0.03)
	$\text{sCR}^{\mathbb{S}}_{\text{KM}}$	2.34(0.10)	0.46(0.02)	0.57(0.04)
	pCR_{KM}	2.05(0.08)	0.40(0.02)	0.46(0.02)
Quadratic	fCR_{KM}	2.48(0.05)	0.48(0.01)	0.60(0.02)
	sCR_{KM}	2.26(0.06)	0.45(0.01)	0.55(0.02)
	pCR_{KM}	2.03(0.08)	0.40(0.02)	0.46(0.03)
-	SVM	2.76(0.05)	0.56(0.01)	0.80(0.03)
	CART	-	0.64(0.04)	0.87(0.04)
	pCR_{L_1}	4.09(0.18)	0.79(0.02)	1.64(0.31)

Table 1.3: % of times different kernels being selected as the optimal kernel based on proposed data driven rule for kernel selection.

	setting i with linear effects			setting ii with interaction effects		
	Linear	Guassian	Quadratic	Linear	Guassian	Quadratic
$h^{(1)} \neq h^{(2)} \neq h^{(3)} \neq h^{(4)}$	100	0	0	0	54	46
$h^{(1)} = h^{(2)} \neq h^{(3)} = h^{(4)}$	100	0	0	0	50	50
$h^{(1)} = h^{(2)} = h^{(3)} = h^{(4)}$	100	0	0	0	40	60

active effects, the procedures with gaussian kernel appear to perform similarly to those with the quadratic kernel with respect to prediction accuracy. This is not surprising since the gaussian kernel is a universal kernel such that its corresponding RKHS is rich enough to approximate any target function (Steinwart, 2002) and hence could capture quadratic effects reasonably well.

Comparing to the three existing methods, our proposed procedures also show great advantage. Across all settings, our proposed $\text{sCR}^{\mathbb{S}}_{\text{KM}}$ method outperforms the SVM. For example, in setting ii, when $h^{(1)} \neq h^{(2)} \neq h^{(3)} \neq h^{(4)}$, SVM has an average $\mathscr{L}_{\text{PS}}$ of 2.32 vs 2.03 from $\text{sCR}^{\mathbb{S}}_{\text{KM}}$; when $h^{(1)} = h^{(2)} = h^{(3)} = h^{(4)}$, SVM has an average $\mathscr{L}_{\text{PS}}$ of 2.76 vs 2.34 from $\text{sCR}^{\mathbb{S}}_{\text{KM}}$. This is in part due to the fact that SVM doesn't consider the ordinal property of the outcome or the underlying sparsity of $h^{(c+1)} - h^{(c)}$. The pCR_{L_1} performs similarly to pCR_{KM} with linear kernel because it only considers linear effect of the predictors. However, when the true underlying effects are non-linear as in setting ii, the pCR_{L_1} performs poorly as expected. The CART method generally provides less accurate prediction compared to both SVM and sCR_{KM} in both linear and non-linear settings. These comparisons imply the precision gain of our method due to imposing sparsity on the differences in the covariate effects of the sequential categories and incorporating potential non-linear effects via kernel selection.

1.3.2 Data Example: Genetic Risk Prediction of Shared Autoimmunity

Autoimmune diseases (ADs), roughly defined as conditions where the immune system attacks self tissues and organs, affect 1 out of 31 individuals in the United States (Jacobson et al., 1997). Although ADs encompass a broad range of clinical manifestations, e.g. joint swelling, skin rash, and vasculitis. Recent studies have uncovered shared genetic risk factors across different ADs (Criswell et al., 2005). Epidemiologic studies corroborate with findings from genetic studies demonstrating that autoimmune diseases co-occur within individuals and families (Somers et al., 2006).

The presence of autoantibodies defines the majority of autoimmune diseases. Cyclic Citrullinated Peptide (CCP) antibodies are associated with rheumatoid arthritis (RA), and assays employing CCP to measure antibodies recognizing citrullinated antigens are used as a diagnostic test for RA (Lee and Schur, 2003). Among RA patients, different levels of CCP also indicate different subtypes of RA and are associated with different disease progressions (Kroot et al., 2000). Positive CCP indicates increased likelihood of erosive disease in RA, and high level of CCP may be useful to identify patients with aggressive disease. Given the shared genetic risk factors across autoimmune diseases, we would expect that subjects with one autoimmune disease, would be at higher risk for other autoimmune diseases. For example, CCP may be positive in patients with other autoimmune diseases such as systemic lupus erythematosus (SLE) (Harel and Shoenfeld, 2006). So patients with other autoimmune diseases or with genetic profiles that are indicative of elevated risk of other autoimmune diseases may have worse RA disease progression, which is partially reflected in the CCP levels.

To study the relationship between CCP levels and measurements of autoimmunity, we applied our methods to a dataset of 1265 rheumatoid arthritis (RA) patients of European decsent nested in an EMR cohort at Partner's Healthcare (Liao et al., 2010). In this RA cohort, all subjects were genotyped for 67 single nucleotide polymorphisms (SNPs) with published associations with RA, Systemic lupus erythematosus (SLE), and celiac disease,

and an aggregate genetic risk score (GRS) is calculated for each of the three diseases based on the number of SNPs for the particular diseases (Liao et al., 2013). These three GRSs represent genetic markers of autoimmunity. In addition to genetic information, billing codes of four ADs, RA, JRA and Psoriatic arthritis (PsA), as well as radiology findings of erosions are also available as predictors. For the CCP levels, we categorized into 4 ordinal categories with the total numbers of patients being 353, 266, 312 and 334, in categories 1, 2, 3 and 4, respectively. To construct and evaluate various prediction methods, we randomly split the data into two independent sets (evenly split within each category) with 633 subjects in the training set and 632 subjects in the validation set.

Applying our proposed procedure to the training data, our adaptive kernel selection rule selected the RBF kernel out of the linear, quadratic and RBF kernels, suggesting the presence of non-linear effects. Prediction models using SVM, pCR_{L_1} and CART were also developed for comparison, and the results are shown in Table 1.4. When applying the proposed prediction rules to the validation set, the $\text{sCR}_{\text{KM}}^{\mathbb{S}}$ method results in the smallest $\mathscr{L}_{\text{PS}}$ (2.58), comparing to the SVM model (3.44) and the pCR_{L_1} model (3.43). As expected, if we enforce the sCR_{KM} with quadratic kernel or linear kernel, the procedure yields a larger $\mathscr{L}_{\text{PS}}$ (3 and 2.9), highlighting the advantage of kernel selection. With respect to the OME and APE, the $\text{sCR}_{\text{KM}}^{\mathbb{S}}$ model also outperforms the three existing methods. The $\text{sCR}_{\text{KM}}^{\mathbb{S}}$ has OME of 0.59, versus 0.67 from SVM, 0.66 from pCR_{L_1} and 0.67 from CART; has APE of 1.02, versus 1.18 from SVM, 1.27 from pCR_{L_1} and 1.22 from CART. In order to evaluate whether the differences in the prediction is significant, we applied the bootstrap procedure to the validation data to estimate the standard errors for the estimated $\mathscr{L}_{\text{PS}}$, OME, and APE, as well as the differences between $\text{sCR}_{\text{KM}}^{\mathbb{S}}$ and other methods with respect to these prediction errors. As shown in Table 1.4, $\text{sCR}_{\text{KM}}^{\mathbb{S}}$ leads to significantly lower prediction errors with p-values < 0.001. Therefore, our $\text{sCR}_{\text{KM}}^{\mathbb{S}}$ method leads to a more accurate model for predicting anti-CCP levels compared with existing methods.

Table 1.4: Prediction performances with respect to average size of the prediction set ($\mathscr{L}_{\text{PS}}$), the overall mis-classification error (OME), and the absolute prediction error (APE); differences in performances between sCR_{KM} and existing methods with standard deviations (SD)

Methods for comparison	Prediction measurements			Differences between sCR_{KM} and other(SD)		
	$\mathscr{L}_{\text{PS}}$	OME	APE	$\mathscr{L}_{\text{PS}}$	OME	APE
sCR_{KM}	2.58(0.09)	0.59(0.02)	1.02(0.04)	-	-	-
SVM	3.5(0.05)	0.67(0.02)	1.18(0.04)	-0.92(0.09)	-0.09(0.02)	-0.16(0.04)
pCR_{L_1}	3.43(0.05)	0.66(0.02)	1.27(0.04)	-0.85(0.10)	-0.07(0.02)	-0.25(0.04)
CART	-	0.67(0.02)	1.22(0.04)	-	-0.08(0.02)	-0.20(0.04)

1.4 Discussion

In this paper, we proposed the sCR_{KM} procedure to construct optimal classification rules for ordinal outcomes. Our proposed method has advantage over existing methods by incorporating potentially non-linear effects while allowing for adaptive selection of optimal kernels. When there is sparsity for the differences in the covariate effects between two sequential categories, our method will also automatically assign the same coefficients to the sequential categories to achieve an optimal balance between model complexity and prediction accuracy. Our numerical studies suggest that when the underlying model is either the fCR_{KM} model or the reduced pCR_{KM} model, our proposed sCR_{KM} method performs similarly to fitting the corresponding model. In the case that the underlying model is in between the full and reduced model, the sCR_{KM} method performs better than fitting either the fCR_{KM} or the pCR_{KM} model. The proposed data driven rule for kernel selection also enables us to choose an optimal kernel for a given dataset. When we select how many components to use in the singular value decomposition of $\mathbb{K}_n$, we choose r_n as the largest r such that the estimated proportion of variation explained by the first r_n eigenfunctions, defined as $\frac{\{\sum_{l=1}^{r_n} a_l\}}{\{\sum_{l=1}^{n} a_l\}}$, is at least $\eta \in (0, 1]$. This selection rule is similar to those considered in the standard PCA literature (Park, 1981). Alternatively, one may treat r_n as an additional tuning parameter and select the appropriate r_n from a certain range to make sure the final

AIC reaches its optimal.

1.5 Appendix A: Algorithm details

To numerically obtain $\widehat{\boldsymbol{\theta}}$ in (9), we first perform a variable transformation by letting $\boldsymbol{\delta}^{(c)}$ to represent the differences between adjacent categories: $\boldsymbol{\delta}^{(c)} = \boldsymbol{\beta}_{(r_n)}^{(c+1)} - \boldsymbol{\beta}_{(r_n)}^{(c)}$, for $c = 1, \ldots, C-2$. Let $\boldsymbol{\Theta} = (\gamma_0^{(1)}, \cdots, \gamma_0^{(C-1)}, \boldsymbol{\beta}_{(r_n)}^{(1)}, \boldsymbol{\delta}^{(1)}, \ldots, \boldsymbol{\delta}^{(C-2)})$ be our new parameters after transformation, which relates to the original parameter vector $\boldsymbol{\theta} = (\gamma_0^{(1)}, ..., \gamma_0^{(C-1)}, \boldsymbol{\beta}_{(r_n)}^{(1)}, \boldsymbol{\beta}_{(r_n)}^{(2)}, ..., \boldsymbol{\beta}_{(r_n)}^{(C-2)}, \boldsymbol{\beta}_{(r_n)}^{(C-1)})^\intercal$ through $\boldsymbol{\theta} = \mathbb{M}\boldsymbol{\Theta}$, where

$$\mathbb{M} = \begin{bmatrix} \mathbb{I}_{C-1} & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & \mathbb{I}_r & 0 & 0 & 0 & \cdots & 0 \\ 0 & \mathbb{I}_r & \mathbb{I}_r & 0 & 0 & \cdots & 0 \\ 0 & \mathbb{I}_r & \mathbb{I}_r & \mathbb{I}_r & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \mathbb{I}_r & \mathbb{I}_r & \mathbb{I}_r & \mathbb{I}_r & \cdots & \mathbb{I}_r \end{bmatrix}$$

Let $\widetilde{\mathbf{X}} = \widetilde{A}^\intercal \mathbb{M}$ and $\widetilde{\mathbf{Y}} = \widetilde{A}^\intercal \widetilde{\boldsymbol{\theta}}$, where $\widetilde{\mathbb{A}} = \widetilde{A}\widetilde{A}^\intercal$. Therefore, (9) is transformed into a linear adaptive group LASSO (gLASSO) problem:

$$\widehat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta}}{\operatorname{argmin}} \left[\frac{1}{2} \|\widetilde{\mathbf{X}}\boldsymbol{\Theta} - \widetilde{\mathbf{Y}}\|_2^2 + \tau_1 \sum_{c=1}^{C-2} \frac{\|\boldsymbol{\Theta}^{(c)}\|_2}{\|\widetilde{\boldsymbol{\Theta}}^{(c)}\|_2} \right] \tag{1.11}$$

where $\widetilde{\boldsymbol{\Theta}} = \mathbb{M}^{-1}\widetilde{\boldsymbol{\theta}}$.

1.6 Appendix B: Parameter tuning

There are three tuning parameters involved in our proposed procedure, ρ, τ_1 and τ_2, where ρ is the parameter for kernel $k(\cdot,\cdot;\rho)$, τ_2 is the tuning parameter for the ridge penalty, and τ_1 is the gLASSO penalty parameter controlling the amount of penalty for the differences between adjacent categories. Commonly used methods for selecting tuning parameters for ridge regression and gLASSO penalties include AIC, BIC, cross-validation, and generalized cross validation (GCV) (Golub et al., 1979; Hastie et al., 2005; Yuan and Lin, 2006; Wang and Leng, 2008). For each given ρ, we obtain an optimal τ_2 based on the GCV criterion (Golub et al., 1979), denoted by $\tau_2(\rho)$. Then with each given ρ and $\tau_2(\rho)$, we obtain the corresponding synthetic data $\{\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}}, \widetilde{\delta}^{(c)}\}$ for fitting the gLASSO penalized least square in (1.11). The tuning parameters τ_1 and ρ are then selected via the AIC criterion. The degree of freedom in the AIC criterion is estimated analogous to those proposed in Yuan and Lin (2006) and Wang and Leng (2008). Specifically, we define $\text{DF}(\rho,\tau_1) = \sum_{c=1}^{C-1} I\{\|\widehat{\delta}^{(c)}(\rho,\tau_1)\| > 0\} + \sum_{c=1}^{C-1} \frac{\|\widehat{\delta}^{(c)}(\rho,\tau_1)\|_2}{\|\widetilde{\delta}^{(c)}(\rho)\|_2}(d_c(\rho)-1)$, where $d_c(\rho)$ is the effective number of parameters in the c^{th} group from the ridge regression, calculated as the sum of the diagonal elements of Hessian matrix $\widetilde{\mathbb{A}}(\rho)$ that correspond to the c^{th} group. We then select the optimal (ρ,τ_1) as the minimizer of $\text{AIC}(\rho,\tau_1) = -2\text{loglik}(\rho,\tau_1) + 2\text{DF}(\rho,\tau_1)$.

1.7 Appendix C: Asymptotic Properties of $\widehat{h}^{(c)}(\cdot)$

Here, when $\mathcal{H}_k$ is finite dimensional, we aim to establish the root-n convergence rate of $\widehat{h}^{(c)}(\mathbf{x})$ and model selection consistency in the sense that $P\{\widehat{h}^{(c)}(\mathbf{x}) = \widehat{h}^{(c+1)}(\mathbf{x})\} \to 1$ when $h^{(c)} = h^{(c+1)}$. To this end, we first note that we can write our penalized likelihood (7) in the same form as in Minnier (2012). It is the summation of $C-1$ independent terms, each of which takes the form:

$$\sum_{i=1}^{n} I(y_i \geq c)\left[D_i^{(c)} \log\{g(\gamma_0^{(c)} + \widetilde{\boldsymbol{\psi}}_i^{\mathsf{T}}\boldsymbol{\beta}_{(r_n)}^{(c)})\} + (1-D_i^{(c)})\log\{1-g(\gamma_0^{(c)} + \widetilde{\boldsymbol{\psi}}_i^{\mathsf{T}}\boldsymbol{\beta}_{(r_n)}^{(c)})\}\right] - \tau_2\|\boldsymbol{\beta}_{(r_n)}^{(c)}\|_2^2$$

Therefore, using the same arguments as given in Minnier (2012), we have

Lemma 1. *$P(r_n = r) \to 1$ and $\|\widetilde{\boldsymbol{\Psi}}(\mathbf{x}) - \boldsymbol{\Psi}(\mathbf{x})\|_2 + n^{-\frac{1}{2}}\|\widetilde{\boldsymbol{\Psi}} - \boldsymbol{\Psi}\|_F + \|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2 = O_p(n^{-\frac{1}{2}})$.*

It also directly implies that $\widetilde{h}^{(c)}(\mathbf{x}) - h^{(c)}(\mathbf{x}) = O_p(n^{-\frac{1}{2}})$ and we may need establish the convergences conditioning on $r_n = r$. In view of this together with the parametrization in (1.11), it suffices to show that $\widehat{\boldsymbol{\delta}}^{(c)} - \boldsymbol{\delta}^{(c)} = O_p(n^{-\frac{1}{2}})$, if $c \in \mathcal{A}$; and $P(\widehat{\boldsymbol{\delta}}^{(c)} = 0) \to 1$, if $c \notin \mathcal{A}$, where $\mathcal{A} = \{c : \boldsymbol{\delta}^{(c)} \neq 0\}$. These are parallel to Theorem 1 and 2 in (Wang and Leng, 2008) where they show the estimation consistency and selection consistency of the adaptive group lasso estimator. The main difference between our problem and the setting considered in Wang and Leng (2008) is that our $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Y}}$ are not directly observed data but are estimated quantities with $\widetilde{\mathbf{X}} = \widetilde{A}^{\intercal}\mathbb{M}$, $\widetilde{\mathbf{Y}} = \widetilde{A}^{\intercal}\widetilde{\boldsymbol{\theta}}$, where $\widetilde{\mathbb{A}} = \widetilde{A}\widetilde{A}^{\intercal}$, so we need to take into account the randomness in $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Y}}$. In their proof, the main arguments rely on two convergences: $n^{-1}\mathbf{X}^{\intercal}\mathbf{X} \to E(\mathbf{X}_i\mathbf{X}_i^{\intercal})$ in probability and $n^{-\frac{1}{2}}\mathbf{X}^{\intercal}(\mathbf{Y} - \mathbf{X}\boldsymbol{\delta}) = O_p(1)$. In our case, the corresponding convergences we need to establish are the probability convergence of $\mathbb{M}^{\intercal}\widetilde{\mathbb{A}}\mathbb{M}$ and $\mathbb{M}^{\intercal}\widetilde{\mathbb{A}}[n^{\frac{1}{2}}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta})] = O_p(1)$. By the Lemma, $n^{\frac{1}{2}}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) = O_P(1)$, and since $\mathbb{M}$ is a constant, it is suffice to show that $\widetilde{\mathbb{A}} = \text{diag}\{\widetilde{\mathbb{A}}^{(1)}, ..., \widetilde{\mathbb{A}}^{(C-1)}\}$ converges to $\mathbb{A} = \text{diag}\{\mathbb{A}^{(1)}, ..., \mathbb{A}^{(C-1)}\}$ in probability, where

$$\begin{aligned}
\widetilde{\mathbb{A}}^{(c)} =& n^{-1}\sum_{i=1}^{n} I(y_i \geq c)\left[\underline{\widetilde{\boldsymbol{\psi}}}_i\underline{\widetilde{\boldsymbol{\psi}}}_i^{\intercal} g(\underline{\widetilde{\boldsymbol{\psi}}}_i^{\intercal}\underline{\widetilde{\boldsymbol{\beta}}}_{(r)}^{(c)})(1 - g(\underline{\widetilde{\boldsymbol{\psi}}}_i^{\intercal}\underline{\widetilde{\boldsymbol{\beta}}}_{(r)}^{(c)}))\right] \\
\mathbb{A}^{(c)} =& E\left\{I(y_i \geq c)\left[\underline{\boldsymbol{\psi}}_i\underline{\boldsymbol{\psi}}_i^{\intercal} g(\underline{\boldsymbol{\psi}}_i^{\intercal}\underline{\boldsymbol{\beta}}_{(r)}^{(c)})(1 - g(\underline{\boldsymbol{\psi}}_i^{\intercal}\underline{\boldsymbol{\beta}}_{(r)}^{(c)}))\right]\right\}
\end{aligned}$$

$\underline{\boldsymbol{\psi}}_i = [1, \psi_i^{\intercal}]^{\intercal}$, $\underline{\boldsymbol{\beta}}_{(r)}^{(c)} = [\boldsymbol{\gamma}_0^{(c)}, \boldsymbol{\beta}_{(r)}^{(c)^{\intercal}}]^{\intercal}$; $\underline{\widetilde{\boldsymbol{\psi}}}_i = [1, \widetilde{\psi}_i^{\intercal}]^{\intercal}$, and $\underline{\widetilde{\boldsymbol{\beta}}}_{(r)}^{(c)} = [\widetilde{\boldsymbol{\gamma}}_0^{(c)}, \boldsymbol{\beta}_{(r)}^{(c)^{\intercal}}]^{\intercal}$.
Since $\|\widetilde{\mathbb{A}} - \mathbb{A}\|_F^2 = \sum_{c=1}^{C-1}\|\widetilde{\mathbb{A}}^{(c)} - \mathbb{A}^{(c)}\|_F^2$, so if we can show the convergence of each of the $C-1$ blocks, we will have convergence for the entire matrix $\widetilde{\mathbb{A}}$. Let $\widetilde{\mathbb{A}}^{\star(c)} = n^{-1}\sum_{i=1}^{n} I(y_i \geq c)\left[\underline{\boldsymbol{\psi}}_i\underline{\boldsymbol{\psi}}_i^{\intercal} g(\underline{\boldsymbol{\psi}}_i^{\intercal}\underline{\boldsymbol{\beta}}_{(r)}^{(c)})(1 - g(\underline{\boldsymbol{\psi}}_i^{\intercal}\underline{\boldsymbol{\beta}}_{(r)}^{(c)}))\right]$, we have $\|\widetilde{\mathbb{A}}^{(c)} - \mathbb{A}^{(c)}\|_F^2 = \|\widetilde{\mathbb{A}}^{(c)} - \widetilde{\mathbb{A}}^{\star(c)} + \widetilde{\mathbb{A}}^{\star(c)} - \mathbb{A}^{(c)}\|_F^2 \leq \|\widetilde{\mathbb{A}}^{(c)} - \widetilde{\mathbb{A}}^{\star(c)}\|_F^2 + \|\widetilde{\mathbb{A}}^{\star(c)} - \mathbb{A}^{(c)}\|_F^2$. Note that since $\widetilde{\mathbb{A}}^{\star(c)} \to \mathbb{A}^{(c)}$ with probability 1 by Law of Large Numbers, so we only need to show $\|\widetilde{\mathbb{A}}^{(c)} - \mathbb{A}^{\star(c)}\|_F^2 \to 0$. To simplify notation, we drop (c) superscripts and the (r) subscripts. We first split $\widetilde{\mathbb{A}} - \widetilde{\mathbb{A}}^{\star}$ into summation of three

parts:

$$\widetilde{\mathbb{A}} - \widetilde{\mathbb{A}}^{\star} = n^{-1}\sum_{i=1}^{n} I(y_i \geq c)\left[(\underline{\widetilde{\boldsymbol{\psi}}}_i - \underline{\boldsymbol{\psi}}_i)\underline{\widetilde{\boldsymbol{\psi}}}_i^{\mathsf{T}} g(\underline{\widetilde{\boldsymbol{\psi}}}_i^{\mathsf{T}}\underline{\widetilde{\boldsymbol{\beta}}})(1 - g(\underline{\widetilde{\boldsymbol{\psi}}}_i^{\mathsf{T}}\underline{\widetilde{\boldsymbol{\beta}}}))\right] \quad \text{(P1)}$$

$$+ n^{-1}\sum_{i=1}^{n} I(y_i \geq c)\left[\underline{\boldsymbol{\psi}}_i(\underline{\widetilde{\boldsymbol{\psi}}}_i - \underline{\boldsymbol{\psi}}_i)^{\mathsf{T}} g(\underline{\widetilde{\boldsymbol{\psi}}}_i^{\mathsf{T}}\underline{\widetilde{\boldsymbol{\beta}}})(1 - g(\underline{\widetilde{\boldsymbol{\psi}}}_i^{\mathsf{T}}\underline{\widetilde{\boldsymbol{\beta}}}))\right] \quad \text{(P2)}$$

$$+ n^{-1}\sum_{i=1}^{n} I(y_i \geq c)\left[\underline{\boldsymbol{\psi}}_i\underline{\boldsymbol{\psi}}_i^{\mathsf{T}}(g(\underline{\widetilde{\boldsymbol{\psi}}}_i^{\mathsf{T}}\underline{\widetilde{\boldsymbol{\beta}}})(1 - g(\underline{\widetilde{\boldsymbol{\psi}}}_i^{\mathsf{T}}\underline{\widetilde{\boldsymbol{\beta}}})) - g(\underline{\boldsymbol{\psi}}_i^{\mathsf{T}}\underline{\boldsymbol{\beta}})(1 - g(\underline{\boldsymbol{\psi}}_i^{\mathsf{T}}\underline{\boldsymbol{\beta}})))\right] \quad \text{(P3)}$$

Assume that $\|\underline{\boldsymbol{\psi}}_i\|_2 \leq R$ and apply the Lemma, (P1) can be bounded since

$$\left\| n^{-1}\sum_{i=1}^{n} I(y_i \geq c)\left[(\underline{\widetilde{\boldsymbol{\psi}}}_i - \underline{\boldsymbol{\psi}}_i)\underline{\widetilde{\boldsymbol{\psi}}}_i^{\mathsf{T}} g(\underline{\widetilde{\boldsymbol{\psi}}}_i^{\mathsf{T}}\underline{\widetilde{\boldsymbol{\beta}}})(1 - g(\underline{\widetilde{\boldsymbol{\psi}}}_i^{\mathsf{T}}\underline{\widetilde{\boldsymbol{\beta}}}))\right]\right\|_F$$
$$\leq n^{-1}\left[Rn^{\frac{1}{2}}\|\underline{\widetilde{\boldsymbol{\Psi}}} - \underline{\boldsymbol{\Psi}}\|_F + \|\underline{\widetilde{\boldsymbol{\Psi}}} - \underline{\boldsymbol{\Psi}}\|_F^2\right] = R \cdot O_p(n^{-1}) + O_p(n^{-2})$$

The term (P2) can also be bounded similarly with

$$n^{-1}\left\|\sum_{i=1}^{n} I(y_i \geq c)\left[\underline{\boldsymbol{\psi}}_i(\underline{\widetilde{\boldsymbol{\psi}}}_i - \underline{\boldsymbol{\psi}}_i)^{\mathsf{T}} g(\underline{\widetilde{\boldsymbol{\psi}}}_i^{\mathsf{T}}\underline{\widetilde{\boldsymbol{\beta}}})(1 - g(\underline{\widetilde{\boldsymbol{\psi}}}_i^{\mathsf{T}}\underline{\widetilde{\boldsymbol{\beta}}}))\right]\right\|_F \leq n^{-1/2}R\|\underline{\widetilde{\boldsymbol{\Psi}}} - \underline{\boldsymbol{\Psi}}\|_F = R \cdot O_p(n^{-1})$$

Since $\|\underline{\widetilde{\boldsymbol{\beta}}} - \underline{\boldsymbol{\beta}}\|_2 = O_p(n^{-1/2})$, $\|\widetilde{\boldsymbol{\Psi}} - \boldsymbol{\Psi}\|_F = O_p(1)$, $\|\underline{\boldsymbol{\psi}}_i\|_2 \leq R$ and $\|\underline{\boldsymbol{\beta}}\|_2 < \infty$, we can easily obtain $(P3) = O_p(n^{-\frac{1}{2}})$. Therefore $\|\widetilde{\mathbb{A}} - \widetilde{\mathbb{A}}^*\|_F \to 0$ in probability and hence $\|\widetilde{\mathbb{A}} - \mathbb{A}\|_F \to 0$ in probability. This, together with the same arguments as given in (Wang and Leng, 2008), implies that $\|\widehat{\boldsymbol{\delta}}^{(c)} - \boldsymbol{\delta}^{(c)}\|_2 = O_p(n^{-\frac{1}{2}})$ when $c \in \mathcal{A}$ and $P(\widehat{\boldsymbol{\delta}}^{(c)} = 0) \to 1$ when $c \notin \mathcal{A}$.

Identifying Predictive Markers for Personalized Treatment Selection

Yuanyuan Shen and Tianxi Cai

Department of Biostatistics

Harvard School of Public Health

2.1 Introduction

The effectiveness and potential risk of a treatment often varies by patient subgroups (Duffy and Crown, 2008; La Thangue and Kerr, 2011). For instance, ER negative breast cancer patients benefit substantially from chemotherapy while ER positive patients do not benefit as compared to receiving tamoxifen alone (IBCSG, 2002). A gene-expression profile appears to highly predictive of whether chemotherapy is beneficial for treating breast cancer patients and is now being further investigated by the TAILORx study (Sparano, 2006; Zujewski and Kamin, 2008). The adverse risk of Abacavir for treating HIV infected patients is strongly associated with the presence of the HLA-B*5701 allele and thus Abacavir was recommended only for patients not carrying this allele (Mallal et al., 2008). Recently, the US Preventive Services Task Force issued new guidelines recommending against routine mammography screening for women under 50 (Nelson et al., 2009). On the other hand, such guidelines may not be appropriate for populations at increased risk and refinement of such recommendations warrants further research.

Many factors including genetics predisposition and environmental influences may play a role in a patient's treatment response. Incorporating information on clinical, biological and genomic markers into personalized prediction of treatment response holds great potential for identifying subgroups of patients who are most likely to benefit or are at high risk for toxicity from a particular therapy. Interventions can then be targeted to well-defined groups that are likely to benefit and at low risk of adverse event.

In recent years, a wide range of statistical methods have been proposed for developing individualized treatment rules (ITRs) based on a set of baseline predictors (Qian and Murphy, 2011; Cai et al., 2011a; Foster et al., 2011; Zhao et al., 2012; Zhang et al., 2012; Zhao et al., 2013). When the number of candidate predictors for deriving ITRs is not small, it is important to only include informative markers since including a large number of unrelated markers may tamper the accuracy of the resulting ITR and lead to unnecessary cost associated with measuring the markers. Variable selection procedures have also been developed for both prediction and decision making (Gunter et al., 2011; Lu et al., 2013; Imai et al., 2013). However, when a large number of candidate markers are available, variable

selection procedures may not work well in identifying informative markers since many of such procedures are not consistent in variable selection and it is generally difficult to identify an appropriate tuning parameter to ensure selection consistency. For such settings, it would be desirable to perform testing on candidate markers and only develop ITRs using markers that are deemed predictive of treatment response.

Standard testing procedures for ITRs consider models that include interactions between the treatment group and the variables of interest and perform a Wald-type test on the interaction term. Rosenblum and van der Laan (2009) showed that even when the model is misspecified, the Wald test still obtains the correct size, if we use the sandwich estimator for the standard error. Despite the robustness property, such an approach suffers from two major limitations. First, the interaction term may not entirely capture markers' ability in predicting subject specific treatment effect (TE). When TE of interest is the treatment difference and the outcome Y is binary, the conditional TE given baseline predictor $\mathbf{X}$, $P(Y = 1 \mid T = 1, \mathbf{X}) - P(Y = 1 \mid T = 0, \mathbf{X})$, may depend on both the main effect and the interaction. For example, when $P(Y = 1 \mid T, \mathbf{X}) = g(\alpha + \beta T + \boldsymbol{\gamma}_0^{\intercal}\mathbf{X} + T\boldsymbol{\gamma}_1^{\intercal}\mathbf{X})$ and $g(\cdot)$ is a distribution function, the conditional TE $g\{\alpha + \beta + (\boldsymbol{\gamma}_1 + \boldsymbol{\gamma}_0)^{\intercal}\mathbf{X}\} - g(\alpha + \boldsymbol{\gamma}_0^{\intercal}\mathbf{X})$ is a function of both the main effect $\boldsymbol{\gamma}_0$ and the interaction effect $\boldsymbol{\gamma}_1$. Second, the standard Wald test restricts attention to linear marker effects. When the markers affect the outcome non-linearly or interactively, the Wald test may have little power in detecting the signal. In this paper, we propose a kernel machine (KM) based score test for identifying markers predictive of TE. The proposed KM testing procedure can effectively incorporate non-linear effects and capture predictors that are predictive of treatment difference. We focus on the treatment difference scale because the value function of an ITR, $\mathcal{I}_{\mathbf{X}} : \mathbf{X} \to \{0, 1\}$, in improving expected population outcome is directly captured by the treatment difference: $E\{\mathcal{I}_{\mathbf{X}}Y^{(1)} + (1 - \mathcal{I}_{\mathbf{X}})Y^{(0)}\} = E\{\mathcal{I}_{\mathbf{X}}(Y^{(1)} - Y^{(0)})\} + E(Y^{(0)})$.

The remaining of the paper is organized as follows. We introduce the KM test for ITR in section 2.2.1. We describe the resampling procedure for approximating the null distribution in section 2.2.2. Additional considerations including combining information from multiple tuning parameters and dimension reduction via kernel principal component analysis (PCA) are given in section 2.2.3. In section 3.3, we present simulation results

suggesting that the proposed procedures out-performs the traditional Wald test in various settings. The proposed procedures are applied to two randomized clinical trials in 2.3.2 and 2.3.3. We conclude with some remarks in section 2.4.

2.2 Treatment Selection Model

2.2.1 Score Statistic for Identifying Important Baseline Predictors for Treatment Selection

Suppose data for analysis comes from a randomized clinical trial (RCT), and consist of independent and identically distributed random variables $\{(Y_i, T_i, \mathbf{X}_i^{\intercal})^{\intercal}, i = 1, \ldots, n\}$, where Y is the disease outcome, T is a binary treatment indicator (1 for new treatment and 0 for standard treatment), and $\mathbf{X}$ represents baseline predictors. Let $Y^{(1)}$ and $Y^{(0)}$ be the counterfactual outcomes under the new and standard treatment, respectively.
To determine whether $\mathbf{X}$ is useful for guiding treatment selection, we quantify the TE of $\mathbf{X}$ based on the conditional treatment difference

$$\Delta(\mathbf{X}) = \mu_1(\mathbf{X}) - \mu_0(\mathbf{X}),$$

where $\mu_k(\mathbf{X}) = E(Y^{(k)}|\mathbf{X})$. Thus $\mathbf{X}$ is not informative for treatment selection if $\mu_1(\mathbf{X}) - \mu_0(\mathbf{X})$ is a constant. Thus, we aim to develop efficient testing procedures for the null hypothesis

$$H_0 : \mu_1(\mathbf{X}) - \mu_0(\mathbf{X}) = \Delta_0, \tag{2.1}$$

where the constant $\Delta_0 = E\{\mu_1(\mathbf{X}) - \mu_0(\mathbf{X})\} = \mu_1 - \mu_0$ and $\mu_k = E(Y^{(k)})$. It is not difficult to see that under H_0,

$$\mathbf{R}_{\boldsymbol{\psi}} = \text{cov}\left\{Y^{(1)} - Y^{(0)}, \boldsymbol{\psi}(\mathbf{X})\right\} = \text{E}\left\{(Y^{(1)} - Y^{(0)} - \Delta_0)\boldsymbol{\psi}(\mathbf{X})\right\} = 0, \quad \text{for any } \boldsymbol{\psi}(\cdot).$$

and thus we propose to test (3.3) by constructing a test statistic summarizing the overall magnitude of $\mathbf{R}_{\boldsymbol{\psi}}$. To this end, we first obtain an empirical estimate of $\mathbf{R}_{\boldsymbol{\psi}}$ based on the observed RCT data. Specifically, by employing an inverse probability weighting (IPW) (Rotnitzky and Robins, 2005) estimator for the counterfactuals, we estimate $\mathbf{R}_{\boldsymbol{\psi}}$ as

$$\widehat{\mathbf{R}}_{\boldsymbol{\psi}} = n^{-1} \sum_{i=1}^{n} \widehat{\delta}_i \boldsymbol{\psi}(\mathbf{X}_i) = n^{-1} \widehat{\Delta}^{\intercal} \boldsymbol{\Psi} \tag{2.2}$$

where $\boldsymbol{\Psi} = [\boldsymbol{\psi}(\mathbf{X}_1), \ldots, \boldsymbol{\psi}(\mathbf{X}_n)]^{\intercal}$, $\widehat{\Delta} = \left[\widehat{\delta}_1, \ldots, \widehat{\delta}_n\right]^{\intercal}$, $\bar{Y}_k = n_k^{-1} \sum_{\{T_i=k\}} Y_i$, $n_k = \sum_{i=1}^{n} I(T_i = k)$,

$$\widehat{\delta}_i = \frac{(Y_i - \bar{Y}_1) I(T_i = 1)}{\hat{\pi}_1} - \frac{(Y_i - \bar{Y}_0) I(T_i = 0)}{\hat{\pi}_0}, \quad \text{and} \quad \widehat{\pi}_k = \frac{n_k}{n} \tag{2.3}$$

In order to test whether (2.2) is close to $\mathbf{0}$, the standard score-type test statistic takes the form of $\widehat{\mathbf{R}}_{\boldsymbol{\psi}}^{\intercal} \widehat{\Sigma}_{\widehat{\mathbf{R}}_{\boldsymbol{\psi}}}^{-1} \widehat{\mathbf{R}}_{\boldsymbol{\psi}}$ and is approximately χ_q^2, where $\widehat{\Sigma}_{\widehat{\mathbf{R}}_{\boldsymbol{\psi}}}$ is the variance-covariance matrix estimate of $\widehat{\mathbf{R}}_{\boldsymbol{\psi}}$ and q is the dimension of $\boldsymbol{\psi}(\mathbf{X})$. However, such a test may suffer from power loss when $\boldsymbol{\psi}(\mathbf{X})$ are correlated. In addition, the χ_q^2 distribution may not approximate the null distribution well especially when the covariance matrix is near singular. We instead summarize the overall effect of $\mathbf{X}$ based on the L_2 norm of (2.2) and propose the test statistic:

$$\widehat{Q}_{\boldsymbol{\psi}} = ||n^{\frac{1}{2}} \widehat{\mathbf{R}}_{\boldsymbol{\psi}}||^2 = n^{-1} \widehat{\Delta}^{\intercal} (\boldsymbol{\Psi} \boldsymbol{\Psi}^{\intercal}) \widehat{\Delta} \tag{2.4}$$

This type of score test statistic has been shown to be a powerful alternative to the standard score test and can be viewed as a variance component test under various settings (Liu et al., 2007; Kwee et al., 2008; Wu et al., 2010; Cai et al., 2011b).

The choice of the basis functions $\psi(\cdot)$ has a significant impact on the power of the resulting test. If the basis functions efficiently capture the non-linear characteristic of the data, one may achieve great power gain comparing to using the original data. However, in practice, it is often difficult to explicitly specify $\boldsymbol{\psi}(\cdot)$ to optimize power since prior knowledge of the underlying functional form is generally not available. We propose to overcome this dif-

ficulty by implicitly specifying the basis functions using the Reproducible Kernel Hilbert Space (RKHS). Let $\mathcal{H}_k$ be a RKHS generated by a given positive definite kernel function $k(\cdot,\cdot;\rho)$, and ρ is some tuning parameter associated with the kernel function (Cristianini and Shawe-Taylor, 2000), where the kernel function $k(\mathbf{x}_1,\mathbf{x}_2;\rho)$ measures the similarity between $\mathbf{x}_1$ and $\mathbf{x}_2$ and different choices of k lead to different RKHS. Some of the popular kernel functions include the gaussian kernel $k(\mathbf{x}_1,\mathbf{x}_2;\rho) = \exp\{-\|\mathbf{x}_1-\mathbf{x}_2\|_2^2/2\rho^2\}$, which can be used to capture complex smooth non-linear effects; the linear kernel $k(\mathbf{x}_1,\mathbf{x}_2;\rho) = \rho + \mathbf{x}_1^\intercal\mathbf{x}_2$ which corresponds to $h(\mathbf{x})$ being linear in $\mathbf{x}$; and the quadratic kernel $k(\mathbf{x}_1,\mathbf{x}_2;\rho) = (\mathbf{x}_1^\intercal\mathbf{x}_2+\rho)^2$, which allows for 2-way interactive effects. By Mercer's Theorem (Cristianini and Shawe-Taylor, 2000), any $h(\mathbf{x}) \in \mathcal{H}_k$ has a $primal\ representation$ with respect to the eigensystem of k. Specifically, under the probability measure of $\mathbf{x}$, k has eigenvalues $\{\lambda_l, l=1,...,\mathcal{J}\}$ with $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_{\mathcal{J}}$ and the corresponding eigenfunctions $\{\phi_l, l=1,...,\mathcal{J}\}$ such that $k(\mathbf{x}_1,\mathbf{x}_2) = \sum_{l=1}^{\mathcal{J}} \lambda_l\phi_l(\mathbf{x}_1)\phi_l(\mathbf{x}_2)$, where $\mathcal{J}$ could be infinity and $\lambda_l > 0$ for any $l < \infty$. The basis functions, $\{\psi_l(\mathbf{x}) = \sqrt{\lambda_l}\phi_l(\mathbf{x}), l=1,...,\mathcal{J}\}$, span the RKHS $\mathcal{H}_k$. These basis functions can potentially be used in (2.2). We note that the kernel functions may depend on the tuning parameter ρ. For the ease of presentation, we suppress ρ from k although procedures for incorporating different choices of ρ in testing will be detailed in Section 2.2.3.

However, the basis functions $\{\psi_l(\cdot)\}$ depend on the unknown distribution of $\mathbf{x}$ and thus is not directly available. To estimate $\{\psi_l(\cdot)\}$, we apply a singular value decomposition to the observed kernel matrix $\mathbb{K}_n = [k(\mathbf{x}_i,\mathbf{x}_j)]_{n\times n}$:

$$\mathbb{K}_n = \widehat{\mathbf{\Phi}}\hat{\Lambda}\widehat{\mathbf{\Phi}}^\intercal = \widehat{\mathbf{\Psi}}\widehat{\mathbf{\Psi}}^\intercal, \quad \text{where} \quad \widehat{\mathbf{\Psi}} = \widehat{\mathbf{\Phi}}\text{diag}\{\lambda_1^{1/2},\cdots,\lambda_n^{1/2}\},$$

where $\hat{\Lambda} = \text{diag}\{a_1,\ldots,a_n\}$, $a_1 \geq \ldots \geq a_n \geq 0$ are the eigenvalues of $\mathbb{K}_n$ and $\widehat{\mathbf{\Phi}} = (\widehat{\boldsymbol{\phi}}_1,\ldots,\widehat{\boldsymbol{\phi}}_n)$ are the corresponding eigenvectors. It has been shown that $\widehat{\mathbf{\Psi}}$ is effectively estimating the basis functions evaluated at the sample points, $\mathbf{\Psi} = [\psi_j(\mathbf{X}_i)]_{n\times n}$ (Koltchinskii and Giné, 2000; Braun et al., 2005). Replacing $\mathbf{\Psi}$ in (2.4) by $\widehat{\mathbf{\Psi}}$ constructed

from the RKHS framework, our KM score test statistic for ITR takes the form

$$\hat{Q}_{\boldsymbol{\psi}} = \frac{1}{n}\widehat{\Delta}^{\intercal}\widehat{\boldsymbol{\Psi}}\widehat{\boldsymbol{\Psi}}^{\intercal}\widehat{\Delta} = \frac{1}{n}\widehat{\Delta}^{\intercal}\mathbb{K}_n\widehat{\Delta}. \tag{2.5}$$

We next detail procedures for approximating the null distribution of the statistic $\hat{Q}_{\boldsymbol{\psi}}$.

2.2.2 Approximating the Null Distribution by Resampling Procedure

To approximate the distribution of (2.5) under H_0, we write in Appendix that

$$\hat{Q}_{\boldsymbol{\psi}} = n^{-1}\int\int k(\mathbf{x},\mathbf{x}')d\widehat{\Theta}(\mathbf{x})d\widehat{\Theta}(\mathbf{x}') \tag{2.6}$$

where $\widehat{\Theta}(\mathbf{x}) = n^{-\frac{1}{2}}\sum_{i=1}^{n}\theta_i(\mathbf{x}) + o_P(1)$, and $\theta_i(\mathbf{x})$ are the influence functions:

$$\theta_i(\mathbf{x}) = \left\{\frac{(Y_i-\mu_1)I(T_i=1)}{\pi_1} - \frac{(Y_i-\mu_0)I(T_i=0)}{\pi_0}\right\}\{I(\mathbf{X}_i\le\mathbf{x}) - \mathcal{F}(\mathbf{x})\} \tag{2.7}$$

where $\pi_k = P(T=k)$ and $\mathcal{F}(\mathbf{x}) = P(\mathbf{X}\le\mathbf{x})$. We show in **??** that $n^{-\frac{1}{2}}\widehat{\Theta}(\mathbf{x})$ converges weakly to zero-mean Gaussian process $G(\mathbf{x})$ and hence

$$\hat{Q}_{\boldsymbol{\psi}} = n^{-1}\int\int k(\mathbf{x},\mathbf{x}')d\widehat{\Theta}(\mathbf{x})d\widehat{\Theta}(\mathbf{x}') \to \int\int k(\mathbf{x},\mathbf{x}')dG(\mathbf{x})dG(\mathbf{x}'), \quad \text{in distribution.}$$

The limiting null distribution of $\hat{Q}_{\boldsymbol{\psi}}$ takes a complex form, making explicit estimation infeasible. We propose to approximate the null distribution of $\hat{Q}_{\boldsymbol{\psi}}$ via perturbation resampling, which has been used successfully in the literature to approximate the distribution of a wide range of regular estimators (Park and Wei, 2003; Cai et al., 2005; Tian et al., 2007). Specifically, for a large number B, we generate independent standard normal random variables, $\{\mathbf{V}^{(b)} = (V_1^{(b)},...,V_n^{(b)}), b = 1,...B\}$, independent of the observed data. For $b = 1,...,B$, let

$$\widehat{\Theta}^{(b)}(\mathbf{x}) = n^{-\frac{1}{2}}\sum_{i=1}^{n}\widehat{\theta}_i(\mathbf{x})V_i^{(b)},$$

be the bth perturbed realization of $\widehat{\Theta}(\mathbf{x})$, where $\widehat{\theta}_i(\mathbf{x}) = \widehat{\delta}_i\{I(\mathbf{X}_i\le\mathbf{x}) - \widehat{\mathcal{F}}(\mathbf{x})\}$ and $\widehat{\mathcal{F}}(\mathbf{x}) =$

$\sum_{l=1}^{n} I(\mathbf{X}_l \leq \mathbf{x})$. Subsequently, we obtain the perturbed counterpart of $\hat{Q}_{\psi}$ as

$$\begin{aligned}\hat{Q}_{\boldsymbol{\psi}}^{(b)} &= \int\int k(\mathbf{x},\mathbf{x}')d\widehat{\Theta}^{(b)}(\mathbf{x})d\widehat{\Theta}^{(b)}(\mathbf{x}') \\ &= n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\widehat{\delta}_i\widehat{\delta}_j V_i^{(b)}V_j^{(b)}\left[k(x_i,x_j) - n^{-1}\sum_{l=1}^{n}\{k(x_l,x_j)+k(x_i,x_l)\} + n^{-2}\sum_{l=1}^{n}\sum_{l'=1}^{n}k(x_l,x_{l'})\right] \\ &= (\widehat{\boldsymbol{\Delta}}\odot\mathbf{V}^{(b)})^{\intercal}\mathbb{K}^{*}(\widehat{\boldsymbol{\Delta}}\odot\mathbf{V}^{(b)}) \qquad (2.8)\end{aligned}$$

where $\mathbb{K}^{*} = \mathbb{K} - \mathbf{e}_n^{\intercal}\mathbb{K} - \mathbb{K}^{\intercal}\mathbf{e}_n + \mathbf{e}_n^{\intercal}\mathbb{K}\mathbf{e}_n$, $\mathbf{e}_n = n^{-1}\mathbf{1}_{n\times 1}$, and for any vectors $\boldsymbol{a}$ and $\mathbf{b}$, $\boldsymbol{a}\odot\mathbf{b}$ denotes element-wise product. The null distribution of $\hat{Q}_{\psi}$ can be approximated by the empirical distribution of $\{\hat{Q}_{\psi}^{(b)}, b = 1,\cdots,B\}$. For an observed $\hat{Q}_{\psi}$, the p-value can be estimated as $\frac{1}{B}\sum_{b=1}^{B} I(\hat{Q}_{\psi}^{(b)} > \hat{Q}_{\psi})$.

2.2.3 Additional Consideration: Scale Parameters and Kernel PCA

Kernels with Scale Parameters Some kernels, such as the Gaussian kernel, involves a scale parameter ρ which has a great impact on the complexity of the resulting RKHS and hence the power of the test. Unfortunately, the parameter ρ is not identifiable under H_0. For the problem of a nuisance parameter ρ disappearing under H_0, Davies (1977) proposed a score test by treating the score statistic as a stochastic process indexed by ρ. Here, we take a similar approach by considering the minimum p-value as the composite test statistic that combines information from multiple choices of ρ. Specifically, let $\{\rho_m, m = 1,...,M\}$ be the list of candidate scale parameters. Let $\hat{Q}_{\boldsymbol{\psi},m}$ and $\widehat{\mathbf{Q}}_{\boldsymbol{\psi},m} = \{\hat{Q}_{\boldsymbol{\psi},m}^{(1)},\cdots,\hat{Q}_{\boldsymbol{\psi},m}^{(B)}\}^{\intercal}$ denote the observed and perturbed test statistic corresponding to kernel $k(\cdot,\cdot,\rho_m)$, respectively, where the same set of perturbation variables $\{\mathbf{V}^{(b)} = (V_1^{(b)},...,V_n^{(b)}), b = 1,...B\}$ are used across all M scale parameters. Let $\hat{S}_m$ denote the empirical survival distribution of $\{\hat{Q}_{\boldsymbol{\psi},m}^{(b)}, b = 1,\cdots,B\}$. Then we define minimum p-value across testing with M scale parameters as

$$\hat{p}_{min} = \min\{\hat{p}_m, m = 1,...,M\}, \quad \text{where} \quad \hat{p}_m = \widehat{S}_m\{\hat{Q}_{\boldsymbol{\psi},m}\}. \qquad (2.9)$$

Although $\hat{p}_m$ is expected to be approximately uniform under H_0, the minimum p-value statistic $\hat{p}_{min}$ is no longer uniformly distributed. Nevertheless, the null distribution of $\hat{p}_{min}$ can be easily approximated using the perturbed realizations $\{\widehat{\mathbf{Q}}_{\boldsymbol{\psi},m}, m = 1, ..., M\}$. Specifically, the empirical distribution of $\{\hat{p}^{(b)}_{min}, b = 1, ..., B\}$ can be used to approximate the null distribution of $\hat{p}_{min}$, where $\hat{p}^{(b)}_{min} = \min\{\hat{p}^{(b)}_m, m = 1, ..., M\}$ and $\hat{p}^{(b)}_m = \widehat{S}_m\{\hat{Q}^{(b)}_{\boldsymbol{\psi},m}\}$.

Kernel PCA When the kernel space $\mathcal{H}_k$ is high dimensional, testing and estimation procedures based on such a space may not be efficient due to the high degrees of freedom (Braun et al., 2005; Cai et al., 2011b). In addition, the null distribution of the test statistic tends to be more difficult to approximate in finite sample, leading to slightly inaccurate type I error (Cai et al., 2011b). One approach to improving the power and maintaining proper size is to effectively reduce the dimensionality. When the eigenvalues of k decay quickly, $\mathcal{H}_k$ can be well approximated by the RKHS spanned by a truncated kernel $k^{(r_n)}(\mathbf{x}_1, \mathbf{x}_2) = \sum_{l=1}^{r_n} \lambda_l \phi_l(\mathbf{x}_1)\phi_l(\mathbf{x}_2)$, for some r_n such that $\sum_{l=r_n+1}^{\mathcal{J}} \lambda_l = o(\sum_{l=1}^{\mathcal{J}} \lambda_l)$. The error $\mathcal{E}_n = \|\mathbb{K}_n - \mathbb{K}_n^{(r_n)}\|$ can be bounded by $O\{\lambda_r + \sum_{l=r_n+1}^{\infty} \lambda_l\}$, where $\mathbb{K}_n^{(r_n)}$ is the kernel matrix constructed from kernel $k^{(r_n)}$ (Braun et al., 2005, Theorem 3.7). In many practical situations with fast decaying eigenvalues for k, r_n is typically fairly small and we can effectively approximate $\mathcal{H}_k$ by a finite dimensional space. Although $\mathbb{K}_n^{(r_n)}$ is generally not attainable directly in practice, we may use kernel PCA to approximate $\mathbb{K}_n^{(r_n)}$ as

$$\widetilde{\mathbb{K}}_n^{(r_n)} = [\widehat{\boldsymbol{\phi}}_1, ..., \widehat{\boldsymbol{\phi}}_{r_n}]\text{diag}\{a_1, ..., a_{r_n}\}[\widehat{\boldsymbol{\phi}}_1, ..., \widehat{\boldsymbol{\phi}}_{r_n}]^\intercal = [\widehat{\boldsymbol{\psi}}_1, ..., \widehat{\boldsymbol{\psi}}_{r_n}][\widehat{\boldsymbol{\psi}}_1, ..., \widehat{\boldsymbol{\psi}}_{r_n}]^\intercal.$$

Replacing $\mathbb{K}_n$ by $\widetilde{\mathbb{K}}_n^{(r_n)}$ in (2.5), we obtain the kernel PCA approximated test statistic

$$\hat{Q}_{PCA} = n^{-1}\widehat{\boldsymbol{\Delta}}^\intercal \widetilde{\mathbb{K}}_n^{(r_n)} \widehat{\boldsymbol{\Delta}} = \sum_{l=1}^{r_n} \|\widehat{\boldsymbol{\Delta}}^\intercal \widehat{\boldsymbol{\psi}}_l\|_2^2, \tag{2.10}$$

Obviously, $\hat{Q}_{PCA}$ reduces to $\hat{Q}_{\boldsymbol{\psi}}$ when $r_n = n$.

Range of ρ It is also important to choose the appropriate range of $\{\rho_m, m = 1, ..., M\}$, since the range will affect the size and power of the procedures. We use a data adaptive approach to select the range by taking into account the eigenvalues decay rate of the kernel for a given ρ, $\alpha(\rho)$, where we assume that $\lambda_j(\rho) = O\{j^{-\alpha(\rho)}\}$. We estimate the decay rate as the slope from fitting a robust linear regression $\log\{a_j(\rho)\} = \alpha \log(j) + \epsilon$ with $j = 1, ..., r_n$. The range of ρ is chosen such that the corresponding estimated $\alpha(\rho)$ is between 1.2 and 2 and the vector $\{\rho_m, m = 1, ..., M\}$ is equally spaced on the logarithm scale within this range. When we select how many components to use in the singular value decomposition of $\mathbb{K}_n$, we choose r_n as the smallest r such that the estimated proportion of variation explained by the first r eigenfunctions, defined as $\{\sum_{l=1}^{r_n} a_l\}/\{\sum_{l=1}^{n} a_l\}$, is at least 0.99.

2.3 Numerical Studies

2.3.1 Simulation Study

We performed extensive simulation study to compare the performances of our proposed procedure to the Wald test with sandwich estimator for covariance matrix. We carried out our procedure with three kernels (i) linear (k_L), (ii) quadratic (k_Q) and (iii) gaussian kernel (k_G). For conciseness, we only present results from the kernel PCA procedure where we select the first r_n eigenvectors that account for 99% of total variation. We studied both continuous and binary outcomes. The predictor $X_{p\times 1}$ was generated from multivariate normal with mean zero, variance 4 and correlation ρ, where we let $p = 5$ and 20, and $\rho = 0.2$ and 0.5. We considered a total sample size of $n = 500$ and 1000. The treatment indicator T is generated from Bernoulli(0.5).

For continuous outcome, we generate Y from the regression model $Y = -35+T+h_0(\mathbf{X})+h_1(\mathbf{X})T + X_1X_2\epsilon$, where ϵ follows a standard normal. Three different settings were con-

sidered for the predictor effect functions $h_0(\mathbf{X})$ and $h_1(\mathbf{X})$:

(ii) Null: $h_0(\mathbf{X}) = 0$; $h_1(\mathbf{X}) = 0$

(ii) Linear Effects: $h_0(\mathbf{X}) = X_3/2$; $h_1(\mathbf{X}) = (X_1 + X_2 + X_5)/3$

(iii) Nonlinear Effects: $h_0(\mathbf{X}) = X_3/2$; $h_1(\mathbf{X}) = (X_1^2 + X_5^2 + X_1X_5 + X_1 + X_5)/2$

For binary outcome, we generated Y from a logistic regression model $\text{logit}\{p(Y = 1|\mathbf{X}, T)\} = 0.3T + h_0(\mathbf{X}) + h_1(\mathbf{X})T$. Three settings were considered for $h_0(\mathbf{X})$ and $h_1(\mathbf{X})$:

(i) Null: $h_0(\mathbf{X}) = 0$; $h_1(\mathbf{X}) = 0$

(ii) Linear Effects: $h_0(\mathbf{X}) = X_3/2$; $h_1(\mathbf{X}) = (X_1 + X_2 + X_5)/5$

(iii) Nonlinear Effects: $h_0(\mathbf{X}) = X_1$; $h_1(\mathbf{X}) = X_5^2/4 + X_3 \times X_5/2 + \Phi(3X_3) \times X_5/3$

where Φ is the cumulative distribution function of standard normal.

In Table 1, we present results for continuous Y when the test is performed at type I error rate of 0.05. Under H_0, the empirical size of all procedures are reasonably close to the nominal level of 0.05. The proposed test with k_Q tend to be be slightly conservative when $p = 20$ and $\rho = 0.2$ due to the high dimensionality of the associated RKHS. Under the alternative with linear effects, our procedure with linear kernel has similar performance as the Wald test when the correlation among predictors is small and $p = 5$. However, as p and the correlation among predictors increase, our proposed procedure with k_L outperforms the Wald test. For example, when $\rho = 0.5$ and $p = 20$, the Wald test has power of 0.357 for $n = 500$ and 0.562 for $n = 1000$; while our proposed method with with k_L has power of 0.499 for $n = 500$ and 0.709 for $n = 1000$. The power loss in the Wald can in part be attributed to the use of a p degree of freedom (DF) when the effective DF in the presence of high correlation could be much lower than p. On the contrary, our proposed test leverages the correlation, resulting a lower effective DF. When the effects are linear, our proposed test with k_Q suffers some power loss when $n = 500$ but has comparable power when $n = 1000$. On the other hand, our KM score test with k_G out performs all

other tests even when the effects are linear. This is not surprising since when the scale parameter ρ is large, the RKHS associated with $k(\cdot, \cdot; \rho)$ approximates the linear space (Cai et al., 2011b) while allowing ρ to vary enables us to choose different basis functions to more efficiently capture the effects. When the underlying effects are non-linear, both the wald test and the KM score test with k_L perform poorly with low power, as expected. The KM score tests with both k_Q and k_G have substantially higher power across all settings. It is interesting to note that although the underlying effects are quadratic, the KM test with k_G has comparable or higher power when p is small. For the larger p of 20, the test with k_Q substantially outperforms k_G. One possible explanation is that the RKHS with k_G may not be an efficient approximation to capture $h_1(\mathbf{X}) - h_0(\mathbf{X})$ when compared to that based on k_Q.

The results for binary outcome are presented in Table 2. All procedures maintain the type I error reasonably well although in this setting the Wald test has a slightly conservative size when $p = 20$. Unlike the setting with continuous outcome, our test is no longer expected to perform similarly to the Wald test even when the effects are linear since the two tests are capturing different aspects of the TE. When $p = 5$, the proposed test and the Wald test perform similarly. However, when $p = 20$, the KM score test with k_L substantially outperform the Wald test. For example, when $p = 20$, $\rho = 0.5$ and $n = 500$, the empirical power is 0.796 for the KM score test and only 0.444 for the Wald test. In this setting, the KM test with k_Q and k_G also perform quite comparably to the test with linear kernel, demonstrating the robustness of the test with non-linear kernels. When the underlying effects are non-linear, the KM test with k_Q generally perform better than the tests assuming linear effects. Since the non-linear signals are mostly quadratic, the KM test with k_Q is generally more powerful than those from k_G although the procedures have similar performances when $p = 5$.

Table 2.1: Sizes and powers for different methods, under various sample size, number of predictors, and correlation among predictors, with continuous outcome

			size		nonlinear		linear	
		method	n=500	n=1000	n=500	n=1000	n=500	n=1000
$\rho = 0.2$	p=5	Wald	0.058	0.048	0.313	0.522	0.557	0.816
		k_L	0.051	0.051	0.364	0.591	0.523	0.793
		k_Q	0.040	0.041	0.805	0.982	0.500	0.704
		k_G	0.036	0.036	0.998	1.000	0.621	0.910
	p=20	Wald	0.051	0.046	0.275	0.401	0.458	0.693
		k_L	0.041	0.046	0.328	0.525	0.530	0.755
		k_Q	0.028	0.035	0.648	0.923	0.443	0.754
		k_G	0.038	0.048	0.401	0.789	0.505	0.771
$\rho = 0.5$	p=5	Wald	0.056	0.050	0.275	0.383	0.454	0.658
		k_L	0.055	0.048	0.314	0.492	0.510	0.712
		k_Q	0.042	0.042	0.878	0.988	0.390	0.584
		k_G	0.042	0.040	0.999	1.000	0.618	0.917
	p=20	Wald	0.039	0.041	0.215	0.319	0.357	0.562
		k_L	0.052	0.046	0.311	0.460	0.499	0.709
		k_Q	0.045	0.039	0.818	0.969	0.390	0.623
		k_G	0.049	0.045	0.723	0.992	0.541	0.790

Table 2.2: Sizes and powers for different methods, under various sample size, number of predictors, and correlation among predictors, with binary outcome

			size		nonlinear		linear	
		method	n=500	n=1000	n=500	n=1000	n=500	n=1000
$\rho = 0.2$	p=5	Wald	0.043	0.048	0.369	0.676	0.811	0.990
		k_L	0.050	0.056	0.397	0.639	0.832	0.991
		k_Q	0.046	0.053	0.755	0.986	0.826	0.988
		k_G	0.054	0.056	0.729	0.992	0.800	0.991
	p=20	Wald	0.028	0.033	0.134	0.346	0.427	0.889
		k_L	0.049	0.047	0.304	0.539	0.690	0.946
		k_Q	0.048	0.046	0.584	0.873	0.643	0.938
		k_G	0.050	0.055	0.338	0.576	0.690	0.944
$\rho = 0.5$	p=5	Wald	0.047	0.049	0.726	0.974	0.844	0.994
		k_L	0.047	0.053	0.903	0.998	0.859	0.994
		k_Q	0.050	0.050	0.981	1.000	0.846	0.986
		k_G	0.058	0.044	0.944	1.000	0.890	0.993
	p=20	Wald	0.026	0.037	0.320	0.802	0.444	0.911
		k_L	0.054	0.049	0.863	0.997	0.796	0.977
		k_Q	0.050	0.056	0.932	0.999	0.755	0.973
		k_G	0.052	0.050	0.865	0.994	0.834	0.989

2.3.2 Example: Predictors Useful for Individualized Treatment of HIV Infected Patients

We apply our method to data from AIDS Clinical Trials Group Protocal 175 (ACTG175), which is a double-blind study that evaluated treatment with either a single nucleoside or two nucleosides in adults infected with human immunodeficiency virus type 1 (HIV-1) (Hammer et al., 1996). The dataset contains 2139 HIV-infected subjects, where subjects were randomized to four different treatment groups: zidovudine (ZDV) monotherapy, ZDV+didanosine (ddI), ZDV+zalcitabine and ddI monotherapy. Following the primary goal of the original study, we compare ZDV monotherapy ($T = 0$) to combination therapies ($T = 1$) and aim to identify baseline predictors that are associated with differential TE. We considered the long term immune response, defined as 96 ($\pm$ 5) week CD4 counts, CD4_{96}, as the continuous outcome which was also used in Tsiatis et al. (2008). To test for predictors for ITR, we included 12 baseline covariates separated into 3 groups: (i) demographic information including age, weight, race and gender; (ii) risk factors including hemophilia status, homosexual activity, antiretroviral history, symptomatic status and history of intravenous drug use; and (iii) functional markers including Karnofsky score, baseline CD4 and baseline CD8 count. The goal is to test whether any group of covariates significantly affects the absolute risk reduction by different treatments, so the variables in the significant group can be used to guide treatment selection in the future. We apply the Wald test, our KM score test with k_L, k_Q and k_G to each of the covariates group. The results for the response being the continuous CD4_{96} as defined are shown in Table 2.3(a). Our proposed method detected functional markers as being significantly predictive of treatment response with p-value about 0.01 and the demographic variables as being marginally significant with p-value 0.07 when the gaussian kernel is employed. On the other hand, the Wald test identified none of the predictor groups as significant.

Table 2.3: P-value for testing the overall effects of different groups of baseline predictors from the Wald test and the proposed KM score test with three kernels: k_L, k_Q and k_G.

(a) Treatment Effect on week 96 CD4 counts ACTG175

	demographic	risk factors	functional markers
Wald	0.18	0.96	0.27
k_L	0.24	0.99	0.02
k_Q	0.25	0.99	0.01
k_G	0.07	0.72	0.01

(b) Treatment Effect on PGA Response with BEST Study

	ICE	PE	HLT	CLT	CH
Wald	0.22	0.12	0.41	0.33	0.20
k_L	0.19	0.04	0.51	0.38	0.18
k_Q	0.05	0.09	0.13	0.06	0.23
k_G	0.18	0.01	0.08	0.19	0.23

2.3.3 Example: Predictors Useful for Treatment of Patients with Advanced Chronic Heart Failure

We also illustrate the proposed procedures using the Beta-Blocker Evaluation of Survival Trial (BEST), which is a randomized clinical trial to investigate if Bucindolol, a beta-blocker, would benefit patients with advanced chronic heart failure (CHF) (of Survival Trial Investigators et al., 2001). The 2-year BEST study had 2708 participants randomized to receive either Bucindolol or Placebo with equal probability. We considered the Physician's Global Assessment (PGA) as the primary response of interest. The PGA takes seven ordinal levels (1-3: different levels of worsening, 4:no change, 5-7: different levels of improvement) and we defined a binary outcome Y if the PGA ≥ 4, reflecting some improvement. For baseline predictors, we considered four groups with grouping information provided in the original study database: (i) Ischemic CHF Etiology (ICE; 6 covariates), including prior myocardio infarction, stenosis, coronary artery disease etiology and so on; (ii) Physical Exam (PE; 14 covariates), including heart rate, blood pressure, weight, height etc; (iii) Hematology Lab Test (HLT; 4 covariates): hematocrit, hemoglobin, platelet, and white blood count; (iv) Chemistry Lab Test (CLT; 19 covariates), including

Clucose, Sodium, Calcium etc; and (v) Cardiac History (CH; 9 covariates), including Duration of CHF, Peripheral Vascular disease etc. The goal is to test whether any of these groups are significantly associated with treatment difference with respect to the binary outcome Y reflecting improvement in PGA.

Results given in Table 2.3(b) suggest that Physical Exam may be a strong predictor of treatment difference with p-value 0.01 from the KM score test with k_G. Results of the KM test with k_L and k_Q are consistent with p-values 0.04 and 0.09, respectively. There is also suggestive evidence that Ischemic and CHF etiology may be associated with treatment response with a marginally significant p-value from the KM test with k_Q although the test is not significant for other kernels. Again, the Wald test failed to reject for any of the predictor groups.

2.4 Discussion

In this paper, we proposed a KM based score test to identify informative baseline predictors that can be useful for individualized treatment selection. Our method is robust due to the model-free construction of the statistic. Our proposed KM test is also generally more powerful than the existing Wald test. Numerical studies suggest that our proposed procedures could substantially outperform the Wald test, especially when testing for a moderate number of predictors that are correlated with each other and/or when the underlying effects are non-linear. Different kernel functions may be preferable for different types of signals. Via the resampling approach, it is not difficult to construct an omnibus test that combines information from multiple kernels, using procedures such as minimum p-value as those illustrated in section 2.2.3.

2.5 Appendix: Convergence of the Proposed Test Statistic

We can write the statistic (2.5) in terms of stochastic process:

$$\hat{Q}_{\boldsymbol{\psi}} = n^{-1}\widehat{\Delta}^{\intercal}\mathbb{K}_n\widehat{\Delta} = \int\int k(\mathbf{x},\mathbf{x}')d\widehat{\Theta}(\mathbf{x})d\widehat{\Theta}(\mathbf{x}') \tag{2.11}$$

where $\widehat{\Theta}(\mathbf{x}) = n^{-\frac{1}{2}}\sum_{i=1}^{n}\frac{(Y_i-\bar{Y}_1)I(T_i=1)}{\hat{\pi}_1}I(\mathbf{x}_i\leq\mathbf{x}) - n^{-\frac{1}{2}}\sum_{i=1}^{n}\frac{(Y_i-\bar{Y}_0)I(T_i=0)}{\hat{\pi}_0}I(\mathbf{x}_i\leq\mathbf{x})$.
In order to derive the influence function of $\widehat{\Theta}(\mathbf{x})$, we write the first part of $\widehat{\Theta}(\mathbf{x})$ in the following way:

$$\begin{aligned}
&n^{-\frac{1}{2}}\sum_{i=1}^{n}\frac{(Y_i-\bar{Y}_1)I(T_i=1)}{\hat{\pi}_1}I(\mathbf{X}_i\leq\mathbf{x})\\
=&n^{-\frac{1}{2}}\sum_{i=1}^{n}\frac{(Y_i-\mu_1)I(T_i=1)I(\mathbf{X}_i\leq\mathbf{x})}{\pi_1} - n^{\frac{1}{2}}\frac{(\bar{Y}_1-\mu_1)}{\pi_1}\left[n^{-1}\sum_{i=1}^{n}I(T_i=1)I(\mathbf{X}_i\leq\mathbf{x})\right] + o_P(1)\\
=&n^{-\frac{1}{2}}\sum_{i=1}^{n}\frac{(Y_i-\mu_1)I(T_i=1)}{\pi_1}\left[I(\mathbf{X}_i\leq\mathbf{x}) - n^{-1}\sum_{i=1}^{n}I(T_i=1)I(\mathbf{X}_i\leq\mathbf{x})\right] + o_P(1)\\
=&n^{-\frac{1}{2}}\sum_{i=1}^{n}\frac{(Y_i-\mu_1)I(T_i=1)}{\pi_1}\left[I(\mathbf{X}_i\leq\mathbf{x}) - \mathcal{F}(\mathbf{x})\right] + o_P(1)
\end{aligned}$$

where $\mathcal{F}(\mathbf{x}) = P(\mathbf{X}_i\leq\mathbf{x})$. Since by a uniform law of large numbers (ULLN) (Pollard, 1990), $n^{-1}\sum_{i=1}^{n}\frac{I(T_i=1)I(\mathbf{X}_i\leq\mathbf{x})}{\pi_1}$ converges in probability to its limit, $\mathcal{F}(\mathbf{x})$, uniformly in $\mathbf{x}$. Therefore,

$$\widehat{\Theta}(\mathbf{x}) = n^{-\frac{1}{2}}\sum_{i=1}^{n}\theta_i(\mathbf{x}) + o_P(1), \tag{2.12}$$

where

$$\begin{aligned}\theta_i =& \frac{(Y_i - \mu_1)I(T_i = 1)}{\pi_1}\left[I(\mathbf{X}_i \leq \mathbf{x}) - \mathcal{F}(\mathbf{x})\right] - \frac{(Y_i - \mu_0)I(T_i = 1)}{\pi_0}\left[I(\mathbf{X}_i \leq \mathbf{x}) - \mathcal{F}(\mathbf{x})\right] \\ =& \left[\frac{(Y_i - \mu_1)I(T_i = 1)}{\pi_1} - \frac{(Y_i - \mu_0)I(T_i = 1)}{\pi_0}\right]\left[I(\mathbf{X}_i \leq \mathbf{x}) - \mathcal{F}(\mathbf{x})\right] \qquad (2.13)\end{aligned}$$

It's not hard to show that $E\theta_i(\mathbf{x}) = 0$. In addition, it follows from a functional central limit theorem (Pollard, 1990) that $\widehat{\Theta}(\mathbf{x})$ converges jointly to a zero mean Gaussian process $G(\mathbf{x})$.

By Lemma A.3 of Bilias et al. (1997) and the strong representation theorem, we have

$$(2.11) \rightarrow \int\int k(\mathbf{x}, \mathbf{x}')dG(\mathbf{x})dG(\mathbf{x}')$$

Identifying Sparse Predictive Markers for Personalized Treatment Selection

Yuanyuan Shen, Ian Barnett and Xihong Lin

Department of Biostatistics

Harvard School of Public Health

3.1 Introduction

When several treatment options are available it is important to select an optimal treatment for patients, since the effectiveness of interventions often varies by patient subgroups. For example, the Trial Assigning Individualized Options for Treatment [Rx] (TAILORx) trial (Sparano, 2006; Zujewski and Kamin, 2008) was developed to integrate the most modern diagnostic tests into clinical decision-making in order to individualize cancer treatments, in which a gene-expression profile appears to be highly predictive of whether chemotherapy is beneficial for treating breast cancer patients. The US Preventive Services Task Force issued new guidelines recommending against routine mammography screening for women under 50 (Nelson et al., 2009). Thus, better understanding of an individual's genomic and other biological characteristic enables more effective response to human variability with improved specificity. Recent advancement of science and technology has led to the discovery of many biological and genetic markers associated with disease outcomes and treatment responses. These new markers combined with traditional clinical assessments hold great potential for identifying subgroups of patients who are most likely to benefit or are at high risk for toxicity from a particular therapy and thus may lead to personalized or tailored medicine.

Motivated by the heterogeneous nature of treatment, a wide range of statistical methods have been developed for individualized treatment rules (ITRs), which is a decision rule that recommends treatments based on an individual's baseline characteristics (Qian and Murphy, 2011; Song and Pepe, 2004; Cai et al., 2011a; Foster et al., 2011; Zhao et al., 2012; Zhang et al., 2012; Zhao et al., 2013). While constructing ITRs, it is important to include only the important markers, since including unrelated predictors will decrease efficiency of the ITRs, while missing important predictors might result in loss of accuracy. Many variable selection procedures have been developed for identifying important baseline predictors for treatment selection, including penalized procedures, detecting important interactive effect and many more (Gunter et al., 2011; Lu et al., 2013; Imai et al.,

2013; Janes et al., 2011). However, many of the variable selection approaches are not consistent in variable selection especially when the number of predictors is big. Instead, one can test for a group of markers' overall effect on treatment effect difference or split the whole set of baseline predictors into meaningful sub-groups, and test for the influence of each group on the effectiveness of treatment. For example, when genome-wide single-nucleotide polymorphism (SNP) data is collected in a clinical trial, and we are interested in whether this genetic information can be used to guide treatment selection, it will be difficult to apply variable selection procedures appropriately in such a high-dimensional setting. We can define sub-groups of SNPs within the same gene, and apply a test for the effect of each gene on the treatment effect difference. We can further construct ITRs based on the important genes. For future applications, we only need to obtain patients' genetic information for the important genes used in the ITRs, so the cost will be reduced. Therefore, it is important to develop a global statistical test to identify informative groups of baseline predictors for treatment selection.

Standard testing procedure for identifying important predictors assume a generalized linear model (GLM) for the data that includes the interaction between treatment and baseline predictors of interest, and a Wald test can be applied to test for the significance of the interaction. Such a procedure suffers from two major limitations. First, if we are interested in the treatment effect difference scale, in the case of linear outcome and the identity link function in the GLM, the effect of baseline predictors on the treatment effect difference is complete captured by the coefficient for the interaction term. However, in the case of binary outcome within non-linear link function, the effect of baseline predictors on the treatment effect difference has a complex form depending on the form of the link function and main effects of baseline predictors as well as interactive effect between treatment and baseline predictors. Therefore, the interaction term may not entirely capture markers' ability in predicting subject specific treatment effect. Second, we might encounter the situation where only a small number of baseline predictors in the group affect the effec-

tiveness of treatment, in which case the Wald test will lose power due to the sparsity of the signal.

The sparsity problem is often encountered in genetic association studies. For example, it has been shown in the analysis of GWAS that there may not be enough power to test single SNPs for marginal associations (Manolio et al., 2009). As a result, region-based analyses have become more popular in genetic association studies (Li and Leal, 2008). Genes, gene networks, and pathways are examples of SNP-sets that contain sparse subsets of SNPs that can contribute to disease risk. Due to the complexity of human disease, methodology that does not require strong marginal SNP effects but is capable of aggregating these small and sparse SNP effects together into a detectable signal is needed. The higher criticism (HC) has been used in SNP-set tests that combine information over all the marginal test statistics, and is ideal for detecting a sparse few disease-associated SNPs out of a much larger pool of unassociated SNPs than comparable methods (Arias-Castro et al., 2011; Wu et al., 2014). The generalized higher criticism (GHC) (Barnett et al., 2015) further takes into account the correlation among predictors and proposes an approach to calculate p-values analytically.

Inspired by the drawbacks of Wald test and the properties of GHC, we propose a scale-independent score test (GHC-ST) to identify important baseline predictors for guiding treatment selection, in which we incorporate the GHC to improve power for detecting groups containing a sparse few signals. When the predictors are correlated, because the Wald test considers each predictor's relationship with treatment effect conditional on the other predictors, signal can be masked by multicollinearity. By instead relying on marginal predictor effects, GHC avoids this problem and can even boost signal by allowing the noise predictors to inherit marginal effects. Simulation results suggest that the proposed procedure out-performs the traditional Wald test in various settings when the signal is sparse. The rest of the paper is organized as follows. We introduce the Wald test and the scale-independent score statistic for identifying important groups of baseline

predictors in section 3.2.1 and 3.2.2, we implement GHC in this context in section 3.2.3, and we proposed a omnibus test combing the ST and GHC-ST in section 3.2.4. Simulation and real data analysis results are in section 3.3 and 3.4, respectively.

3.2 Identifying Informative Baseline Predictors for Treatment selection

3.2.1 Wald Test for Identifying Informative Baseline Predictors for Treatment Selection

Suppose data for analysis comes from a randomized clinical trial (RCT), and consist of independent and identically distributed random variables $\{(Y_i, T_i, \mathbf{X}_i^\intercal)^\intercal, i = 1, \ldots, n\}$, where Y is the disease outcome, T is a binary treatment indicator (1 for new treatment and 0 for standard treatment), and $\mathbf{X}$ represents p-dimensional baseline predictors. let α_0 be the intercept, β_0 be the coefficient of the main effect of treatment T, $\boldsymbol{\beta}_1$ be the p-dimensional coefficient for the main effect of baseline predictions, and $\boldsymbol{\beta}_2$ be the p-dimensional coefficient for the interaction between treatment and baseline predictors. The following regression model with link function $g()$ is assumed for the data:

$$E(Y|T, X) = g(\alpha_0 + \beta_0 T + \boldsymbol{\beta}_1^\intercal \mathbf{X} + T\boldsymbol{\beta}_2^\intercal \mathbf{X}). \tag{3.1}$$

The goal is to examine whether $\mathbf{X}$ is informative for guiding treatment selection, or in other words, whether the difference between two treatments' effects differs with different levels of baseline predictors. We can check for such a dependency by examining the interactive effect between $\mathbf{X}$ and T, which leads to the following hypothesis testing problem:

$$H_0 : \boldsymbol{\beta}_2 = \mathbf{0},$$

and a traditional Wald test can be applied for such testing. We first get the MLE of $\boldsymbol{\beta} = (\alpha_0, \beta_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ and its variance-covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$: $\widehat{\boldsymbol{\beta}} = \text{argmax}_{\boldsymbol{\beta}}\, \widehat{\ell}(\boldsymbol{\beta}; \mathbf{X}, T)$ and $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}} = \left\{ \frac{\partial^2}{\partial \boldsymbol{\beta}^2} \widehat{\ell}(\widehat{\boldsymbol{\beta}}; \mathbf{X}, T, Y) \right\}^{-1}$, where $\widehat{\ell}(\boldsymbol{\beta}; \mathbf{X}, T, Y) = n^{-1} \sum_{i=1}^{n} \log f(\mathbf{x}_i, t_i, y_i | \boldsymbol{\beta})$ is the likelihood for the observed data. The Wald test statistic for $\boldsymbol{\beta}_2$ takes the form $\widehat{\boldsymbol{\beta}}_2^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}_2}^{-1} \widehat{\boldsymbol{\beta}}_2$, where $\widehat{\boldsymbol{\beta}}_2$ is the sub-vector of $\widehat{\boldsymbol{\beta}}$ corresponding to the interaction term and $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}_2}$ is the sub-matrix of $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}$ corresponding to $\boldsymbol{\beta}_2$. The statistic follows a χ^2_p distribution under the null hypothesis.

3.2.2 Score Test for Identifying Important Baseline Predictors for Treatment Selection

As was discussed in the introduction, when we are interested in treatment effect difference scale and g is nonlinear, the interaction in (3.1) may not entirely capture markers' ability in predicting subject-specific treatment effect. So we want to quantify the treatment effect of $\mathbf{X}$ based on conditional treatment difference $\Delta(\mathbf{X}) = \mu_1(\mathbf{X}) - \mu_0(\mathbf{X})$, where $\mu_k(\mathbf{X}) = E(Y^{(k)}|\mathbf{X}), k = 0, 1$, and $Y^{(1)}$ and $Y^{(0)}$ are the counterfactual outcomes under new and standard treatment, and examine whether this quantity is $\mathbf{X}$-dependent. Thus we proposed the following hypothesis testing statement:

$$H_0 : \mu_1(\mathbf{X}) - \mu_0(\mathbf{X}) = \Delta_0, \tag{3.2}$$

where $\mu_k = E\mu_k(\mathbf{X}), k = 1, 2$, $\Delta_0 = \mu_1 - \mu_0$, since if $\mathbf{X}$ is not informative for treatment selection, the treatment difference is a constant with respect to $\mathbf{X}$. It's not hard to see that under (3.2), the covariance between $\mathbf{X}$ and treatment difference is $\mathbf{0}$:

$$\boldsymbol{\gamma} = \text{cov}\left\{Y^{(1)} - Y^{(0)}, \mathbf{X}\right\} = \text{E}\left\{(Y^{(1)} - Y^{(0)} - \Delta_0)\mathbf{X}\right\} = \mathbf{0},$$

and $\boldsymbol{\gamma}$ is a p dimensional vector with each element being the covariance between treatment T and each of the p baseline predictors. Since $\boldsymbol{\gamma}$ is intuitively a measure of the dependency

between treatment difference and $\mathbf{X}$, we proposed to focus on the following hypothesis testing problem:

$$H_0 : \boldsymbol{\gamma} = \mathbf{0} \tag{3.3}$$

And we aim to find a sample estimate of $\boldsymbol{\gamma}$, $\hat{\boldsymbol{\gamma}}$ and a test statistic summarizing the overall magnitude of $\hat{\boldsymbol{\gamma}}$. We first obtain an empirical estimate of $\boldsymbol{\gamma}$ based on the observed RCT data. Specifically, by employing an inverse probability weighting (IPW) estimator (Rotnitzky and Robins, 2005) for the counterfactuals, we estimate $\boldsymbol{\gamma}$ as

$$\widehat{\boldsymbol{\gamma}} = n^{-1} \sum_{i=1}^{n} \widehat{\boldsymbol{\delta}}_i \mathbf{X}_i \tag{3.4}$$

where $\widehat{\boldsymbol{\delta}}_i = \frac{(Y_i - \bar{Y}_1) I(T_i=1)}{\hat{\pi}_1} - \frac{(Y_i - \bar{Y}_0) I(T_i=0)}{\hat{\pi}_0}$, $\bar{Y}_k = n_k^{-1} \sum_{\{T_i=k\}} Y_i$, $n_k = \sum_{i=1}^{n} I(T_i = k)$, and $\hat{\pi} = \frac{n_k}{n}$, $k = 1, 2$.

The large sample property of $\widehat{\boldsymbol{\gamma}}$ can be obtained through the influence functions:

$$n^{\frac{1}{2}} \widehat{\boldsymbol{\gamma}} = n^{-\frac{1}{2}} \sum_{i=1}^{n} \boldsymbol{\theta}_i(\mathbf{X}) + o_p(1),$$

where $\boldsymbol{\theta}_i(\mathbf{X}) = \left\{ \frac{(Y_i - \mu_1) I(T_i=1)}{\pi_1} - \frac{(Y_i - \mu_0) I(T_i=1)}{\pi_0} \right\} \{\mathbf{X}_i - E(\mathbf{X})\}$. By law of large numbers, it is straightforward to get the convergence property: $n^{\frac{1}{2}} \boldsymbol{\gamma} \to N(\mathbf{0}, \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\gamma}}})$, where

$$\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\gamma}}} = \text{Var}\{\boldsymbol{\theta}_i(\mathbf{X})\} = \text{E}\boldsymbol{\theta}_i(\mathbf{X}) \boldsymbol{\theta}_i(\mathbf{X})^{\intercal}$$

An empirical estimate of $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\gamma}}}$ is:

$$\widehat{\Sigma}_{\widehat{\boldsymbol{\gamma}}} = n^{-1} \sum_{i=1}^{n} \widehat{\boldsymbol{\theta}}_i(\mathbf{X}) \widehat{\boldsymbol{\theta}}_i(\mathbf{X})^{\intercal}, \tag{3.5}$$

where $\widehat{\boldsymbol{\theta}}_i(\mathbf{X}) = \widehat{\boldsymbol{\delta}}_i(\mathbf{X}_i - \bar{\mathbf{X}})$, $\bar{\mathbf{X}} = n^{-1} \sum_{i=1}^{n} \mathbf{X}_i$.

We can then summarize the overall effect of $\mathbf{X}$ based on L_2 norm of (3.4) and propose the test statistic $\hat{Q} = ||n^{\frac{1}{2}}\hat{\boldsymbol{\gamma}}||_2^2$. In order to obtain the p-value for $\widehat{Q}$, we propose to approximate the null distribution of $\widehat{Q}$ via perturbation resampling, which has been used successfully in the literature to approximate the distribution of a wide range of regular estimators (Park and Wei, 2003; Cai et al., 2005; Tian et al., 2007). Specifically, for a large number B (=1000), we generate independent standard normal random variables, $\{\mathbf{V}^{(b)} = (V_1^{(b)}, ..., V_n^{(b)}), b = 1, ..., B\}$, independent of the observed data. For $b = 1, ..., B$, let

$$\widehat{\boldsymbol{\gamma}}^{(b)} = n^{-1} \sum_{i=1}^{n} \widehat{\boldsymbol{\theta}}_i(\mathbf{X}) V_i^{(b)},$$

be the bth perturbed realization of $\widehat{\boldsymbol{\gamma}}$, and calculate the L_2 norm $\widehat{Q}^{(b)} = ||n^{\frac{1}{2}}\widehat{\boldsymbol{\gamma}}^{(b)}||_2^2$, $b = 1, ..., B$. For an observed $\widehat{Q}$, the p-value can be estimated as $\frac{1}{B}\sum_{b=1}^{B} I(\widehat{Q}^{(b)} > \hat{Q})$. We call the testing procedure based on $\hat{Q}$ the score test (ST).

3.2.3 Sparsity Assumption and Incorporating Generalized Higher Criticism

In many real world problems, only a sparse few of the group of baseline predictors are informative. We can take this into account by equivalently assuming only a small number of elements in $\boldsymbol{\gamma}$ are non-zero, and formulate the following hypothesis testing problem:

$$\begin{aligned} &H_0 : \boldsymbol{\gamma} = \mathbf{0} \\ &H_1 : \boldsymbol{\gamma} \neq \mathbf{0}, \quad \boldsymbol{\gamma} \text{ is sparse} \end{aligned} \tag{3.6}$$

Under hypothesis testing statement (3.6), the Wald test and the ST might result in reduced power due to the sparsity, since their test statistic allow for the noise to drown the signal. Therefore, we propose implementing statistical testing procedure that detects sparse signals. The higher criticism (HC) (Donoho and Jin, 2004) combines information over all the

marginal test statistics within a set, and is ideal for detecting a sparse few signals out of a much larger pool of candidate variables, and has been shown to be powerful for high sparsity situations. The HC has been successfully applied in SNP-set test where there are a sparse few disease associated SNPs out of a much larger pool of unassociated SNPs, and achieves greater power than comparable methods (Arias-Castro et al., 2011). In our case, $\widehat{\gamma}$ is an asymptotic multivariate normal random variable, with a mean vector $\mathbf{0}$ under H_0, variance-covariance matrix estimate (3.5), and sparse under H_1. If we assume the p elements in $\widehat{\gamma}$ are independent, let $\bar{\Phi}(z)$ be the survival function of the standard normal distribution and $S(z) = \sum_{j=1}^{p} I_{\{|\gamma_j| \geq z\}}$, the HC test statistic is

$$Z_{HC} = \sup_{z>0} \left(\frac{S(z) - 2p\bar{\Phi}(z)}{[2p\bar{\Phi}(z)\{1 - 2\bar{\Phi}(z)\}]^{1/2}} \right) \tag{3.7}$$

If we allow $p \to \infty$, Z_{HC} converges to a Gumbel distribution with a very slow rate of $O\{(\log p)^{-1/2}\}$. HC is original designed for high dimensional settings with independence variables, and for the case where there exists non-identity correlations among $\hat{\gamma}_i$, $i = 1, \cdots, p$, an innovated higher criticism test statistic (iHC) is proposed to use transformed $\hat{\boldsymbol{\gamma}}^{\star} = U^{-1}\hat{\boldsymbol{\gamma}} \sim MVN(\mathbf{0}, \mathbf{I_p})$, where $\mathbf{U}\mathbf{U}^{\intercal} = \hat{\boldsymbol{\Sigma}}$ is the Cholesky decomposition and $\hat{\boldsymbol{\Sigma}}$ is the variance covariance matrix estimate of $\hat{\boldsymbol{\gamma}}^{\star}$ (Hall et al., 2010).

Barnett et al. (2015) discussed about the drawbacks of the original HC and iHC approach. The slow rate of convergence of the HC statistic causes the size of the test to be drastically incorrect when using the asymptotic distribution to calculate p-values for p as large as a million (Barnett and Lin, 2013). When correlation exists among covariates, the iHC also has the drawback of having to first transform the marginal test statistics from $\hat{\boldsymbol{\gamma}}$ into $\hat{\boldsymbol{\gamma}}^{\star}$. In the presence of even moderately small correlation, this can lead to significant loss of power due to the noise diluting the sparse signals after being mixed in the transformation. In addition, for stronger correlations the matrix inverse operation can be quite unstable. Barnett et al. (2015) used the original statistic $\hat{\boldsymbol{\gamma}}$ and analytically calculate the

variance covariant matrix of $S(t)$, and proposed the generalized higher criticism (GHC) test statistic as

$$Z_{GHC} = \sup_{z>0} \left(\frac{S(z) - 2p\bar{\Phi}(z)}{\widehat{Var}S(z)^{1/2}} \right) \tag{3.8}$$

where $\widehat{Var}S(z)$ is the sample estimate for the variance of $S(z)$ under H_0. A p-value calculation for GHC in finite p settings that does not rely on asymptotic is also proposed.

The conditions of the GHC is met in our previous section with the asymptotic normal property of $\widehat{\boldsymbol{\gamma}}$ as well as the sample estimate for the variance-covariance matrix (3.5), and the p-value calculation procedure in Barnett et al. (2015) is applied. We call this test GHC-ST. By implementing the GHC procedure, we expect to observe improved power over the Wald test and the ST in the situations where only a small number of baseline predictors are informative.

3.2.4 The Omnibus Test

As is demonstrated in section 3.2.2 and 3.2.3, the strengths and weakness of ST are very different from those of GHC-ST, making the two tests natural complements. This motivates the combining of the two tests into a robust omnibus test. Letting $\widehat{\tau}_G$ be the p-value of the GHC-ST test and letting $\widehat{\tau}_S$ be the p-value of the ST test, we define the omnibus test statistic to be $\widehat{\tau}_{OMNI} = \min\{\widehat{\tau}_G, \widehat{\tau}_S\}$. We reject H_0 for small values of $\widehat{\tau}_{OMNI}$.

The dependence between $\widehat{\tau}_G$ and $\widehat{\tau}_S$ is not trivial to characterize, and so the analytic distribution of $\widehat{\tau}_{OMNI}$ under H_0 is difficult to obtain. Instead we opt to approximate its distribution through simulation of the joint null distributions of $\widehat{\tau}_S$ and $\widehat{\tau}_G$ by a perturbation resampling procedure. We perturb the $\widehat{\boldsymbol{\gamma}}$ as described in (3.2.3), and carry out both ST and GHC-ST on $\widehat{\boldsymbol{\gamma}}^{(b)}$ and result in $\widehat{\tau}_S^{(b)}$ and $\widehat{\tau}_G^{(b)}$, $b = 1, ..., B$. The null distribution of $\widehat{\tau}_{OMNI}$ is approximated by $\widehat{\tau}_{OMNI}^{(b)}$, where $\widehat{\tau}_{OMNI}^{(b)} = \min\{\widehat{\tau}_G^{(b)}, \widehat{\tau}_S^{(b)}\}$, $b = 1, ..., B$, and the p-value of $\widehat{\tau}_{OMNI}$ can be estimated as $\frac{1}{B}\sum_{b=1}^{B} I(\widehat{\tau}_{OMNI}^{(b)} \leq \widehat{\tau}_{OMNI})$.

3.3 Simulation Study

We carried out a simulation study to compare the performances of the Wald test, the ST, our proposed GHC-ST method and the Omnibus test, under settings with different sparsity of signals, and correlation structures among predictors. The treatment indicator T is generated from Bernoulli(0.5), and the baseline predictor $\mathbf{X}_{p\times 1}$ was generated from MVN(0, $\mathbf{\Sigma}_{p\times p}$). $\mathbf{\Sigma}_{p\times p}$ takes two types of structure: (i) AR-1 correlation: $\Sigma_{i,j}^{AR-1} = 4\rho^{|i-j|}$; (ii) block exchangeable (assuming the first p_1 predictors are informative):

$$\mathbf{\Sigma}^{Block} = \begin{bmatrix} 4\rho_1 + 4(1-\rho_1)\mathbb{I}_{p_1} & \rho_2 \\ \rho_2 & 4\rho_3 + 4(1-\rho_3)\mathbb{I}_{p-p_1} \end{bmatrix},$$

where ρ_1 measures the correlation among informative predictors, ρ_3 measures the correlation among noise predictors, and ρ_2 measures the correlation of each pair of informative predictor and noise predictor. We studied both continuous and binary outcome, and we considered a total sample size of $n = 1000$. The results are based on 1000 simulations for power calculation, and 5000 simulations for size calculation.

First, the AR-1 correlation structure is adopted. For continuous outcome, Y is generated from the regression model $Y = 2T + h_0(\mathbf{X}) + h_1(\mathbf{X})T + \epsilon$, where ϵ follows a standard normal distribution. We considered the H_0 setting and 7 H_1 settings for the predictor effect functions $h_0(\mathbf{X})$ and $h_1(\mathbf{X})$:

$$\begin{aligned} &\text{Null:} && h_0(\mathbf{X}) = 0; && h_1(\mathbf{X}) = 0 \\ &\text{Alternative:} && h_0(\mathbf{X}) = X_1/20; && h_1(\mathbf{X}) = \frac{1}{\sqrt{p_1}}\sum_{i=1}^{p_1} X_i/10, ; \ p_1 = 1, ..., 7 \end{aligned}$$

For the alternative settings, we fix the total number of predictor to be 35, the overall signal strength (L_2 norm of the coefficient vector), and and examine how the relative power of

different methods vary with diluted interactive signal.

For binary outcome, Y is generated from logistic regression model $\text{logit}\{P(Y = 1)\} = 0.3T + h_0(\mathbf{X}) + h_1(\mathbf{X})T$. Again, we considered the H_0 setting and 7 H_1 settings:

$$\begin{aligned} &\text{Null:} && h_0(\mathbf{X}) = 0; \quad h_1(\mathbf{X}) = 0 \\ &\text{Alternative:} && h_0(\mathbf{X}) = X_1/5;\ h_1(\mathbf{X}) = \frac{1}{\sqrt{p_1}} \sum_{i=1}^{p_1} X_i/5,;\ p_1 = 1, ..., 7 \end{aligned}$$

And we carried out simulations for $\rho = 0.2$ and $\rho = 0.5$.

Secondly, the block exchangeable correlation structure and the binary outcome is adopted. We fix the total number of predictors to be 30, and $h_0(\mathbf{X}) = X_1/2.5$; $h_1(\mathbf{X}) = \frac{1}{\sqrt{2}}(X_1/2.5 + X_2/2.5)$. We carried out four combinations of $\rho_1, \rho_3 = 0$ or 0.4, and varied ρ_2 for non-negative multiples of 0.05 or 0.01 that result in positive definite variance-covariance matrix $\mathbf{\Sigma}^{Block}$.

In table 3.1 and table 3.2, we present sizes for Wald test, ST, GHC-ST and the Omnibus test. Under H_0, the empirical sizes of all procedures are reasonably close to the nominal level of 0.05. We also carried out size simulation with sample size $n = 500$ (results not shown), in which case the ST, GHC-ST and Omnibus test maintain the correct size, but the Wald test has very conservative size due to the poor model fitting. It demonstrated the disadvantage of Wald test when we have a bigger number of predictors.

The power curves under the AR-1 correlation structure is shown in figure 3.1. We observe that the GHC-ST performed better than the ST when the number of informative predictors is small; while when the number of informative predictors increases though keeping the overall signal strength unchanged, the ST outperformed GHC-ST, with a cross-over point depending on ρ. Also notice that after the cross-over point, the power difference between ST and GHC-ST increases as p_1 increases. This is because the ST is more and more powerful with decreasing of sparsity, so the ST which includes information from all

Table 3.1: Sizes for Wald test, ST, GHC-ST and Omnibus test with continuous outcome, under different correlation and total number of predictors

correlation	method	p=5	p=10	p=15	p=20	p=25	p=30	p=35
$\rho = 0.5$	Wald	0.05	0.06	0.05	0.06	0.05	0.05	0.05
	ST	0.05	0.05	0.05	0.05	0.06	0.05	0.05
	GHC-ST	0.05	0.05	0.06	0.05	0.06	0.05	0.05
	Omnibus	0.05	0.05	0.05	0.05	0.06	0.05	0.05
$\rho = 0.2$	Wald	0.05	0.05	0.05	0.05	0.05	0.06	0.06
	ST	0.05	0.05	0.05	0.04	0.05	0.04	0.04
	GHC-ST	0.05	0.05	0.05	0.05	0.05	0.05	0.05
	Omnibus	0.05	0.05	0.06	0.05	0.05	0.05	0.05

Table 3.2: Sizes for Wald test, ST, GHC-ST, and Omnibus test with binary outcome, under different correlation and number of predictors

correlation	method	p=5	p=10	p=15	p=20	p=25	p=30	p=35
$\rho = 0.5$	Wald	0.04	0.04	0.05	0.04	0.04	0.04	0.04
	ST	0.05	0.05	0.05	0.05	0.05	0.05	0.05
	GHC-ST	0.05	0.05	0.05	0.05	0.05	0.05	0.05
	Omnibus	0.05	0.05	0.05	0.05	0.05	0.05	0.05
$\rho = 0.2$	Wald	0.05	0.05	0.05	0.04	0.04	0.04	0.04
	ST	0.05	0.05	0.06	0.04	0.05	0.04	0.05
	GHC-ST	0.05	0.05	0.05	0.05	0.05	0.05	0.05
	Omnibus	0.05	0.05	0.06	0.05	0.05	0.05	0.05

the predictors start to gain power. The omnibus test is able to take advantage of the more powerful test, and the power stays close to the more powerful test regardless of p_1. The Wald test has the lowest power, which might be due to the big number of predictors as well as the high correlation among the predictors.

Figure 3.2 shows under block exchangeable correlation structure, the power changes with ρ_2, with fixed ρ_1 and ρ_3. We observed similar power curve pattern as Figure 3.1. There is a significant power increase with increasing ρ_2 for the ST; while the GHC-ST is relatively robust in terms of correlation between informative and noise predictors. It shows that when there is high correlation between informative predictors and noise predictors, the exchangeable correlation structure allow the informative predictors to share signal with the noise predictors. Since the ST statistic takes into account signal from all the predictors, the power of ST is very sensitive to how significantly the informative predictors contribute to the primary source of variability in the full set of predictors. If the noise drives this variability, as is the case when signals are sparse, ρ_2 is low, and ρ_3 is large, then the power of ST is greatly diminished. In this case, as ρ_2 increases the informative predictors are allowed a greater share in this primary direction of variability, the gains in power are drastic. The power of GHC-ST is robust to these artifacts of the correlation structure because GHC-ST thresholds only the largest components of $\widehat{\gamma}$. This means that even if the noise predictors drive the variability, their corresponding components of $\widehat{\gamma}$ will still be small and so they will be ignored after thresholding.

We expect GHC-ST to be more powerful than ST when signals is sparse. Intuition that the higher criticism is powerful for sparse signals is borne out of the traditional treatments of the higher criticism, which only considered low correlation settings (Donoho and Jin, 2004; Arias-Castro et al., 2011). However, we have observed that in some sparse cases ST outperforms GHC-ST. The reason for this counterintuitive result is due to the strong correlation structures we consider. When the signals are sparse, the noise can inherit a marginal association with outcome if it is strongly correlated with a signal. As a result,

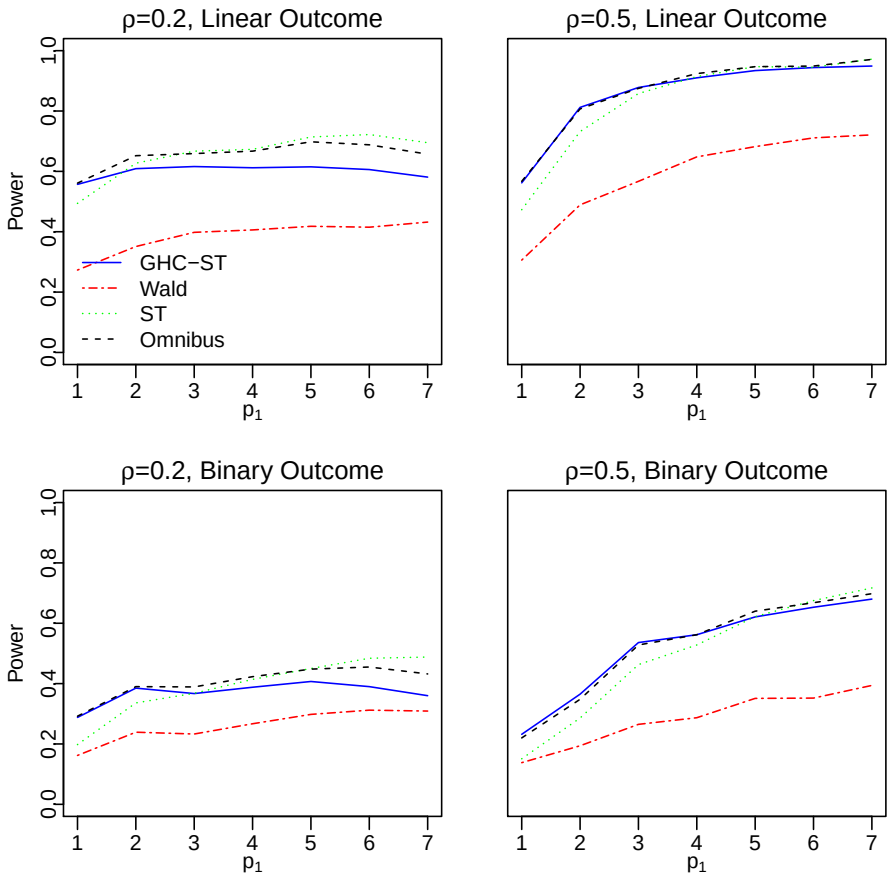

Figure 3.1: Power of four methods, with 30 predictors and fixed overall effect size spread over different numbers of predictors

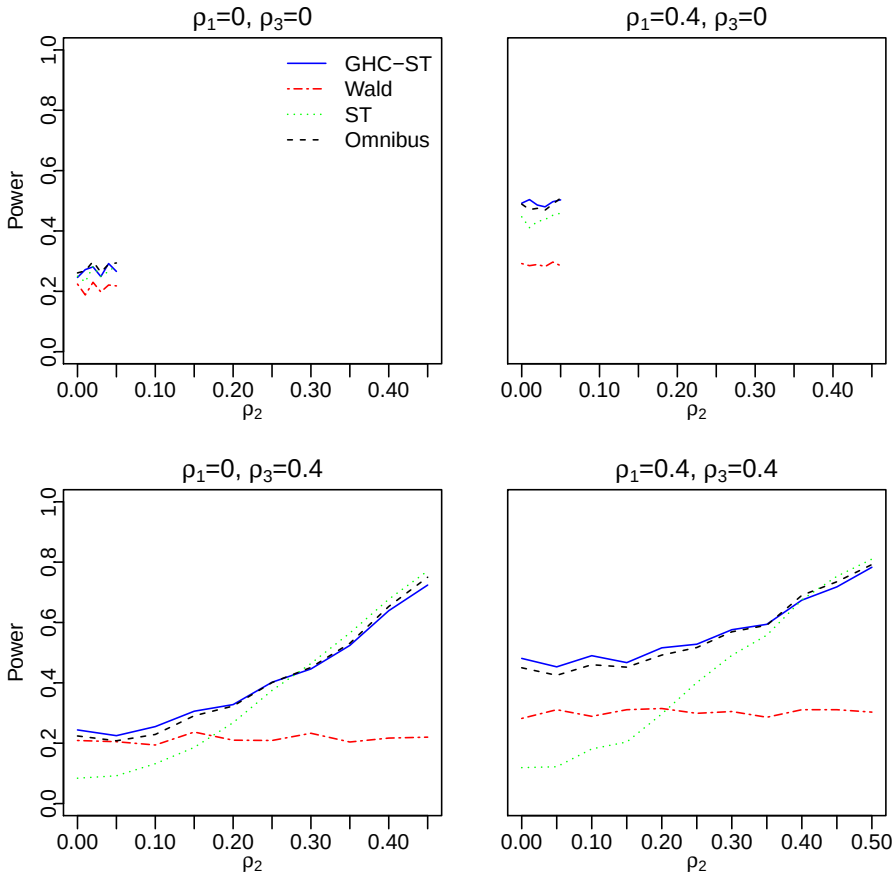

Figure 3.2: Power curve for increase of correlation between informative markers and non-informative markers

a sparse signal can appear dense in the presence of some stronger correlation structures, leading to an improvement in the power of ST, even boosting it over that of GHC-ST in some cases. The power of GHC-ST is higher with $\rho_1 = 0.4$ when compared to $\rho_1 = 0$, which is caused by the sharing of the signals across informative predictors, boosting the marginal test statistics of the informative predictors.

3.4 Data Example: Detecting Informative Baseline Predictors for HIV Treatment Selection

We apply our method to data from AIDS Clinical Trials Group Study 175 (ACTG 175), which is a double-blind study that evaluated treatment with either a single nucleoside or two nucleosides in adults infected with human immunodeficiency virus type 1 (HIV-1) (Hammer et al., 1996). The dataset contains 2139 HIV-infected subjects, where subjects were randomized to four different treatment groups: zidovudine (ZDV) monotherapy, ZDV+didanosine (ddI), ZDV+zalcitabine and ddI monotherapy. Following the primary goal of the original study, we compare ZDV mono therapy ($T = 0$) to combination therapies ($T = 1$), and aim to identify baseline predictors that associated with differential treatment effectiveness.

In the study, the primary endpoint was time to the first $\geq 50\%$ decline in CD4 count, and AIDS-defining event or death. The results from survival analysis on the primary endpoint, as well as other analysis such as adverse events, responses of CD4 Cell counts showed the combination therapies slows the progression of HIV disease and is superior to ZDV mono therapy. There has been subsequent discussion about the efficacy of the combination therapies for HIV (Laskey and Siliciano, 2014; Salam and Pozniak, 2014). With the development of personalized medicine, people focus more on the individualization of different treatments by integrating individual' clinical and genetic information, rather than using the "average" clinical trial result to assign treatment. Thus, from the ACTG175

study data, people are interested in building ITR for HIV based on the baseline information measured in the study. For example, based on the 12 baseline covariates provided in the study, which include five continuous variables: age (years), weight(kg), Karnofsky score(scale of 0-100), CD4 count (cells/mm^3) at baseline and CD8 count (cells/mm^3) at baseline, and seven binary covariates: hemophilia (hemo; 0=no, 1=yes), homosexual activity (homo; 0=no, 1=yes), history of intravenous drug use (intra; 0=no, 1=yes), race (0=white, 1=non-white), gender (0=female, 1=male), antiretroviral history (anti; 0=naive, 1=experienced), and symptomatic status (symp; 0=asymptomatic, 1=symptomatic), Geng et al. (2014); Ma et al. (2015) derived the optimal treatment regime to maximize the mean log survival time and to select important predictors that are needed for deriving the optimal treatment regime. Lu et al. (2011) aimed to find the optimal treatment to maximize the expected CD4 count at 20 ± 5 weeks post-baseline by applying their proposed method to identify variables that are involved in the decision rule as well as building the optimal treatment strategy. Their analysis didn't identify any important baseline covariate that is informative for treatment selection between the ZDV mono therapy and the combination therapies.

We apply our methods to the ACTG 175 data, trying to test whether the 12 baseline covariates as listed above as a group can be used for treatment selection. We can subsequently use the 12 baseline covariates to build ITR if any significant signal is detected. Following Tsiatis et al. (2008), Zhang et al. (2008) and Lu et al. (2013), we chose the CD4 count (cells/mm^3) at 96 ± 5 weeks post-baseline as the primary continuous response Y^{con}. We also defined a responder versus non-responder binary outcome by defining $Y^{bin} = 1$, if the CD4 count at 96 ± 5 weeks is bigger than 500 cells/mm^3 and at least 50 cells/mm^3 increase existing comparing 96 ± 5 weeks to the baseline and $Y^{bin} = 0$ otherwise. We first did Wald test for interaction between each of the 12 variables and the treatment. In particular, in model 3.1, we let the baseline predictor variable $\mathbf{X}$ to contain each of the 12 variables one at a time and carry out the Wald test. The p-values resulted from single

variable test are in table 3.3.

The signal variable effect Wald tests show that only weight and CD8 count at baseline have significant effect on the treatment effect difference, in both continuous outcome case and the binary outcome case. We then applied the Wald test, ST, and GHC-ST to the 12 covariates as a group. When we treated the CD4 count at 96 ± 5 weeks as continuous outcome, the Wald test of 12 covariates as a group resulted in a p-value of 0.55, the ST had a p-value of 0.09, and the GHC-ST had a p-value of 0.01. When we threshold the outcome and use the binary outcome Y^{bin}, the Wald test resulted in a p-value of 0.38, while the ST had a p-value of 0.06 and the GHC-ST had a p-value of 0.05. The results well demonstrated that when we have sparse signals among a bigger number of total predictors, the Wald test has low power in detecting the signal, well the ST and GHC-ST can better detect the signal by either implicitly leverage the effective degree of freedom or by thresholding the signal and excluding the noise. Comparing ST to GHC-ST, the GHC-ST came up with a slightly more significant result, which marginally showed the advantage of applying the GHC procedure in the testing.

Table 3.3: P-values of Wald test for interaction between each baseline covariate and treatment

	age	weight	race	gender	hemo	homo	anti	symp	intra	Karnofsky	CD4	CD8
Y^{con}	0.85	0.02	0.98	0.85	0.87	0.99	0.99	0.88	0.95	0.17	0.63	0.02
Y^{bin}	0.99	0.07	0.42	0.74	0.35	0.48	0.47	0.52	0.35	0.12	0.59	0.01

3.5 Discussion

In this paper, we proposed a scale-independent score test procedure to identify informative baseline predictors for treatment selection. We also incorporated GHC to detect sparse signals. Our method has advantage over the existing Wald test through the model-free construction of the statistic and thresholding signals via the GHC procedure. Our

numerical studies suggest that our proposed method maintains a more accurate size comparing to the Wald test, and also, under situations with sparse signal or strong correlation among predictors, our proposed procedure achieves a much larger power comparing to the Wald test.

References

ANANTH, C. V. and KLEINBAUM, D. G. (1997). Regression models for ordinal responses: a review of methods and applications. *International journal of epidemiology* **26** 1323–1333.

ARCHER, K. and WILLIAMS, A. (2012). L 1 penalized continuation ratio models for ordinal response prediction using high-dimensional datasets. *Statistics in Medicine* **31** 1464–1474.

ARIAS-CASTRO, E., CANDÈS, E. J., PLAN, Y. ET AL. (2011). Global testing under sparse alternatives: Anova, multiple comparisons and the higher criticism. *The Annals of Statistics* **39** 2533–2556.

BARNETT, I. and LIN, X. (2013). Analytic p-value calculation for the higher criticism test in finite p problems. *Manuscript* .

BARNETT, I., MUKHERJEE, R. and LIN, X. (2015). The generalized higher criticism for testing snp-sets in genetic association testing. *Manuscript* .

BILIAS, Y., GU, M., YING, Z. ET AL. (1997). Towards a general asymptotic theory for cox model with staggered entry. *The Annals of Statistics* **25** 662–682.

BISHOP, C. M. ET AL. (2006). *Pattern recognition and machine learning*, vol. 1. springer New York.

BRAUN, M. L. ET AL. (2005). *Spectral properties of the kernel matrix and their relation to kernel*

methods in machine learning. Ph.D. thesis, PhD thesis, Friedrich-Wilhelms-Universität Bonn, 2005.

BREIMAN, L. and SPECTOR, P. (1992). Submodel selection and evaluation in regression: the x-random case. *International Statistical Review/Revue Internationale de Statistique* **60** 291–319.

CAI, T., TIAN, L., SOLOMON, S. D. and WEI, L. (2008). Predicting future responses based on possibly mis-specified working models. *Biometrika* **95** 75–92.

CAI, T., TIAN, L. and WEI, L. (2005). Semiparametric box–cox power transformation models for censored survival observations. *Biometrika* **92** 619–632.

CAI, T., TIAN, L., WONG, P. H. and WEI, L. (2011a). Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics* **12** 270–282.

CAI, T., TONINI, G. and LIN, X. (2011b). Kernel machine approach to testing the significance of multiple genetic markers for risk prediction. *Biometrics* **67** 975–986.

CARDOSO, J. S. and DA COSTA, J. F. P. (2007). Learning to classify ordinal data: The data replication method. *Journal of Machine Learning Research* **8** 6.

CHU, W. and KEERTHI, S. S. (2005). New approaches to support vector ordinal regression. In *Proceedings of the 22nd international conference on Machine learning*. ACM.

CRISTIANINI, N. and SHAWE-TAYLOR, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.

CRISWELL, L. A., PFEIFFER, K. A., LUM, R. F., GONZALES, B., NOVITZKE, J., KERN, M., MOSER, K. L., BEGOVICH, A. B., CARLTON, V. E., LI, W. ET AL. (2005). Analysis of families in the multiple autoimmune disease genetics consortium (madgc) collection: the ptpn22 620w allele associates with multiple autoimmune phenotypes. *The American Journal of Human Genetics* **76** 561–571.

DAVIES, R. B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **64** 247–254.

DONOHO, D. and JIN, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics* 962–994.

DUFFY, M. J. and CROWN, J. (2008). A personalized approach to cancer treatment: how biomarkers can help. *Clinical chemistry* **54** 1770–1779.

FAULKENBERRY, G. D. (1973). A method of obtaining prediction intervals. *Journal of the American Statistical Association* **68** 433–435.

FOSTER, J. C., TAYLOR, J. M. and RUBERG, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in medicine* **30** 2867–2880.

GALIMBERTI, G., SOFFRITTI, G. and DI MASO, M. (2012). Classification trees for ordinal responses in r: The rpartscore package. *Journal of Statistical Software* **47** 1–25.

GENG, Y., ZHANG, H. H. and LU, W. (2014). On optimal treatment regimes selection for mean survival time. *Statistics in medicine* .

GOLUB, G. H., HEATH, M. and WAHBA, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21** 215–223.

GUNTER, L., ZHU, J. and MURPHY, S. (2011). Variable selection for qualitative interactions. *Statistical methodology* **8** 42–55.

HALL, P., JIN, J. ET AL. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *The Annals of Statistics* **38** 1686–1732.

HAMMER, S. M., KATZENSTEIN, D. A., HUGHES, M. D., GUNDACKER, H., SCHOOLEY, R. T., HAUBRICH, R. H., HENRY, W. K., LEDERMAN, M. M., PHAIR, J. P., NIU, M. ET AL. (1996). A trial comparing nucleoside monotherapy with combination therapy

in hiv-infected adults with cd4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine* **335** 1081–1090.

HAREL, M. and SHOENFELD, Y. (2006). Predicting and preventing autoimmunity, myth or reality? *Annals of the New York Academy of Sciences* **1069** 322–345.

HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J. and FRANKLIN, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer* **27** 83–85.

HSU, C.-W. and LIN, C.-J. (2002). A comparison of methods for multiclass support vector machines. *Neural Networks, IEEE Transactions on* **13** 415–425.

IMAI, K., RATKOVIC, M. ET AL. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics* **7** 443–470.

JACOBSON, D. L., GANGE, S. J., ROSE, N. R. and GRAHAM, N. M. (1997). Epidemiology and estimated population burden of selected autoimmune diseases in the united states. *Clinical immunology and immunopathology* **84** 223–243.

JANES, H., PEPE, M. S., BOSSUYT, P. M. and BARLOW, W. E. (2011). Measuring the performance of markers for guiding treatment decisions. *Annals of internal medicine* **154** 253–259.

JESKE, D. R. and HARVALLIE, D. A. (1988). Prediction-interval procedures and (fixed-effects) confidence-interval procedures for mixed linear models. *Communications in Statistics-Theory and Methods* **17** 1053–1087.

KIMELDORF, G. S. and WAHBA, G. (1970). A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics* **41** 495–502.

KOLTCHINSKII, V. and GINÉ, E. (2000). Random matrix approximation of spectra of integral operators. *Bernoulli* 113–167.

KROOT, E.-J. J., DE JONG, B. A., VAN LEEUWEN, M. A., SWINKELS, H., VAN DEN HOOGEN, F. H., VAN'T HOF, M., VAN DE PUTTE, L., VAN RIJSWIJK, M. H., VAN VENROOIJ, W. J. and VAN RIEL, P. L. (2000). The prognostic value of anti–cyclic citrullinated peptide antibody in patients with recent-onset rheumatoid arthritis. *Arthritis & Rheumatism* **43** 1831–1835.

KWEE, L. C., LIU, D., LIN, X., GHOSH, D. and EPSTEIN, M. P. (2008). A powerful and flexible multilocus association test for quantitative traits. *The American Journal of Human Genetics* **82** 386–397.

LA THANGUE, N. B. and KERR, D. J. (2011). Predictive biomarkers: a paradigm shift towards personalized cancer medicine. *Nature reviews Clinical oncology* **8** 587–596.

LASKEY, S. B. and SILICIANO, R. F. (2014). A mechanistic theory to explain the efficacy of antiretroviral therapy. *Nature Reviews Microbiology* .

LAWLESS, J. and FREDETTE, M. (2005). Frequentist prediction intervals and predictive distributions. *Biometrika* **92** 529–542.

LEE, D. and SCHUR, P. (2003). Clinical utility of the anti-ccp assay in patients with rheumatic diseases. *Annals of the rheumatic diseases* **62** 870–874.

LI, B. and LEAL, S. M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics* **83** 311–321.

LIAO, K. P., CAI, T., GAINER, V., GORYACHEV, S., ZENG-TREITLER, Q., RAYCHAUDHURI, S., SZOLOVITS, P., CHURCHILL, S., MURPHY, S., KOHANE, I. ET AL. (2010). Electronic

medical records for discovery research in rheumatoid arthritis. *Arthritis care & research* **62** 1120–1127.

LIAO, K. P., KURREEMAN, F., LI, G., DUCLOS, G., MURPHY, S., GUZMAN, R., CAI, T., GUPTA, N., GAINER, V., SCHUR, P. ET AL. (2013). Associations of autoantibodies, autoimmune risk alleles, and clinical diagnoses from the electronic medical records in rheumatoid arthritis cases and non–rheumatoid arthritis controls. *Arthritis & Rheumatism* **65** 571–581.

LIU, D., LIN, X. and GHOSH, D. (2007). Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. *Biometrics* **63** 1079–1088.

LU, W., ZHANG, H. H. and ZENG, D. (2011). Variable selection for optimal treatment decision. *Statistical methods in medical research* 0962280211428383.

LU, W., ZHANG, H. H. and ZENG, D. (2013). Variable selection for optimal treatment decision. *Statistical methods in medical research* **22** 493–504.

MA, J., HOBBS, B. P. and STINGO, F. C. (2015). Statistical methods for establishing personalized treatment rules in oncology. *BioMed Research International* **2015**.

MALLAL, S., PHILLIPS, E., CAROSI, G., MOLINA, J., WORKMAN, C., TOMAZIC, J., JAGEL-GUEDES, E., RUGINA, S., KOZYREV, O., CID, J. ET AL. (2008). HLA-B* 5701 screening for hypersensitivity to abacavir. *New England Journal of Medicine* **358** 568.

MANOLIO, T. A., COLLINS, F. S., COX, N. J., GOLDSTEIN, D. B., HINDORFF, L. A., HUNTER, D. J., MCCARTHY, M. I., RAMOS, E. M., CARDON, L. R., CHAKRAVARTI, A. ET AL. (2009). Finding the missing heritability of complex diseases. *Nature* **461** 747–753.

IBCSG (2002). (The International Breast Cancer Study Group) endocrine responsiveness and tailoring adjuvant therapy for postmenopausal lymph node negative breast cancer: A randomized trial. *J Natl Cancer Inst* **94** 1054–65.

MIKA, S., SCHÖLKOPF, B., SMOLA, A., MÜLLER, K.-R., SCHOLZ, M. and RÄTSCH, G. (1999). Kernel pca and de-noising in feature spaces. *Advances in neural information processing systems* **11** 536–542.

MINNIER, J. N. (2012). Inference and prediction for high dimensional data via penalized regression and kernel machine methods. *ProQuest Dissertations and Theses* 113Copyright - Copyright ProQuest, UMI Dissertations Publishing 2012.
URL http://search.proquest.com.ezp-prod1.hul.harvard.edu/docview/102776280

NELSON, H., TYNE, K., NAIK, A., BOUGATSOS, C., CHAN, B. and HUMPHREY, L. (2009). Screening for breast cancer: an update for the US Preventive Services Task Force. *Annals of Internal Medicine* **151** 727.

OF SURVIVAL TRIAL INVESTIGATORS, B.-B. E. ET AL. (2001). A trial of the beta-blocker bucindolol in patients with advanced chronic heart failure. *The New England journal of medicine* **344** 1659.

PARK, S. H. (1981). Collinearity and optimal restrictions on regression parameters for estimating responses. *Technometrics* **23** 289–295.

PARK, Y. and WEI, L. (2003). Estimating subject-specific survival functions under the accelerated failure time model. *Biometrika* **90** 717–723.

POLLARD, D. (1990). Empirical processes: theory and applications. Ims.

QIAN, M. and MURPHY, S. A. (2011). Performance guarantees for individualized treatment rules. *Annals of statistics* **39** 1180.

RASMUSSEN, C. E. (2004). Gaussian processes in machine learning. In *Advanced Lectures on Machine Learning*. Springer, 63–71.

ROSENBLUM, M. and VAN DER LAAN, M. J. (2009). Using regression models to analyze randomized trials: Asymptotically valid hypothesis tests despite incorrectly specified models. *Biometrics* **65** 937–945.

ROTNITZKY, A. and ROBINS, J. M. (2005). Inverse probability weighting in survival analysis. *Encyclopedia of Biostatistics* .

SALAM, A. P. and POZNIAK, A. L. (2014). Current antiretroviral therapy. *Pulmonary Complications of HIV* **66** 12.

SCHÖLKOPF, B., MIKA, S., SMOLA, A., RÄTSCH, G. and MÜLLER, K.-R. (1998). Kernel pca pattern reconstruction via approximate pre-images. In *ICANN 98*. Springer, 147–152.

SCHÖLKOPF, B. and SMOLA, A. J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)*. The MIT Press.

SOMERS, E. C., THOMAS, S. L., SMEETH, L. and HALL, A. J. (2006). Autoimmune diseases co-occurring within individuals and within families: a systematic review. *Epidemiology* **17** 202–217.

SONG, X. and PEPE, M. S. (2004). Evaluating markers for selecting a patient's treatment. *Biometrics* **60** 874–883.

SPARANO, J. A. (2006). Tailorx: trial assigning individualized options for treatment (rx). *Clinical breast cancer* **7** 347–350.

STEINWART, I. (2002). On the influence of the kernel on the consistency of support vector machines. *The Journal of Machine Learning Research* **2** 67–93.

SUN, B.-Y., LI, J., WU, D. D., ZHANG, X.-M. and LI, W.-B. (2010). Kernel discriminant learning for ordinal regression. *Knowledge and Data Engineering, IEEE Transactions on* **22** 906–910.

TIAN, L., CAI, T., GOETGHEBEUR, E. and WEI, L. (2007). Model evaluation based on the sampling distribution of estimated absolute prediction error. *Biometrika* **94** 297–311.

TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. and KNIGHT, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67** 91–108.

TSIATIS, A. A., DAVIDIAN, M., ZHANG, M. and LU, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Statistics in medicine* **27** 4658–4677.

WANG, H. and LENG, C. (2007). Unified lasso estimation by least squares approximation. *Journal of the American Statistical Association* **102**.

WANG, H. and LENG, C. (2008). A note on adaptive group lasso. *Computational Statistics & Data Analysis* **52** 5277–5286.

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, Z., SUN, Y., HE, S., CHO, J., ZHAO, H., JIN, J. ET AL. (2014). Detection boundary and higher criticism approach for rare and weak genetic effects. *The Annals of Applied Statistics* **8** 824–851.

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68** 49–67.

ZHANG, B., TSIATIS, A. A., DAVIDIAN, M., ZHANG, M. and LABER, E. (2012). Estimating optimal treatment regimes from a classification perspective. *Stat* **1** 103–114.

ZHANG, M., TSIATIS, A. A. and DAVIDIAN, M. (2008). Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics* **64** 707–715.

ZHAO, L., TIAN, L., CAI, T., CLAGGETT, B. and WEI, L.-J. (2013). Effectively selecting a target population for a future comparative study. *Journal of the American Statistical Association* **108** 527–539.

ZHAO, Y., ZENG, D., RUSH, A. J. and KOSOROK, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association* **107** 1106–1118.

ZUJEWSKI, J. A. and KAMIN, L. (2008). Trial assessing individualized options for treatment for breast cancer: the tailorx trial .