Personalized Plans with Multiple Treatments

Chathura Siriwardhana, Meng Zhao, Somnath Datta and K.B. Kulasekera (kb.kulasekera@louisville.edu)

Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, Kentucky 40202

Abstract

In this work we propose a method for treatment assignment based on individual covariate information for a patient. Our method covers more than two treatments and it can be applied with a broad set of models and it has very desirable large sample properties. An empirical study using simulations and a real data analysis show the applicability of the proposed procedure.

Key Words: Design variables; Personalized Treatments, Single Index Models

1. Introduction

Designing optimal treatment regimes based on individual patient characteristics has gained a momentum over the last few years (see for example van’t Veer and Bernards, 2008; Varquez, 2013). Dynamic treatment regimes that are geared towards the “best” outcome for a patient based on his/her genetic and genomic markers are of high importance. Rather limited literature on this topic mainly deals with deciding between two treatments based on patient characteristics. Assuming without any loss of generality that a larger outcome is better, the methods developed in the literature essentially determine the larger conditional expectation of the outcome given the set of markers for the patient. Cai et al. (2011) use a smoothed sub-group mean in the comparison of two treatments. Here the subgroups are determined via a set of contours (scores) that define overall similarities among patients. For continuous responses, these scores have been defined via linear models. Qian and Murphy (2011) discuss a two step procedure that is based on an estimation of a conditional mean followed by a maximization of that mean over a set of possible treatments. In a different approach for treatment assignments Zhao et al. (2012) consider an optimization technique to select between two treatments
where the binary optimization procedure is within a class of pre-specified model functions. Drawing parallels to the support vector machine technology, these authors show decision optimality of the treatment selection procedure within the binary framework showing that the procedure discussed in Qian and Murphy (2011) is inferior to theirs in the two treatments case. In a more recent article, Zhang et al. (2012) use a robust conditional mean estimation method to alleviate possible wrong model postulation when one estimates the conditional mean for each patient’s profile. Schulte et al. (2014) provide details of using Quality learning (Q-learning) and Advantage learning (A-learning) concepts in devising sequential rules based on a set of pre-specified decision points. The optimality of the decision algorithm, based on the conditional sequential mean, has been discussed by these authors. While mathematically and computationally tedious, it gives a sequential decision rule that self updates the changing patient behavior in switching to a different treatment. Additional references on dynamic treatment regimes can be found in Schulte et al. (2014). Treatment selection based on observational studies has been treated by many authors. Readers are referred to Robins et al. (2004, 2008) and references therein for additional details of such procedures.

In many treatment selection situations clinicians have more than two treatments to select from and the decision of assigning the treatment protocol based on individual patient characteristics is highly desirable. In this work, we discuss the $K$ treatment option ($K \geq 2$) scenario where we compare quantities that are suitable approximations to true conditional probabilities of outcome variable of each treatment dominating other treatments given patient specific scores constructed from covariates. In particular, rather than estimated marginal conditional expectations which can be estimated without sampling from the joint distribution of outcomes for $K$ treatments for an individual, we examine estimated conditional probability of each treatment dominating the others based on $K$ independent pairs of outcomes and covariates, one for each treatment. We choose the optimal treatment as the one that has the highest estimated probability of dominating every one else for a given patient score. In our approach, scores are defined via a set of Single Index Models (SIMs) or Partially Linear Single Index Models (PLSIMs) and our scoring system simplifies to the same type of scoring as in Cai et al. (2011) if $K = 2$. The method we propose is general where the SIMs (PLSIMs) can be quantile regression models rather than mean regression models, thus allowing a broad class of error structures. Empirical evaluations of this new mechanism using a detailed simulation study to assess the accuracy of treatment selection show that the proposed method is comparable with existing methodology in the two treatment option with linear models, it has a higher accuracy in the two treatment case with SIMs and performs very well in the multiple treatment case. Furthermore, we applied our method to an existing dataset with multiple treatment arms to examine the use of treatment assignment based on patient characteristics. The results show that one arm is highly preferred over the others for patients in this study with respect to the primary outcome variable which was a blood count. We also assessed possible gains or losses of patient survival had the patients were assigned according to the rule proposed here. Interestingly our study
reveals that there could have been an advantage in terms of survival also to have used our selection method in the treatment assignment.

The remainder of the article is organized as follows. In Section 2, we discuss the proposed methodology. Section 3 includes simulation results followed by a real data illustration. The main body of the paper ends with a discussion in Section 4. Outlines of all proofs are deferred to an Appendix.

2. Treatment Selection

In this section we describe the proposed procedure and list some of its desirable large sample properties. Let \((Y_i^*, X)\) be the hypothetical (counterfactual) response and covariate pair for treatment \(i\), \(i = 1, \ldots, K\) where larger values of the response are indicative of better outcomes and \(X\) is a vector of \(r\) covariates. Assume further that a patient’s covariate value \(X\) is used to obtain a lower dimensional composite patient score \(U(X)\). In practice one cannot observe the whole vector \((Y_1^*, \ldots, Y_K^*)\) for a single patient. However, using iid observations of type \((\tilde{Y}_i, X_i, A_i)\), \(i = 1, \ldots, n\) where \(A_i\) is the binary treatment indicator for two treatments and \(\tilde{Y}_i\) is the observed response for the \(i\)th patient, previous authors have proposed the estimated difference in conditional means given a score \(U\) to compare two treatments. For example, Zhang et al. (2012) use robust estimators of \(E[Y_1^*|A = 0, U(X)] - E[Y_2^*|A = 1, U(X)]\) where \(U(X) = X\) and \(A = 0, 1\) assign treatments 1 and 2 respectively.

In our approach, we consider pairs of independent observations \((Y_k^*, X_k)\) from the marginal distribution of \((Y_k^*, X)\), \(k = 1, \ldots, K\) to extend the treatment selection for \(K\) treatments using a set of probabilities defined as

\[
p_i(u) = P[Y_i > \max_{i \neq j} Y_j|U(X_k) = u; k = 1, \ldots, K]; i = 1, \ldots, K
\]

for a suitable score defined via a score function \(U\). Note that in (1), the \(Y\)s do not denote the set of true counterfactuals for a patient (given the set of \(X\)) but are independently distributed with the same marginal distributions (given the set of \(X\)). Although the function \(p_i(u)\) does not use the joint distribution of \((Y_1^*, \ldots, Y_K^*, X)\) for a patient with covariate value \(X\), we argue that \(p_i\) above nevertheless gives a measure of dominance for the \(i\)th treatment over the others and hence can be used in selecting the best treatment. This is an alternative to measures based on conditional expectations which require restrictive moment assumptions on the error distribution for all inference aspects in a regression context, the natural framework of handling such data. In our approach, for a given set of functions \(p_1(\cdot), \ldots, p_K(\cdot)\), we define the best treatment for patients with a score \(U_0\) as the treatment given by

\[
k^*(U_0) = \arg \max_{1 \leq i \leq K} \{p_i(U_0)\}.
\]
This procedure can be thought of as maximizing a value function that is the joint conditional expectation of an indicator of one treatment dominating the others given the score rather than evaluating $E[Y|U]$ for each treatment and picking the largest. For example, in Zhao et al (2012), the best treatment was in principle defined as the index corresponding to the larger of $E[Y_1|U]$ and $E[Y_2|U]$ where $Y_1$ and $Y_2$ are the responses for each treatment. In practice, we propose to use estimators of $p_i(U_0)$ based on clinical data and then choose the best treatment as the one that is given by the corresponding estimator of $k^*(U_0)$.

The above approach can be meaningfully used for any set of models that is appropriate for relating responses and covariates provided that those models define an ordering of the above $p_i$s for at least one score so that one of the treatments stands out. If several treatments have the same largest $p_i$ value for a given score, one may pick one of those at random. As shown below, one set of models that can provide such an ordering are Single Index Models (SIMs). In the sequel we base our discussions on Single Index Models relating response $Y_i$ for the $i$th treatment and covariates $X_i$ via

$$Y_i = g_i(\beta_i'X_i) + \epsilon_i$$

for $i = 1, \ldots, K$ where each $\beta_i$ is a $r$-vector of parameters, $g_i$s are unknown link functions for which we assume some reasonable smoothness conditions to hold, and $\epsilon$ is an error term with $E[\epsilon|X] = 0$. This model can also be taken as a quantile regression model with suitable modifications. In methods based on conditional means, one would ideally use $E[Y_i|X]$ to select the best treatment. However, when $X$ has very high dimension, a natural choice is to use a composite score $U(X)$ that has a much smaller dimension. We show in the sequel that if $g_i(\beta_i'X) > g_j(\beta_j'X)$ for all $i \neq j$, then the corresponding $p_i(u) > p_j(u), i \neq j$ for the realization $U = u$ for our proposed score. Hence, using $p_i$s to choose the best treatment is somewhat more general than using conditional expectations. Although the properties of the proposed approach discussed in the sequel are for mean SIMs, they all also hold for quantile SIMs models. Additionally, those properties extend to PLSIMs as the parameters of the linear part of PLSIMs can be estimated at a $\sqrt{n}$ rate (see for example Liang et al., 2010). Hence, our approach is applicable for a very wide class of models. For notational simplicity, we only list properties of the procedure for conditions that are appropriate for mean SIMs. Modifications in these conditions needed for other models are minimal.

Our data are of the following form. Let $Y_{ij}$ indicate the $j$th responses from a group of $n_i$ individuals under treatment $i$ with covariate values $X_{ij}, j = 1, \ldots, n_i$. The sample sizes $n_i$ are assumed to satisfy the condition that $n_i/N$ tends to a positive number where $N = \sum n_i$. Then, for this data, relationship (3) is written as

$$Y_{ij} = g_i(\beta_i'X_{ij}) + \epsilon_{ij}, j = 1, \ldots, n_i.$$  

(4)
Our approach to define an appropriate overall score $U$ is first to use a reasonable model to obtain a treatment specific score for each patient. The score for treatment $i$ measures how favorable it is for a patient to receive this treatment when compared to if he or she were to receive other treatments. To be more specific, we first define

$$S_i (X) = g_i (\beta'_i X) - \max_{j \neq i} \{ g_j (\beta'_j X) \}.$$  

Next, we define the overall score to be the combination of the maximum of these treatment specific scores, and an index that indicates for which treatment the maximum has been achieved for the particular covariate value. That is, we define

$$S (X) = \max_i \{ S_i \}$$

$$\delta (X) = \arg \max_i \{ S_i \} .$$  

(5)

Then, for a patient with covariate value $X$ we define the patient score as $U(X) = (S(X), \delta(X))'$. Note that the score $U(X)$ reduces to the score used in the two treatment case by Cai et al. (2011) if we restrict $g$s to be linear.

In practice one does not know the error distributions and model functions for models defined in (3) and therefore we cannot directly calculate $p_i$s at a given score $u$. Thus, to apply the proposed selection method, we first need to estimate each $p_i$ using a standard function estimation method. This requires observed $Y_{ij}$ values as well as observed $U$ values corresponding to those responses. However, $U$s defined above are hypothetical scores for a covariate value $X$ as we do not know $g$s and $\beta$s. Hence, in estimating $p_i$s, we propose to use “estimated” $U(X_{ij})$ values, $\hat{U}(X_{ij})$, say, corresponding to responses $Y_{ij}, j = 1, ..., n_i; i = 1, ..., K$.

Now, to obtain $\hat{U}(X_{ij})$ values, suitable estimators of link functions $g$s and index vectors $\beta$s can be used to construct estimators $\hat{S}(X)$ and $\hat{\delta}(X)$ of $S(X)$ and $\delta(X)$, respectively. There is a vast literature on estimating the link function and the index vector of a single index model (see, for example, Hristache et al., 2001, Yu and Ruppert, 2002 and references therein) allowing us to use one out of a several available reasonable estimation methods to estimate the $g$s and the $\beta$s. We used the procedure given in Hristache et al. (2001) in our simulations and data analysis in the sequel. In the sequel these estimators will be generically denoted by $\hat{g}_i$ and $\hat{\beta}_i$, respectively, for $i = 1, \ldots, K$.  

5
In particular, for any given vector \( \mathbf{x} \), let
\[
\hat{S}_i(\mathbf{x}) = \hat{g}_i \left( \hat{\beta}_i' \mathbf{x} \right) - \max_{j \neq i} \left\{ \hat{g}_j \left( \hat{\beta}_j' \mathbf{x} \right) \right\}
\]
\[
\hat{S}(\mathbf{x}) = \max_i \left\{ \hat{S}_i(\mathbf{x}) \right\}
\]
\[
\hat{\delta}(\mathbf{x}) = \arg \max_i \left\{ S_i(\mathbf{x}) \right\}
\]
and
\[
\hat{U}(\mathbf{x}) = (\hat{S}(\mathbf{x}), \hat{\delta}(\mathbf{x})).
\]  

(6)

We randomly select an index \( \hat{\delta} \) in the unlikely event that multiple treatments produce the same \( \hat{S} \). Now, we construct our estimator for \( p_i(u), i = 1, ..., K \) at a given \( u = (s, d)' \) as follows. Define
\[
\mathcal{J} = \{(j_1, \ldots, j_K) | j_i \in \{1, \ldots, n_i \}, i = 1, \ldots, K \}.
\]
and, for \( J \in \mathcal{J} \) we let
\[
\hat{w}_J(s) = \prod_{i=1}^{K} \frac{1}{h_i} w \left( \frac{s - \hat{S}(X_{ij_i})}{h_i} \right)
\]
where \( w \) is a kernel function with \( w \geq 0 \) and \( \int w(t) dt = 1 \), and \( h_i \)'s are a set of smoothing parameters. Also, let
\[
\hat{\eta}_J(d) = \prod_{k=1}^{K} I \left[ \hat{g}_d \left( \hat{\beta}_d' X_{kj_k} \right) = \max_m \left\{ \hat{g}_m \left( \hat{\beta}_m' X_{kj_k} \right) \right\} \right] = \prod_{k=1}^{K} I \left( \hat{\delta}(X_{kj_k}) = d \right).
\]

Now, taking an approach similar to the construction of conditional \( U \)-statistics (Stute, 1991), an estimator of \( p_i(u), i = 1, ..., K \) can be defined as
\[
\hat{p}_i(u) = \frac{\sum_{J \in \mathcal{J}} I \left[ Y_{ij_i} > \max_{k \neq i} \{ Y_{kj_k} \} \right] \hat{w}_J(s) \hat{\eta}_J(d)}{\sum \hat{w}_J(s) \hat{\eta}_J(d)}.
\]  

(7)

For a realization \( \mathbf{x}_0 \) of the covariate \( \mathbf{X} \), if we knew the corresponding realization of the score, \( u_0 = (S(\mathbf{x}_0), \delta(\mathbf{x}_0))' \), we can estimate \( p_i(u_0) \) by \( \hat{p}_i(u_0) \). However, due to the aforementioned reasons, we can only find an estimate \( \hat{u}_0 \) of \( u_0 \) using (6) above. Thus, we use \( \hat{p}_i(\hat{u}_0) \) as our estimate of \( p_i(u_0) \) for \( i = 1, ..., K \). Finally, the estimated best treatment for a patient with estimated score \( \hat{u}_0 \) is defined as
\[
\hat{k}^* = \arg \max_{1 \leq i \leq K} \{ \hat{p}_i(\hat{u}_0) \}.
\]  

(8)
Under reasonable conditions stated below, we can show that

\[ \hat{p}_i(\hat{u}_0) \to p_i(u_0) \quad (9) \]

in probability for each \( i \). Hence, if for some \( k^* \), \( p_{k^*}(u_0) > \max_{j \neq k^*} \{p_j(u_0)\} \), then the treatment selection procedure described above is consistent since the best treatment is defined as the treatment corresponding to the largest \( p_i \) and, given the property \( \hat{p}_i(\hat{u}_0) \to p_i(u_0) \), our procedure selects the best treatment with probability tending to one. The ordering of the \( p_i \)s depends on models that relate the responses and the covariates.

Bandwidth selection for estimating the link functions and \( p_i \)s is a challenging issue. Method suggested in Wand and Jones (1995) seemed to perform reasonably well in our simulations and data analysis. However, these choices may not be optimal. We do not investigate the optimal bandwidth selection issue in this work.

2.1 Theoretical Properties

In this section we list a few results that show the consistency of the proposed procedure. We begin by introducing some conditions that are needed to develop these theoretical results. In the sequel we assume that the random variables \((Y_i, X_i), i = 1, ..., K\) are independent and further assume that \( X_i, i = 1, ..., K \) are iid. Let \( F(s, d) \) be the common joint distribution function of \((S(X_1), \delta(X_1))\). We define \( T_i = g_i(\beta'X_1), \quad T = (T_1, \ldots, T_K)' \) and we let \( f_T(t) \) be the joint pdf of \( T \). We need following additional assumptions.

**Assumption 1.** \( F(s, d) \) is absolutely continuous in \( s \) for fixed \( d \) and has a density function \( f(s, d) \), which is bounded.

**Assumption 2.** The kernel function \( w \) is symmetric, has bounded support, Riemann-integrable, nonnegative, bounded away from zero at \( 0 \), and has bounded derivative and finite total variation.

**Assumption 3.** \( f_T \) is continuous.

**Assumption 4.** The errors \( \epsilon_{ij}, \quad j = 1, \ldots, n_i; \quad i = 1, \ldots, K \) are i.i.d with a continuous pdf \( f_\epsilon(\epsilon) \) and \( f_\epsilon(0) > 0 \).

**Remark 1.** All distributional assumptions above are very reasonable and easily satisfied for many error distributions. Assumptions regarding the kernel function \( w \) are standard in nonparametric smoothing literature.

The following lemma shows that the orderings of the \( p_i \)s exist under models specified in (3). The proofs of some of these results use techniques similar to those used in the
proofs of generalized $U$-statistics theory (Stute, 1991). However, since the generalized $U$-statistics theory is not directly applicable here, we give outlines of the proofs in the Appendix.

**Lemma 1.** Under Assumption 1-4 and models (3), for a realization $u = (s, d)'$ of the score $U(X)$ defined above, functions $p_i(u), i = 1, ..., K$ are continuous in $s$ and $p_d(u) > \max_{1 \leq k \leq K; k \neq d} p_k(u)$.

The above lemma shows that under the SIM structure, if there is a link function dominating others at a given covariate value, then there is a corresponding $p$ function that dominates the other $p$ functions over a non trivial set of scores. We now illustrate the consistency of $\hat{p}_i(\hat{u}_0)$ as an estimator for $p_i(u_0)$ at a given score $u_0$.

For our next result which shows that the estimator $\hat{p}(\hat{u}_0)$ converges to $\tilde{p}_i(u_0)$ we need the following assumption. In light of Remark 2 in the Appendix where the proof of Lemma 2 is provided and uniform convergence properties of nonparametric estimators of the link function in Single Index Models (see Wang and Yang (2007) and references therein), we see that this is a reasonable assumption.

**Assumption 5.** For each $i = 1, \ldots, K$, smoothing parameters $h_i \propto N^{-1/5}$ and 

$\sup_{x \in S_X} \left| \hat{g}_i \left( \hat{\beta}_i, x \right) - g_i (\beta_i, x) \right| = O_p \left( N^{-2/5} \log N \right)$.

Now we have the following.

**Theorem 1.** Under Assumptions 1-5, for $u_0$ and $\hat{u}_0$ defined above, we have $\hat{p}(\hat{u}_0) - p(u_0) = o_p(1)$ as $N \to \infty$.

This result shows that the selection of the appropriate treatment is consistent where we define consistency as being able to identify the index associated with the largest $p$ function in (1). In the next section we will provide an empirical assessment of the proposed procedure.

3. Empirical Studies

In this section we present a detailed simulation study that investigates the properties of the proposed procedure in finite samples.

We conducted a series of simulations with the proposed procedure under various settings. Primarily, we focused on the accuracy of treatment assignment of a new (test) observation by using estimated values of the $p_i$ functions from a set of training data. This simulation study was performed for both the two and multiple ($K > 2$) treatment
groups cases. Results for the two groups cases were compared with the corresponding results for existing methods. However, such comparisons were not possible with multiple treatments since there is currently no other method covering more than two treatments. We select our model sets such that each model in a set dominates other competing models for some combination of covariate values; in other words, none of considered models fully dominate other models within the whole covariate space. This signifies, subjects with distinct covariates vectors, could experience corresponding highest response from different treatments illustrating the personalized medicine concept.

In our study, we first simulated \( K \) independent samples with sample size \( n \) (\( n = 50 \) or \( n = 100 \)) per group. The components of the \( r \) dimensional covariate vectors \( \mathbf{X} \) were generated independently from a \( U(-1, 1) \) distribution, where \( r \) ranged from 3 to 8. Using various link functions and index vectors, where a selected few are listed in Tables (1)-(3), we obtained the treatment responses from model (3). Here the errors were generated from \( N(0, \sigma^2) \) and \( DE(0, \sigma) \) where the dispersion parameter \( \sigma \) was chosen from the set \{0.1, 0.2, 0.3, 0.4, 0.5, 1.0\}. We have considered the performance under both linear and nonlinear regression models. We discuss additional details of the structures of these models in the sequel.

Once the \( K \) samples were generated, we estimated the corresponding SIMs followed by an estimation of scores at each covariate value. SIMs were estimated by the procedure given in Hristache et al. (2001) using Epanechnikov kernels (see Polzehl, 2013). Then, a new covariate value \( \mathbf{X}_0 \) was generated in the same manner as previous covariates above, and for its corresponding estimated score \( \hat{u}_0 \), we calculated \( \hat{p}_i(\hat{u}_0) \) for \( i = 1, ..., K \). The kernel function in this estimation was taken to be a \( U(-1, 1) \) probability density function (pdf). We chose the bandwidths by the algorithm given by Wand and Jones (1995) for each \( i, i = 1, ..., K \). We then generated \( K \) new responses, \( Y^*_i \), each with mean \( g_i(\beta'_i \mathbf{X}_0) \) for \( i = 1, ..., K \), corresponding to this \( \mathbf{X}_0 \) using model (3) where the errors were generated independently from the same error distribution that was used to generate the \( K \) original samples. We define the treatment assignment to be correct if

\[
\arg \max \{ \hat{p}_i(\hat{u}_0) \} = \arg \max \{ Y^*_i \}.
\]

We repeated this procedure 1000 times for each model and error distribution combination. The frequency of correct treatment assignment for a selected set of cases are given in the Tables (1)-(3) and a few additional tables are provided in the supplemental materials.

In the analysis of the two groups case (Table 1), we used \( N(0, \sigma^2) \) errors with \( \sigma = 0.1 \) and 0.2. We also compared these results with corresponding results for the two groups assignment methods proposed by Cai et al. (2011), Zhang et al. (2012), and Zhao et al. (2012). We report the number of cases in which their selection (using the highest conditional mean) matched with the group with the largest response. We chose to
compare only with these three methods because these methods highly differ in their approaches and dominate other existing methods in the literature for the two groups case. Here, we highlighted the settings in which other methods underperformed against our method by an asterisk sign. Out of 48 cases the new method competed well with the existing methods in 40 cases. Clearly, the proposed method has a high accuracy in nonlinear models compared to the three existing treatment selection methods. In the case of linear treatment models, which is represented by Model 1 in Table 1, the new method performed comparably to the best method. Model 4 in Table 1 was chosen to demonstrate the robustness of the proposed method, where the requirement of SIM’s is violated. Even in these cases, the accuracy remained fairly high, showing that the proposed method is rather robust.

<table>
<thead>
<tr>
<th>Models(regression function)</th>
<th>Error SD (σ)</th>
<th>Per Group Size (n)</th>
<th>Proposed Method</th>
<th>Cai’s Method</th>
<th>Zhao’s Method</th>
<th>Zhang’s Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) (1.5X₁ - 0.1X₂ + 2X₃ + 2X₄ - 1.5X₅ - 1.6X₆)/√73.37; Group 1 (2X₁ + 1.6X₂ + 2.2X₃ + 3.5X₄ + 1.2X₅ + 1.5X₆)/√27.34; Group 2</td>
<td>0.1</td>
<td>50</td>
<td>864</td>
<td>889</td>
<td>889</td>
<td>882*</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>50</td>
<td>794</td>
<td>817</td>
<td>775*</td>
<td>801</td>
</tr>
<tr>
<td>(2) sin{π(X₁ - 0.3X₂ - X₃)/√700}; Group 1</td>
<td>0.1</td>
<td>50</td>
<td>900</td>
<td>683*</td>
<td>670*</td>
<td>683*</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>50</td>
<td>726</td>
<td>724*</td>
<td>728*</td>
<td></td>
</tr>
<tr>
<td>(3) sin{π/4 + π/π(X₁ + X₂ + X₃)/√5.66}; Group 1 sin{π/2 + π/1.8X₁ - 1.3X₂ + 0.8X₃ + X₄ - 1.2X₅ - X₆)/√507}; Group 2</td>
<td>0.1</td>
<td>50</td>
<td>850</td>
<td>708*</td>
<td>693*</td>
<td>718*</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>50</td>
<td>850</td>
<td>708*</td>
<td>693*</td>
<td>718*</td>
</tr>
<tr>
<td>(4) sin{π(X₁ + X₂ + X₃)/√3}; Group 1 sin{π/2 + π/π(X₁ + X₂ + X₃)/√7}; Group 2</td>
<td>0.1</td>
<td>50</td>
<td>920</td>
<td>743*</td>
<td>694*</td>
<td>804*</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>50</td>
<td>900</td>
<td>741*</td>
<td>703*</td>
<td>796*</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>100</td>
<td>935</td>
<td>704*</td>
<td>764*</td>
<td>842*</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>100</td>
<td>924</td>
<td>774*</td>
<td>749*</td>
<td>834*</td>
</tr>
</tbody>
</table>

Table 1: Frequencies of correct treatment assignments in 1000 test cases by four competing algorithms in the two groups case. The regression models used in the simulations include linear and nonlinear SIM models, as well as models that are not SIM. Cases where the proposed method (ours) outperformed a competing method is denoted by *.

We studied the multiple treatment groups case for \( K = 3 \) and 4, using a variety of models generated from several nonlinear model families. All considered cases produced results that are generally anticipated in a study of this nature. Cases involving highly nonlinear curves with minor differences in the mean value function performed somewhat poor compared with cases where the nonlinearity is less severe or the differences between the signals is higher. Our discussion in the sequel focuses on two families

\[
Y_i = g_i\left\{ \pi k_i + \pi(\beta'X) \right\} + \varepsilon, \ i = 1, \ldots, K, \quad \text{(Type I)},
\]

and

\[
Y_i = g_i\left\{ \pi k_i + \pi(\beta'_iX) \right\} + \varepsilon, \ i = 1, \ldots, K, \quad \text{(Type II)}.
\]
In each type above, \( g_i \) is either a \( sine \) or a \( cosine \) function. In Type I models, the same single index vector \( \beta \) has been used for the treatment groups where the \( g_i \) function varies across the groups. In our simulations we chose this common vector to be \( C' = (1/\sqrt{r},...,1/\sqrt{r})_{1 \times r} \) (Table 2). In Type II models we used a variety of \( \beta_i \) index vectors whose components were selected in an arbitrary fashion. These components are given in Table (4). For example, in the three treatment case with \( r = 3 \), \( \beta_1 = (1.5, 1.6, 0.9)' \), \( \beta_2 = (0.8, 0.6, 0.7)' \), and \( \beta_3 = (1.8, 2.1, 0.8)' \).

If several models are close to each other within the whole covariate domain, a high classification error (i.e., incorrect treatment assignment) can be expected due to the lack of functional separations. In general, the functional behavior of a multi covariate nonlinear model cannot be easily visualized. Type-I models used here have relatively substantial functional differences compared to some Type II models for each \( K \). Tables [2] and [3] show the correct assignment frequencies for a representative set of multi-groups cases. Again, the results for all examined cases were very similar to the few presented here.

Examination of the results reveal high assignment accuracy for large sample sizes and low error variability. In general, we observed fairly high accuracies for low covariate dimensions. The presented simulation results are based on \( sine \) and \( cosine \) functions which are bounded in \((-1,1)\). Hence, an increment in \( \sigma \) by 0.1 adds a relatively large noise to a model. Consequently, as expected, we observed a decline in the correct assignment frequency as \( \sigma \) is increased. The results for the three groups case for both Type I and II models are somewhat comparable whereas the results for Type II models for four groups case were lower compared to those corresponding to Type I models. As indicated in the previous paragraph, we believe these differences are due to relative lack of separation in the model functions.

4. ACTG-175 HIV Clinical Trial

In this section we illustrate our proposed method using a real clinical trial dataset.

The data resulted from the ACTG 175 clinical trial (Hammer et al. 1996). This trial was a randomized, double-blinded, placebo-controlled clinical trial that was conducted for comparing antiviral medications for HIV-1 patients whose T-cell CD4 counts were in the range of 200 to 500 per cubic millimeter. The dataset (Juraska et al. 2012) contains information on 2136 HIV-1 infected individuals who were randomized into four treatment arms; those treated with Zidovudine (arm-0), combination of Zidovudine and Didanosine (arm-1), combination of Zidovudine and Zalcitabine (arm-2), Didanosine (arm-3). Arms 0, 1, 2, and 3 contain 532, 519, 524, and 561 patients, respectively. The
Table 2: Frequencies of correct treatment assignments in 1000 test cases by the proposed method in multiple groups case \((K > 2)\), using Type I nonlinear regression models, with \(C' = \left(1/\sqrt{r}, \ldots, 1/\sqrt{r}\right)_{1 \times r}\).

<table>
<thead>
<tr>
<th>Number of Groups</th>
<th>Models (regression function)</th>
<th>Dimension (r)</th>
<th>Per Group Size</th>
<th>Normal Error SD ((\sigma))</th>
<th>DE Error SD ((\sigma))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Three</td>
<td>(\sin \left(\frac{\pi}{C'} X\right)): Group 1</td>
<td>3</td>
<td>50</td>
<td>0.1, 0.3, 0.5</td>
<td>0.1, 0.3, 0.5</td>
</tr>
<tr>
<td></td>
<td>(\cos \left(\frac{\pi}{C'} X\right)): Group 2</td>
<td>5</td>
<td>50, 100</td>
<td>0.1, 0.3, 0.5</td>
<td>0.1, 0.3, 0.5</td>
</tr>
<tr>
<td></td>
<td>(\sin \left(\frac{\pi}{C'} X\right)): Group 3</td>
<td>8</td>
<td>50, 100</td>
<td>0.1, 0.3, 0.5</td>
<td>0.1, 0.3, 0.5</td>
</tr>
<tr>
<td>Four</td>
<td>(\sin \left(\frac{\pi}{C'} X\right)): Group 1</td>
<td>3</td>
<td>50</td>
<td>0.1, 0.3, 0.5</td>
<td>0.1, 0.3, 0.5</td>
</tr>
<tr>
<td></td>
<td>(\cos \left(\frac{\pi}{C'} X\right)): Group 2</td>
<td>5</td>
<td>50, 100</td>
<td>0.1, 0.3, 0.5</td>
<td>0.1, 0.3, 0.5</td>
</tr>
<tr>
<td></td>
<td>(\sin \left(\frac{\pi}{C'} X\right)): Group 3</td>
<td>8</td>
<td>50, 100</td>
<td>0.1, 0.3, 0.5</td>
<td>0.1, 0.3, 0.5</td>
</tr>
</tbody>
</table>

Table 3: Frequencies of correct treatment assignments in 1000 test cases by the proposed method in multiple groups case \((K > 2)\), using Type II nonlinear regression models. Selected \(\beta\) vectors are shown in Table (4).

<table>
<thead>
<tr>
<th>Number of Groups</th>
<th>Models (regression function)</th>
<th>Dimension (r)</th>
<th>Per Group Size</th>
<th>Normal Error SD ((\sigma))</th>
<th>DE Error SD ((\sigma))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Three</td>
<td>(\sin \left(\frac{\pi}{\tau} (\beta_1 X)\right)): Group 1</td>
<td>3</td>
<td>50, 100</td>
<td>0.1, 0.3, 0.5</td>
<td>0.1, 0.3, 0.5</td>
</tr>
<tr>
<td></td>
<td>(\cos \left(\frac{\pi}{\tau} (\beta_2 X)\right)): Group 2</td>
<td>5</td>
<td>50, 100</td>
<td>0.1, 0.3, 0.5</td>
<td>0.1, 0.3, 0.5</td>
</tr>
<tr>
<td></td>
<td>(\sin \left(\frac{\pi}{\tau} (\beta_3 X)\right)): Group 3</td>
<td>8</td>
<td>50, 100</td>
<td>0.1, 0.3, 0.5</td>
<td>0.1, 0.3, 0.5</td>
</tr>
<tr>
<td>Four</td>
<td>(\sin \left(\frac{\pi}{\tau} (\beta_1 X)\right)): Group 1</td>
<td>3</td>
<td>50, 100</td>
<td>0.1, 0.3, 0.5</td>
<td>0.1, 0.3, 0.5</td>
</tr>
<tr>
<td></td>
<td>(\cos \left(\frac{\pi}{\tau} (\beta_2 X)\right)): Group 2</td>
<td>5</td>
<td>50, 100</td>
<td>0.1, 0.3, 0.5</td>
<td>0.1, 0.3, 0.5</td>
</tr>
<tr>
<td></td>
<td>(\sin \left(\frac{\pi}{\tau} (\beta_3 X)\right)): Group 3</td>
<td>8</td>
<td>50, 100</td>
<td>0.1, 0.3, 0.5</td>
<td>0.1, 0.3, 0.5</td>
</tr>
</tbody>
</table>

severity of HIV progression is measured through a decline in CD4 counts. This trial periodically measured a patient’s CD4 count as the clinical outcome. In our analysis, we considered the log transformed CD4 count of a patient after 20 weeks of treatment as the clinical response. As covariates, we used the log-CD4 count at baseline, age, weight, and the number of months a patient received the pre-antiviral therapy.

We applied the proposed treatment assignment strategy to the data from all four arms of the study. In addition, to compare with several existing two-group methods, we provide another illustration. In each situation, we randomly selected 200 patients from each arm as “training” data to estimate the SIMs. Remaining patients were considered as new (test) patients. After fitting SIMs to training data we estimated the scores for the test cases and estimated the corresponding \(p_i\) functions at those scores to assign each
<table>
<thead>
<tr>
<th>Treatment groups</th>
<th>Group</th>
<th>Number of covariates</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>$\beta_3$</th>
<th>$\beta_4$</th>
<th>$\beta_5$</th>
<th>$\beta_6$</th>
<th>$\beta_7$</th>
<th>$\beta_8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Three</td>
<td>1</td>
<td>3</td>
<td>1.5</td>
<td>1.6</td>
<td>0.9</td>
<td>1.2</td>
<td>1.4</td>
<td>-1.5</td>
<td>1.2</td>
<td>1.6</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3</td>
<td>1.0</td>
<td>1.4</td>
<td>0.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>5</td>
<td>1.0</td>
<td>1.4</td>
<td>0.8</td>
<td>0.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>8</td>
<td>1.0</td>
<td>1.4</td>
<td>0.8</td>
<td>0.6</td>
<td>-1.1</td>
<td>0.8</td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td></td>
<td>1.3</td>
<td>1.7</td>
<td>0.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td>1.3</td>
<td>1.7</td>
<td>0.7</td>
<td>0.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td></td>
<td>1.3</td>
<td>1.7</td>
<td>0.7</td>
<td>0.9</td>
<td>1.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td></td>
<td>1.3</td>
<td>1.7</td>
<td>0.7</td>
<td>0.9</td>
<td>1.1</td>
<td>-1.3</td>
<td>-0.1</td>
<td>0.9</td>
</tr>
<tr>
<td>Four</td>
<td>1</td>
<td>3</td>
<td>0.8</td>
<td>0.6</td>
<td>0.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3</td>
<td>1.2</td>
<td>1.4</td>
<td>0.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>5</td>
<td>1.2</td>
<td>1.4</td>
<td>0.9</td>
<td>1.5</td>
<td>0.9</td>
<td>1.1</td>
<td>1.4</td>
<td>1.2</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>8</td>
<td>1.2</td>
<td>1.4</td>
<td>0.9</td>
<td>1.5</td>
<td>0.9</td>
<td>1.1</td>
<td>1.4</td>
<td>1.2</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td></td>
<td>0.2</td>
<td>0.3</td>
<td>0.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td>0.2</td>
<td>0.3</td>
<td>0.8</td>
<td>0.6</td>
<td>0.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td></td>
<td>0.2</td>
<td>0.3</td>
<td>0.8</td>
<td>0.6</td>
<td>0.3</td>
<td>0.1</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td></td>
<td>0.2</td>
<td>0.3</td>
<td>0.8</td>
<td>0.6</td>
<td>0.3</td>
<td>0.1</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td>1.8</td>
<td>2.1</td>
<td>0.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td></td>
<td>1.8</td>
<td>2.1</td>
<td>0.8</td>
<td>0.7</td>
<td>0.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td></td>
<td>1.8</td>
<td>2.1</td>
<td>0.8</td>
<td>0.7</td>
<td>0.9</td>
<td>1.3</td>
<td>1.1</td>
<td>1.3</td>
</tr>
</tbody>
</table>

Table 4: $\beta$ vectors of Type II models, for model dimensions ($r$) 3, 5, and 8.

test patient to the best treatment group suggested by the largest estimated $p_i$ value, as dictated by our method.

The calculation of the estimator (7) is computationally tedious for larger sample sizes ($>100$), especially, in multiple treatment cases ($K > 2$). To ease the computational burden, an ad hoc approximation was used for the four treatment case. For each training set, we picked a random location in the list of patients. Then, we selected five consecutive individuals starting from that position from each group and estimate $p_k(\hat{u})$, $k = 1, ..., K$, at a given $u$ based on these $5 \times K$ observations (i.e., 5 per group). This procedure was repeated 1000 times. We define $\hat{p}_{k,m}(\hat{u})$ to be the estimator (7) based on these selected observations for the $m$th set of subsamples. In the estimation of $\hat{p}_{k,m}(\hat{u})$ functions, we use only $5 \times K$ observations for each $m$. Hence, to avoid the possibility of getting zero denominators we used Gaussian kernels, instead of uniform kernels. We obtained an overall approximation $\hat{p}_k^*(\hat{u})$ of $\hat{p}_k(\hat{u})$ as,

$$
\hat{p}_k^*(\hat{u}) = \frac{1}{1000} \sum_{m=1}^{1000} \hat{p}_{k,m}(\hat{u}).
$$

To investigate the appropriateness of such an ad hoc subsampling scheme, we performed several additional simulations using both the original full sample estimation method and
the above ad hoc method with several subsample sizes (details not reported). We noted that the frequencies of accurate assignment are closely matched for the two methods, suggesting this type of an ad hoc method is a reasonable approach in practice.

We report the results for the two group comparisons first. When we used the proposed method, out of 651 test patients, only 23 were assigned to arm 0, suggesting that possibly a large number of patients would have experienced a more favorable outcome from arm 1. We also applied the two-group assignment methods proposed by Cai et al. (2011), Zhao et al. (2012), and Zhang et al. (2012), for the same training and test data. These methods also assigned lesser number of patients to arm 0, than the actual assignment by the randomized trial. We present these results in Table (5). For example, in Table (5), the (0,0) cell for the Proposed Method indicates that only 10 out of 332 patients who were actually treated in arm-0 would have been assigned to arm 0 had we used the proposed method.

<table>
<thead>
<tr>
<th>Original Assignment</th>
<th>New Assignment</th>
<th>Proposed Method</th>
<th>Cai’s Method</th>
<th>Zhao’s Method</th>
<th>Zhang’s Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>322</td>
<td>7</td>
<td>325</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 5: Two groups treatment assignment summary for ACTG-175 trial, by four methods.

In the multiple treatments assignment setting, we have a total of 1336 patients in the test set. Among them, we assigned the majority: 895 to arm-1 whereas 182 and 259 patients are assigned to arms 2 and 3, respectively. Notably, the new method did not assign any patients to arm-0, seemingly suggesting that one of the other arms dominate arm 0 with respect to our scoring mechanism. These results are summarized in Table (6). We noticed that, a large number of patients (1014) are proposed to be assigned to a different treatment arm than their actual assignment. Based on these allocations, it appears that the majority in the study would have benefited from arm 1.

4.1 Examination of the survival aspect

The proposed treatment selection method is an attempt to assign patients to receive the optimal outcome based on their score. Given that the above analysis shows that the optimal assignments based on patient characteristics are different from actual assignments towards a higher CD4 count, we might also be able to argue that such an assignment rule should also improve the expected value of the related survival time conditional on
Table 6: Four groups treatment assignment summary for ACTG-175 clinical trial, by the proposed method.

<table>
<thead>
<tr>
<th>Original Assignment</th>
<th>Proposed Assignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td></td>
</tr>
</tbody>
</table>

In the dataset, there are three types of events: (i) when an individual’s CD4 count drops less than 50% of his/her pretreatment count, (ii) an event indicating progression to AIDS, (iii) death. Thus, the term “survival time” would denote an event time in the above sense. In addition, there was right censoring present in the data. Now, consider the $i$th subject in the test set with covariate value $X_i$ who is assigned to a particular arm by an assignment mechanism. Suppose the individuals estimated score is $\hat{u}_i = \{\hat{S}(X_i), \hat{\delta}(X_i)\}$. Let $k^*_i$ be the group the procedure would assign this patient based on his/her estimated score $\hat{u}_i$ and let $k_i$ be the treatment group he was assigned in the original trial. Conditional on $\hat{u}_i$, we estimated the difference in the survival times in the two groups, as

$$\Delta_i = E(t_{k^*_i|\hat{u}_i}) - E(t_{k_i|\hat{u}_i}).$$

For a fixed $k$, we consider a symmetric neighborhood of width $2h$ centered around $\hat{S}(X_i)$,

$$N_h = \{\hat{S}(X_i) - h, \hat{S}(X_i) + h\},$$

where $h$ was the bandwidth chosen by the procedure given in Wand and Jones (1995) for scores for all patients. Next, we selected a subgroup of patients from the whole set (training and test), whose covariate values $X$ satisfy (i) patient was originally treated in arm $k$ and (ii) $\hat{S}(X) \in N_h$ and (iii) the score satisfies $\hat{\delta}(X) = \hat{\delta}(X_i)$. Suppose the size of the above subgroup is $d$. If $d < 10$, we increased the width of the neighborhood $N_h$ in multiples of $h$ (i.e., $3h, 4h$ etc.) and recalculated $d$ till it is at least 10. After that the Kaplan Meier estimator was calculated using the survival times of those individuals in $N_h$.

Our estimator of the expected survival time for each group, i.e., $E(t_{k_i|\hat{u}_i})$, $k_i = 1, ..., K$, was the area covered under the corresponding Kaplan-Meier curve. For a given $\hat{u}_i$, we then find the estimated survival gain $\hat{\Delta}_i$ from the proposed selection as the difference between the two estimated expectations, $\hat{E}(t_{k^*_i|\hat{u}_i}) - \hat{E}(t_{k_i|\hat{u}_i})$. Finally we estimate the
overall treatment selection efficiency as the averaged $\hat{\Delta}_i$s for all test patients,

$$\rho = \frac{1}{N} \sum_{i=1}^{N} \hat{\Delta}_i,$$

(10)

where $N$ is the number in the test set. Note that a positive value for $\rho$ indicates an overall effective treatment selection. Table (7) gives these $\rho$ values for the proposed procedure with two and multiple treatments cases along with the resulting estimated survival gains for methods proposed by Cai et al. (2011), Zhao et al. (2012), and Zhang et al. (2012), for the two-groups application.

<table>
<thead>
<tr>
<th>Two Groups Assignments</th>
<th>Four Groups Assignments by Proposed Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed Method</td>
<td>Cai's Method</td>
</tr>
<tr>
<td>$\rho$</td>
<td>63.7</td>
</tr>
<tr>
<td>$\rho_m$</td>
<td>57.4</td>
</tr>
</tbody>
</table>

Table 7: Observed $\rho$ and $\rho_m$ by four treatment selection methods, under different treatment possibilities.

Additionally, we consider the marginal survival functions and define,

$$\Delta'_i = E(t_{k^*_i}) - E(t_{k_i}),$$

where $E(t_{k^*_i})$ and $E(t_{k_i})$ are corresponding marginal expected survival times of new ($k^*_i$) and actual ($k_i$) arms. Again using the area under the marginal Kaplan Meier estimates, we calculate estimated values of $\Delta'_i$, $i = 1, ..., N$. Similar to (10), we obtain $\rho_m$ using these marginal estimates. Corresponding $\rho_m$'s are also reported in Table (7). Since the proposed treatment selection is based on a scoring scheme, we argue that examining the score dependent survival outcome would be a more reliable approach. This is confirmed by the fact $\rho_m \leq \rho$ in all cases.

5. Discussion

In this article we proposed a novel personalized treatment plan to select the optimal treatment from a set of multiple treatments. This method is a single step procedure which can be easily applied. The proposed method is based on semi parametric Single Index Models which, add great flexibility in modeling real life situations. Furthermore, this method can also be used for quantile regression SIMs providing additional model flexibility compared with existing methods based on conditional expectations. Our empirical studies show that the proposed method performs very satisfactorily in selecting
the optimal treatment in a multiple treatment setting while outperforming existing methods for the two treatment case for non-linear models which are more realistic in practical situations. In addition, as our simulations showed, the method is rather robust against departures from SIMs. We show that the proposed method has desirable theoretical properties. Our analysis of a real clinical trials dataset which has the multiple treatment option reveals a possible connection between optimal treatment selection and a gain in patient survival.

This article deals with complete responses. However, censoring is very common in practice. An extension of the proposed methodology to a covariate dependent censoring setting and various lifetime aspects such as multi state models is forthcoming. Our study is addressing the optimal treatment selection based on a single response. However, there are numerous circumstances where the optimality is desired with respect to multiple criteria. For example, a treatment may have to be selected to maximize the survival rates but minimize after effects and maximize the quality of life in terms of temporary side effects. In such cases we have a multi criteria optimization problem. This opens up another interesting future research avenue.
Bibliography


Appendix

In this Appendix we provide outlines of the proofs of the technical results.

Proof. Proof of Lemma 1

First, we prove (ii), the continuity of $p_i(u)$ in $s$ under regularity conditions. Let

$$f_\epsilon (\epsilon) = \prod_{i=1}^{K} f_\epsilon (\epsilon_i),$$

where $\epsilon = (\epsilon_1, \ldots, \epsilon_K)'$, and $\epsilon = (\epsilon_1, \ldots, \epsilon_K)'$. For a $K$-vector $a = (a_1, \ldots, a_K)'$ and a subset $I$ of $\{1, \ldots, K\}$, let $a_{\{I\}}$ denote the $(K - ||I||)$-vector obtained from $a$ by removing the $i$th element from $a$ for an $i \in I$, and, let $a_{\{I\}}$ denote the $||I||$-vector consisting of $a_i$ for $i \in I$. Without loss of generality, assume $d = 1$. Let $t_i = (t_{i1}, \ldots, t_{iK})'$ be vectors in $\mathbb{R}^K$ and $t_j (s) = (t_{j1}(s), t_{j2}(s), \ldots, t_{jK}(s))'$ where

$$t_{ij}(s) = \begin{cases} \max_{k>1} t_{ik} + s, & \text{if } j = 1, \\ t_{ij}, & \text{if } j \neq 1. \end{cases}$$

It can be shown that

$$p_i(u) = \frac{\int \cdots \int_{A_i^{(s)}} f_\epsilon (\epsilon) \prod_{j=1}^{K} f_T (t_j (s)) d\epsilon dt_1(1) \cdots dt_K(1)}{\int \cdots \int \prod_{j=1}^{K} f_T (t_j (s)) dt_1(1) \cdots dt_K(1)}$$

(12)

where $\epsilon = (\epsilon_1, \ldots, \epsilon_K)'$, $\epsilon = (\epsilon_1, \ldots, \epsilon_K)'$, and $A_i^{(s)} = \{ \epsilon \in \mathbb{R}^K | \epsilon_i + t_{ii} (s) > \max_{k \neq 1, i} \{ \epsilon_k + t_{kk} (s) \} \}$. Let $s' = s + \Delta s$, $u' = (s', 1)$ and $e_i = (0, \ldots, 1, \ldots, 0)$. We then have

$$\int \cdots \int_{A_i^{(s')}} f_\epsilon (\epsilon) \prod_{j=1}^{K} f_T (t_j (s')) d\epsilon dt_1(1) \cdots dt_K(1) - \int \cdots \int_{A_i^{(s)}} f_\epsilon (\epsilon) \prod_{j=1}^{K} f_T (t_j (s)) d\epsilon dt_1(1) \cdots dt_K(1)$$

$$= \int \cdots \int_{A_i^{(s)}} f_\epsilon (\epsilon - \Delta se_1) \prod_{j=1}^{K} f_T (t_j (s) + \Delta se_1) d\epsilon dt_1(1) \cdots dt_K(1)$$

$$= \int \cdots \int_{A_i^{(s)}} \left( f_\epsilon (\epsilon - \Delta se_1) \prod_{j=1}^{K} f_T (t_j (s) + \Delta se_1) - f_\epsilon (\epsilon) \prod_{j=1}^{K} f_T (t_j (s)) \right) d\epsilon dt_1(1) \cdots dt_K(1),$$

(13)
and
\[
\int \ldots \int \prod_{j=1}^{K} f_T(t_j(s')) \, dt_{1(1)} \ldots dt_{K(1)} - \int \ldots \int \prod_{j=1}^{K} f_T(t_j(s)) \, dt_{1(1)} \ldots dt_{K(1)}
\]
\[
= \int \ldots \int \left( \prod_{j=1}^{K} f_T(t_j(s) + \Delta s e_1) - \prod_{j=1}^{K} f_T(t_j(s)) \right) \, dt_{1(1)} \ldots dt_{K(1)}
\]

(14)

By Assumption 3, we have
\[
\lim_{\Delta s \to 0} p_i(u') - p_i(u) = 0
\]
proving the continuity of \( p_i \).

Now we show that \( p_1(u) > p_k(u) \) for any \( k > 1 \). Since the denominator of the right hand side of (12) is not affected by \( i \), we only need to show the inequality for the numerator. In the following discussion, we use the assumption that \( \epsilon_1, \ldots, \epsilon_K \) are iid random variables with the common pdf \( f_\epsilon \). Consider we have
\[
I_1 = \int \ldots \int A_1^{[1]} \, f_\epsilon(\varepsilon) \prod_{j=1}^{K} f_T(t_j(s)) \, d\varepsilon \, dt_{1(1)} \ldots dt_{K(1)}
\]
\[
= \int \ldots \int P \left( \epsilon_1 + t_{11}(s) \geq \max_{j>1} \{ \epsilon_j + t_{jj}(s) \} \right) \prod_{j=1}^{K} f_T(t_j(s)) \, d\varepsilon \, dt_{1(1)} \ldots dt_{K(1)}
\]

Now, we have
\[
P \left( \epsilon_1 + t_{11}(s) \geq \max_{j>1} \{ \epsilon_j + t_{jj}(s) \} \right)
\]
\[
= P \left( \epsilon_1 \geq \epsilon_k + t_{kk}(s) - t_{11}(s) \text{ and } \epsilon_1 \geq \max_{j>1, j \neq k} \{ \epsilon_j + t_{jj}(s) - t_{11}(s) \} \right)
\]
\[
= P \left( \epsilon_k \geq \epsilon_1 + t_{kk}(s) - t_{11}(s) \text{ and } \epsilon_k \geq \max_{j>1, j \neq k} \{ \epsilon_j + t_{jj}(s) - t_{11}(s) \} \right)
\]
\[
\geq P \left( \epsilon_k \geq \epsilon_1 + t_{k1}(s) - t_{1k}(s) \text{ and } \epsilon_k \geq \max_{j>1, j \neq k} \{ \epsilon_j + t_{jj}(s) - t_{1k}(s) \} \right)
\]
\[
= P \left( \epsilon_k \geq \epsilon_1 + t_{11}'(s) - t_{kk}(s) \text{ and } \epsilon_k \geq \max_{j>1, j \neq k} \{ \epsilon_j + t_{jj}'(s) - t_{kk}'(s) \} \right)
\]
\[
= P \left( \epsilon_k + t_{kk}'(s) \geq \max_{j \neq k} \{ \epsilon_j + t_{jj}'(s) \} \right)
\]
where \( t'_{ij} = t_{ij} \) if \( i \neq 1, k \) and \( t'_{1j} = t_{kj}, t'_{kj} = t_{1j} \). Thus, we have

\[
I_1 \geq \int \ldots \int P \left( \epsilon_k + t'_{kk} (s) \geq \max_{j \neq k} \{ \epsilon_j + t'_{jj} (s) \} \right) \prod_{j=1}^{K} f_T (t'_j (s)) d\epsilon dt'_1 \ldots dt'_{K(1)}
\]

\[
= \int \ldots \int \int_{A_s^{(k)}} f_\epsilon (\varepsilon) \prod_{j=1}^{K} f_T (t_j (s)) d\varepsilon dt_1 \ldots dt_{K(1)}
\]

Considering the assumption of \( f_\epsilon (0) > 0 \), it can be shown that the above inequality is strict. This results in

\[
p_1 (u) > p_k (u).
\]

Now we show an intermediate result that would be used in proving Theorem 1. First, for any given \( u = (s, d)' \) we define

\[
\tilde{\eta}_J (d) = \prod_{k=1}^{K} \left[ g_d (\beta'_d X_{k, j}) = \max_m \{ g_m (\beta'_m X_{k, j}) \} \right]
\]

and

\[
\tilde{w}_J (s) = \prod_{i=1}^{K} \frac{1}{h_i} w \left( \frac{s - S (X_{i, j})}{h_i} \right).
\]

Now, let

\[
\tilde{p}_i (u) = \frac{\sum_{J \in \mathcal{J}} I \left[ Y_{ij} > \max_{k \neq i} \{ Y_{k, j} \} \right] \tilde{w}_J (s) \tilde{\eta}_J (d)}{\sum_{J \in \mathcal{J}} \tilde{w}_J (s) \tilde{\eta}_J (d)}, \quad (15)
\]

The following lemma shows that \( \tilde{p}_i \) above behaves almost as \( p_i \) for any \( u \) in large samples.

**Lemma 2.** Under Assumptions 1–4, for \( u = (s, d)' \) such that \( f (s, d) > 0 \) for \( s \) in an open interval containing \( s \), we have \( \tilde{p}_i (u) \to p_i (u) \) if \( h_i \to 0 \) and \( Nh_i \to \infty \), \( i = 1, \ldots, K \).

The following lemmas are needed for the proof of Lemma 2.

**Lemma 3.** Let \( U \) and \( V \) be positive random variables, defined on a probability space \((\Omega_1, \mathcal{F}_1, P_1)\), and \( A, B \) be a subsets of \( \Omega_1 \). We have (i) \( \text{Var} (U I_A) \leq \text{Var} (U) + E^2 (U) \), (ii) \( |\text{Cov} (U I_A, V I_B)| \leq |\text{Cov} (U, V)| + E (U) E (V) \).
Proof.

\[ \text{Var} (UI_A) = E (U^2 I_A) - E^2 (UI_A) \]
\[ \leq E (U^2) - E^2 (U) + E^2 (U) - E^2 (UI_A) \]
\[ = \text{Var} (U) + E^2 (U) - E^2 (UI_A) \]
\[ \leq \text{Var} (U) + E^2 (U) \]

This proves (i).

\[ \text{Cov} (UI_A, VI_B) = E (UVI_AI_B) - E (UI_A) E (VI_B) \]
\[ \leq E (UV) - E (UI_A) E (VI_B) - E (U) E (V) + E (U) E (V) \]
\[ = \text{Cov} (U, V) + E (U) E (V) - E (UI_A) E (VI_B) \]
\[ \leq \text{Cov} (U, V) + E (U) E (V) \leq |\text{Cov} (U, V)| + E (U) E (V). \]

Also,

\[ \text{Cov} (UI_A, VI_B) = E (UVI_AI_B) - E (UI_A) E (VI_B) \geq -E (UI_A) E (VI_B) \]
\[ \geq -E (U) E (V) - |\text{Cov} (U, V)|, \]

proving (ii). \qed

Now, we define

\[ \mu_{w,d,k} (h,s) = \sum_{d_1=1}^{K} \int \frac{1}{h} w \left( \frac{s - s_1}{h} \right) I (d_1 = d) f (s_1, d_1) ds_1, \]
\[ = \int \frac{1}{h} w \left( \frac{s - s_1}{h} \right) f (s_1, d) ds_1 \]
\[ \mu_w = \int w (s) ds \]

and

\[ \sigma_{w,d,k}^2 (h,s) = \sum_{d=1}^{K} \int \left( \frac{1}{h} w \left( \frac{s - s_1}{h} \right) I (d_1 = d) \right)^2 f_k (s_1, d) ds_1 - \mu_{w,d,k}^2 (h,s) \]
\[ = \int \left( \frac{1}{h} w \left( \frac{s - s_1}{h} \right) \right)^2 f_k (s_1, d) ds_1 - \mu_{w,d,k}^2 (h,s) \]
\[ \sigma_w^2 = \int w^2 (s_1) ds_1. \]
It can be verified that
\[
\lim_{h \downarrow 0} \mu_{w,d_0,k} (h, s_0) = f_k (s_0, d_0) \mu_w \quad (16)
\]
and
\[
\lim_{h \downarrow 0} h \sigma^2_{w,d_0,k} (h, s_0) = f_k (s_0, d_0) \sigma^2_w \quad (17)
\]

Straight forward calculations yield the following result.

**Lemma 4.** For \(J, J' \in \mathcal{J}\), let
\[
A (J,J') = \{1 \leq k \leq K : j_k = j'_k\}, \quad \text{and} \quad B (J,J') = \{1, \ldots, K\} \setminus A (J,J').
\]
Then, for \(J, J' \in \mathcal{J}\), we have
\[
E (\tilde{w}_J (s) \tilde{\eta}_J (d)) = \prod_{k=1}^K \mu_{w,d,k} (h_k, s) \quad (18)
\]
\[
\text{Cov} (\tilde{w}_J (s) \tilde{\eta}_J (d), \tilde{w}_{J'} (s) \tilde{\eta}_{J'} (d)) = \prod_{k \in A (J,J')} \sigma^2_{w,d,k} (h_k, s) \prod_{k \in B (J,J')} \mu^2_{w,d,k} (h_k, s) \quad (19)
\]

Now we prove Lemma 2.

**Proof.** First we analyze the numerator pf \(\tilde{p}_i\). It can be shown that
\[
E \left( I \left[ Y_{ij} > \max_{k \neq i} \{ Y_{kj} \} \right] \tilde{w}_J (s) \tilde{\eta}_J (d) \right) = p_i \prod_{k=1}^K f_k (s, d) + O_p (\max \{ h_k \}) \quad (20)
\]

The proof of this result is a standard procedure for kernel estimation of smooth functions.
and is omitted here. Next, consider the variance.

\[
Var \left( \sum_{J \in \mathcal{J}} I \left[ Y_{ij} > \max_{k \neq i} \{ Y_{ijk} \} \right] \bar{w}_J (s) \bar{\eta}_J (d) \right) = \sum_{J \in \mathcal{J}} \sum_{J' \in \mathcal{J}} Cov \left( I \left[ Y_{ij} > \max_{k \neq i} \{ Y_{ijk} \} \right] \bar{w}_J (s) \bar{\eta}_J (d) , I \left[ Y_{ij'} > \max_{k \neq i} \{ Y_{ijk'} \} \right] \bar{w}_{J'} (s) \bar{\eta}_{J'} (d) \right) \\
= \sum_{J, J' \in \mathcal{J}, A(J, J') \neq \emptyset} Cov \left( I \left[ Y_{ij} > \max_{k \neq i} \{ Y_{ijk} \} \right] \bar{w}_J (s) \bar{\eta}_J (d) , I \left[ Y_{ij'} > \max_{k \neq i} \{ Y_{ijk'} \} \right] \bar{w}_{J'} (s) \bar{\eta}_{J'} (d) \right) \\
\leq \sum_{J, J' \in \mathcal{J}, A(J, J') \neq \emptyset} \left| Cov \left( I \left[ Y_{ij} > \max_{k \neq i} \{ Y_{ijk} \} \right] \bar{w}_J (s) \bar{\eta}_J (d) , I \left[ Y_{ij'} > \max_{k \neq i} \{ Y_{ijk'} \} \right] \bar{w}_{J'} (s) \bar{\eta}_{J'} (d) \right) \right| \\
\leq \sum_{J, J' \in \mathcal{J}, A(J, J') \neq \emptyset} \left( \prod_{k \in A(J, J')} \sigma^2_{w, d, k} (h_k, s) \prod_{k \in B(J, J')} \mu^2_{w, d, k} (h_k, s) + \prod_{k=1}^K \mu^2_{w, d, k} (h_k, s) \right)
\]

Thus, by (18) and (19), for large \( N \),

\[
Var \left( \sum_{J \in \mathcal{J}} I \left[ Y_{ij} > \max_{k \neq i} \{ Y_{ijk} \} \right] \bar{w}_J (s) \bar{\eta}_J (d) \right) \leq \sum_{J, J' \in \mathcal{J}, A(J, J') \neq \emptyset} c_1 \prod_{k \in A(J, J')} \frac{1}{h_k}, \tag{21}
\]

where \( c_1 \) is a value that is not dependent on \( N \). The right-hand side of Equation (21) is the sum of \( \prod n_k (\prod n_k - \prod (n_k - 1)) \) terms. The number of terms for which \( \|A(J, J')\| = r \) is of order \( O \left( N^{rN2(K-r)} \right) \). These terms are of the form \( c_1 \) divided by the product of \( r \) of the \( h_k \)’s, and thus, the sum of these terms is of order \( o \left( N^{2K} \right) \). Therefore we conclude,

\[
Var \left( \sum_{J \in \mathcal{J}} I \left[ Y_{ij} > \max_{k \neq i} \{ Y_{ijk} \} \right] \bar{w}_J (s) \bar{\eta}_J (d) \right) = o \left( N^{2K} \right) .
\]

Combining (20) and (21) we have that

\[
\frac{1}{\prod n_k} \sum_{J \in \mathcal{J}} I \left[ Y_{ij} > \max_{k \neq i} \{ Y_{ijk} \} \right] \bar{w}_J (s) \bar{\eta}_J (d) \xrightarrow{P} p_i \prod_{k=1}^K f (s, d) . \tag{22}
\]
Following a similar procedure, we can show that

\[
\frac{1}{\prod n_k} \sum_{J \in J} \tilde{w}_J(s) \tilde{n}_J(d) \to \prod_{k=1}^K f(s, d).
\]  

(23)

Combining (22) and (23) we have the desired result.

\[ \square \]

Remark 2. From the proof of lemma 1 it can be seen that, to achieve optimal rate of convergence for variances of both the numerator and denominator of the right-hand side of (15), the bandwidth \( h_k \) need to be of order \( N^{-1/5} \) for \( k = 1, \ldots, K \).

We will introduce some additional notation before we prove Theorem 1.

Define

\[
\hat{F}^{(n)}_i(s, d) = \frac{1}{n_i} \sum_{j=1}^{n_i} I \left( \hat{S}(X_{ij}) \leq s, \delta(X_{ij}) = d \right)
\]

and

\[
\tilde{F}^{(n)}_i(s, d) = \frac{1}{n_i} \sum_{j=1}^{n_i} I \left( S(X_{ij}) \leq s, \delta(X_{ij}) = d \right).
\]

Furthermore, let \( y = (y_1, \ldots, y_K)' \), \( s = (s_1, \ldots, s_K)' \), \( d = (d_1, \ldots, d_K)' \), \( F^{(n)}(s, d) = F^{(n)}_1 \times \ldots \times F^{(n)}_K \), \( \tilde{F}^{(n)}(s, d) = \tilde{F}^{(n)}_1 \times \ldots \times \tilde{F}^{(n)}_K \), and finally, let \( F = F \times F_2 \times \ldots \times F_K \). We need the following assumption that is a very reasonable assumption from empirical distribution results.

Assumption 6. \( \left\| \tilde{F}_i - F_i \right\|_\infty = o_p(N^{-1/2} \log(N)) \).

Proof. Proof of Theorem 1

Let \( v(s) = \prod_{k=1}^K \frac{1}{h_k} \tilde{w} \left( \frac{s - s_k}{h_k} \right) I(s_{d_0} > \max_{k \neq d_0} s_k) \). By assumption 2, we have that \( v(s) \) is Riemann-integrable. Let \( S_{i(j)} \) denote the \( j \)th largest of \( S(X_{i1}), \ldots, S(X_{in_i}) \). Similarly we can define \( \hat{S}_{i(j)} \). Since \( \sup_{x \in S} \left| \hat{g}_i (\beta' x) - g(\beta, x) \right| = O_p(N^{-2/5} \log N) \), we have, \( \sup_{x \in S} \left| S(x) - \hat{S}(x) \right| = O_p(N^{-2/5} \log N) \). Additionally, \( \max_{i,j} \left| \hat{S}(X_{ij}) - S(X_{ij}) \right| = O_p(N^{-2/5} \log N) \). After some tedious calculations we can deduce that the above also implies \( \max_{i,j} \left| \hat{S}_{i(j)} - S_{i(j)} \right| = O_p(N^{-2/5} \log N) \). Combine this with 6, and the fact that \( \hat{d}_0 \to d_0 \) (by the fact that \( s_0 \) is positive and the second part of Assumption 5), we can find sets \( A_N \subset \Omega \) and positive numbers \( a_N \propto N^{-2/5} \log N \) such that for \( \omega \in A_N \), \( \max_{i,j} \left| \hat{S}_{i(j)} - S_{i(j)} \right| + \left\| \tilde{F}^{(n)} - F^{(n)} \right\|_\infty \leq a_N \) and \( \hat{d}_0 = d_0 \). Define \( q_{ij} \),
\[ i = 1, \ldots, K; j = 1, \ldots, n_i \] to be values such that \( F(q_{ij}, d) = \frac{j}{n_1 F(\infty, d) + 1} \). With condition 1, we can find \( b_N \propto N^{-1/2} \log N \) such that \( \max_{i,j} |q_{ij} - \bar{S}_{ij}| \leq b_N \) if \( q_{ij} \in (s_0 - d_1, s_0 + d_1) \), a neighborhood of \( s_0 \). Thus, by properly redefining \( a_N \propto N^{-2/5} \), we can assume when \( \omega \in A_N \), \( \max_{i,j} |q_{ij} - \hat{S}_{ij}| \leq a_N \). By condition 2, without loss of generality we can assume that \( w \left( \frac{s - s_0}{h_k} \right) = 0 \) for \( s \) outside \((s_0 - d_1, s_0 + d_1)\). Let \( c_1, \ldots, c_m \) be such that \( c_i - c_{i-1} = 4b_N \), and the support of \( w \) is within \( \left( \frac{c_1 - s_0 + 2b_N}{h_k^{(2)}}, \frac{c_m - s_0 - 2b_N}{h_k^{(2)}} \right) \).

Define \( c_i' = \frac{c_i + c_{i+1}}{2} \) for \( i = 1, \ldots, m - 1 \). For \( k = 0, \ldots, 2^K - 1 \), let \( I_k \) be collection of \( K \)-dimensional intervals of the form \([c_{i_1}^* - c_{i_1}^*, c_{i_1}^*] \times \cdots \times [c_{i_K}^* - c_{i_K}^*, c_{i_K}^*] \), where \( c_{i_j}^* \) is either \( c_{i_j} \) or \( c_{i_j}' \) depending the whether the \( j \)th position of the \( k \) when written as a binary number is 0 or 1. For \( I \in I_k \), let \( \tilde{v}_{I,h} \left( \bar{w}_{I,h} \right) \) be the supremum (infimum) of \( v(s) \) over \( s \in I \). For \( J = (j_1, \ldots, j_K) \in \mathcal{J} \), let \( I_J^{(k)} \) denote the \( I \) in \( I_k \) such that \( s = (q_{ij_1}, \ldots, q_{ij_K}) \in I \). By condition 2, we have

\[
\frac{1}{n_1 \cdots n_K} \sum_{J \in \mathcal{J}} \left( \tilde{v}_{I_J^{(k)}}, h - \bar{w}_{I_J^{(k)}}, h \right) \leq \frac{1}{n_1 \cdots n_K} \prod_{k=1}^{2^K - 1} \left[ 4b_N n_k \| f \|_\infty \right] \sum_{I \in I_k} \left( \bar{v}_{I}, h - \bar{v}_{I}, h \right) = O \left( (4b_N)^K \sum_{I \in I_k} \left( \bar{v}_{I}, h - \bar{v}_{I}, h \right) \right) \to 0
\]

for \( k = 0, \ldots, 2^K - 1 \). Note that the first inequality in the above expression is due to the fact that the number of \( q_{ij} \)'s that falls into \([c_k, c_{k+1}] \) \([c_k', c_{k+1}'] \) is less than \([4b_N n_i \| f \|_\infty] \). When \( \omega \in A_N \), for \( J = (j_1, \ldots, j_K) \in \mathcal{J} \), we have that \( S = (S(X_{ij_1}), \ldots, S(X_{ij_K})) \) must be in one of the \( I_J^{(k)} \)'s. Thus,

\[
\frac{1}{n_1 \cdots n_K} \left( \sum_{J \in \mathcal{J}} I \left[ Y_{ij_i} > \max_{k \neq i} \{ Y_{ij_k} \} \right] \tilde{w}_J (s) \tilde{\eta}_J (d) - \sum_{J \in \mathcal{J}} I \left[ Y_{ij_i} > \max_{k \neq i} \{ Y_{ij_k} \} \right] \bar{w}_J (s) \bar{\eta}_J (d) \right) \leq \frac{1}{n_1 \cdots n_K} \sum_{k=0}^{2^K - 1} \sum_{J \in \mathcal{J}_k} \left( \left( \bar{v}_{I_J} - \bar{v}_{I_J}, h \right) \right) \to 0
\]

Similarly, we can show that when \( \omega \in A_N \),

\[
\frac{1}{n_1 \cdots n_K} \left( \sum_{J \in \mathcal{J}} \tilde{w}_J (s) \tilde{\eta}_J (d) - \sum_{J \in \mathcal{J}} \bar{w}_J (s) \bar{\eta}_J (d) \right) \to 0
\]

These combined with (22) and (23) give us the desired result. \( \Box \)