, *pp*. 1–15

Latent Supervised Learning for Estimating Treatment Effect Heterogeneity

BY SUSAN WEI

Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, U.S.A.

susanwe@live.unc.edu

AND MICHAEL R. KOSOROK

Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, U.S.A.

kosorok@unc.edu

SUMMARY

It is oft observed in medicine that what works for one patient may not work for another. Determining for whom a treatment works and does not work is of great clinical interest. We propose a methodology to estimate treatment effect heterogeneity, i.e. to ascertain for which subpopulations a treatment is effective or harmful. The model studied assumes the relationship between an outcome of interest (e.g. blood pressure, cholesterol, survival) and a set of covariates (e.g. treatment, age, gender) is modified by a linear combination of a set of features (e.g. gene expression). Specifically a threshold on the linear combination divides the population into two subpopulations with different responses to treatment. Techniques from Latent Supervised Learning, a novel machine learning idea, is applied for model estimation, i.e. estimation of the linear combination and the corresponding threshold. Consistency of the estimator is established. In simulations the proposed methodology demonstrates high classification accuracy in a wide array of settings. Three data analysis examples are presented to illustrate the efficacy and applicability of the proposed methodology.

Some key words: Biomarker; Cox model; Empirical processes; Generalized linear model; Glivenko-Cantelli; Personalized medicine; Sieve maximum likelihood; Sliced inverse regression; Subgroup analysis; Survival analysis; Treatment interaction.

1. INTRODUCTION

Treatment effect heterogeneity is often observed in medical studies, i.e. different treatments having different effects on different individuals. For instance, a treatment can be beneficial for all the subpopulations but with varying magnitudes, or of more interest, a treatment is only beneficial for certain subpopulations. Subgroup analysis is a commonly used approach to estimate for which subpopulations a treatment is beneficial or harmful. However, its misuse is well-documented and subgroup analysis remains quite controversial in the realm of medical research (Rothwell, 2005; Lagakos, 2006).

In this paper we study a parsimonious model where the relationship between the outcome of interest and a set of covariates is modified by a linear combination of a set of features. Specifically, there exists a separating hyperplane in the feature space that divides the population into

two subgroups with different regression coefficients. For example, the outcome of interest could be survival time, the covariates could include a treatment indicator and a set of confounding variables, and the set of features could be gene expression variables. The model then postulates the existence of two subpopulations whose treatment response differ according to a linear combination of gene expression values.

We first present the model for the case when the outcome of interest belongs to the exponential family of distributions. The analogous model for Y being a right-censored survival time is presented later in Section 3. Recall the standard generalized linear model is characterized by the following features:

Property 1. a linear predictor $\eta = \beta^T U$.

Property 2. a differentiable one-to-one link function g which specifies the relationship between the mean $E(Y) = \mu$ and the linear predictor: $g(\mu) = \eta$.

Property 3. a variance function V which specifies the relationship between the mean and the variance: $Var(Y) = \phi V(\mu)$ where ϕ represents the dispersion parameter.

We introduce several additional components to the the standard generalized linear model. First, we decompose the independent variable into two parts, a joint and individual one. Let $U \in \mathbb{R}^d$ denote the joint component and $Z \in \mathbb{R}^{d'}$ denote the individual component. Per the setting motivated above, the term joint refers to the fact that there is a single regression coefficient for U while the term individual reflects the condition that the regression coefficient for Z depends on the features $X \in \mathbb{R}^p$. This dependency is only through the value of the indicator $1\{\omega^T X - \gamma \geq 0\}$. Here ω is of the same dimension as X and has unit length, $||\omega|| = 1$. Specifically, the linear predictor has the following form:

$$\eta_i = (\beta_1 + (\beta_2 - \beta_1)1\{\omega^T x_i - \gamma \geq 0\})^T z_i + \delta^T u_i. \tag{1}$$

The primary interest is to estimate the separating hyperplane determined by ω and γ. This in turn gives estimates of the regression coefficients β_1, β_2 and δ.

A subset of the feature X is allowed to be part of the joint variable or the individual variable or both. In order to ensure identifiability, U and Z cannot contain the same variables, however. To be as general as possible, the intercept term is included in the individual component Z and not in the joint component U.

The independent variables Z and U are often low-dimensional. In applications of primary interest to us, Z is the treatment variable and U the confounding variables. The X features can be higher-dimensional such as gene expression profiles.

Existing methodologies for estimating treatment effect heterogeneity have focused on treatment-covariate interaction. Taking a variable selection viewpoint, the methods proposed in Imai & Strauss (2011) and Gunter et al. (2011) seek to identify variables with large interaction effects with treatment for further analysis and validation. The personalized medicine methods in Qian & Murphy (2011) and Zhao et al. (2012) focus on designing optimal treatment regimes for each individual and thus indirectly estimate treatment effect heterogeneity. More commonly used methods such as Boosting (LeBlanc & Kooperberg, 2010), Bayesian Additive Regression Trees (Chipman et al., 2010), and other tree-based approaches (Su et al., 2009) focus on prediction and can be difficult to interpet.

In the model proposed here, we are interested in estimating the treatment-subgroup interaction effect as opposed to treatment-covariate interactions. Importantly, in the setting studied here, the subgroup term is unknown a priori and as such we cannot apply the above methodologies for model estimation. The proposed model is worthwhile to study for its simplicity, ease of inter-

pretability, and parsimony. Similar to the models studied in Imai & Strauss (2011) and Gunter et al. (2011), variable selection is in some ways built into our framework. Namely, the coefficients of ω in Model (1) are rough indicators of the importance of a feature in terms of how much it drives the separation in treatment responses between the two subpopulations. The main advantage of the proposed model over those studied previously is the fact that a general classification rule is learned for identifying subpopulations with treatment effect heterogeneity. This is in contrast to the methods in Imai & Strauss (2011) and Gunter et al. (2011) which can only go so far as to give soft characterizations of the subpopulations with different treatment responses instead of a hard classification rule.

We apply techniques from a novel machine learning idea called Latent Supervised Learning for model estimation. Latent Supervised Learning bridges the gap between unsupervised and supervised learning. The basic idea is to use a continuous surrogate variable to supervise the learning of a binary classifier. The following Gaussian classification problem was considered in Wei & Kosorok (2013):

$$Y \sim N(\mu_1, \sigma_1^2) \text{ when } \omega^T X - \gamma \geq 0$$

and

$$Y \sim N(\mu_2, \sigma_2^2) \text{ when } \omega^T X - \gamma < 0.$$

This model is a special case of Model (1) studied here. To see this, set $Z = 1$ and U to be empty in (1). Besides introducing regressors, we also allow Y to be any member of the exponential family of distributions rather than restricting its distribution to be Gaussian. Finally, we allow the mean response to be related to the linear predictor through a link function g. These additions require adaptation of the original Latent Supervised Learning methodology.

2. METHODOLOGY

2·1. *Sieve maximum likelihood estimation*

This section focuses on model estimation for the case when Y is in the exponential family of distributions. The special case when Y is a right censored survival time is treated thoroughly in Section 3. We also restrict our attention to the case when there is no overlap between Z and X. The case when there is overlap is postponed to the Appendix.

The log likelihood of the data, denoted L_n, can be parametrized in terms of ω alone. Given ω and γ, the regression coefficients β_1, β_2 and δ in (1) can be found via the standard Fisher scoring method employed in generalized linear model estimation. Next, given ω, the cutpoint γ in the separating hyperplane can be found via a simple grid search. Thus we only need concern ourselves with a single parameter function $L_n(\omega)$.

A natural approach to estimate ω is via maximum likelihood. However, direct maximization over $\mathbb{R}^p$ is computationally challenging when the dimension of the feature X is large. Instead we consider maximization over a data-driven approximating space that grows dense as the sample size increases. We will refer to such a sequence of approximating spaces as a sieve, following the terminology in Grenander (1981).

A sieve maximum likelihood approach was also used in two previous papers on Latent Supervised Learning, first in the Gaussian classification problem studied in Wei & Kosorok (2013) and next in the survival time classification problem in a technical report by the same authors available online at Bepress (University of North Carolina, Chapel Hill, Department of Biostatistics). In these papers, a preliminary sieve based on information in the covariate space is first constructed

and the sieve is next improved by incorporating the response variable Y. The methodology employed here is adapted from this general strategy.

The simple sieve is constructed as follows. The convex hull of the point cloud in the X feature space is first computed. For each possible binary enumeration of the points on the convex hull, calculate the direction which connects the means of the two classes. Note this is the normal vector to the separating hyperplane produced by the simple binary classifier known as the centroid method in Hastie et al. (2001). If the number of points on the convex hull is very large, randomly select a subset of m points. We recommend m no bigger than 10. The collection of the 2^m directions trained in this manner shall be referred to as the simple sieve.

Remark 1. In Wei & Kosorok (2013) and the subsequent technical report, the X space was first partitioned into several regions and the simple sieve direction was calculated within each region. This is much more computationally intensive than what is proposed here.

Remark 2. Convex hull computation can be difficult in high dimensions. If necessary, first reduce the dimension of the X feature space via, say Principal Components Analysis. Interestingly, in very high dimensions, all points are on the convex hull, see Hall et al. (2005). Thus, a random number of points can be chosen instead of the actual computation of the convex hull when this is the case.

For each ω direction in the simple sieve, Sliced Inverse Regression Li (1991) is performed on the bivariate $(Y, \omega^T X)$. For simplicity assume X has already been standardized to have mean zero and unit covariance. The improved sieve is created as follows.

Step 1. Slice the range of Y into several non-overlapping regions. Follow this by slicing on the range of $\omega^T X$ within each slice of Y. Let I_h denote the h-th slice for $h = 1, \ldots, H$.

Step 2. Calculate the weighted sample covariance matrix

$$\hat{V}_n(\omega) = \sum_{h=1}^{H} \hat{p}_h \hat{m}_h(\omega)' \hat{m}_h(\omega) \tag{2}$$

where $\hat{m}_h(\omega)$ is the unbiased estimate of $m_h(\omega) = E(X \mid (Y, \omega^T X) \in I_h)$ based on the sample average of X for $(Y, \omega^T X) \in I_h$ and similarly for $\hat{p}_h(\omega)$, the unbiased estimate of the quantity $p_h(\omega) = pr((Y, \omega^T X) \in I_h)$.

Remark 3. In Wei & Kosorok (2013) and the subsequent technical report, the bivariate $(Y, 1\{\omega^T X - \gamma \geq 0\})$ is sliced. Here we eliminate the need to first estimate γ. This results in a computational improvement.

Remark 4. The number of slices H need not increase with n. We have found that slicing Y into roughly $n/10$ slices and then further slicing $\omega^T X$ into two slices works well in practice.

Under certain conditions we will detail in the next section, the largest eigenvector $\hat{\nu}_n(\omega)$ of the weighted sample covariance matrix $\hat{V}_n(\omega)$ in (2) is guaranteed to be consistent for ω_0. Let $\hat{\Omega}_n$ be the collection of directions $\hat{\nu}_n(\omega)$ where ω ranges over the simple sieve.

The final estimate for ω_0 is

$$\hat{\omega}_n = \arg\max_{\omega \in \hat{\Omega}_n} L_n(\omega). \tag{3}$$

2·2. *Asymptotic results*

We establish the asymptotic properties of the sieve maximum likelihood estimator in this section. Let P be the probability measure generating the data under Model (1) conditional on the variables U and Z. Let $\mathbb{P}_n$ be the empirical measure. The regression coefficients $\beta_1, \beta_2 \in \mathbb{R}^{d'}, \delta \in \mathbb{R}^d$ are collected into the variable ψ. The direction vector ω is constrained to have unit length. The unknown parameters can be collected as $\theta \equiv (\psi, \omega, \gamma)$ and the subscript zero will be used to denote the true parameter values.

Since Y is a member of the exponential family, we can write its density in the following form

$$f(y|\zeta) = c(y)\exp(\zeta y - b(\zeta)) \tag{4}$$

where we have written the distribution in the canonical form (or natural form). We can further simplify the expression by using the canonical link function g which gives $\eta = \zeta$ in which case the density in (4) can be rewritten as

$$f(y|\eta) = c(y)\exp(\eta y - b(\eta)) \tag{5}$$

Remark 5. The original definition of Nelder & Wedderburn (1972) introduces an additional nuisance parameter in (4). The maximum likelihood estimator of θ remains unchanged. Thus, without loss of generality, we content ourselves to the simpler form in (4).

From Equation 5, it is easy to see that maximizing the log likelihood of the data is the same as maximizing $M_n(\theta) \equiv \mathbb{P}_n m_\theta$, where

$$\begin{aligned} m_\theta(y,x) &\equiv \eta y + b(\eta) \\ &= [(\beta_1 + (\beta_2 - \beta_1)1\{\omega^T x - \gamma \leq 0\})^T z + \delta^T u]y \\ &+ b((\beta_1 + (\beta_2 - \beta_1)1\{\omega^T x - \gamma \leq 0\})^T z + \delta^T u) \end{aligned} \tag{6}$$

Let $\hat{\theta}_n$ be the sieve maximizer of $M_n(\theta)$, where $\hat{\theta}_n \equiv (\hat{\psi}_n, \hat{\omega}_n, \hat{\gamma}_n)$.

The following conditions will be needed:

Condition 1. The parameter space of the regression coefficients ψ is compact.

Condition 2. The cuptoint γ is known to lie in a bounded interval $[a, b]$.

Condition 3. The univariate random variable $\omega^T X$ has a strictly bounded and positive density f where $||\omega|| = 1$.

Condition 4. For any $b \in \mathbb{R}^p$, the conditional expectation $E(b^T X|\omega_0^T X)$ is linear in $\omega_0^T X$.

Condition 5. The change-line regression coefficients $\beta_{1,0} \neq \beta_{2,0}$.

Condition 6. The covariate X has a continuous distribution.

Conditions 1 and 2 guarantee the existence of $\hat{\theta}_n$. Condition 3 is used in establishing semicontinuity of the theoretical objective function. Condition 4 is a key assumption in Li (1991) and is satisfied when the distribution of X is Gaussian or more generally, elliptically symmetric. The last two conditions guarantee the model is identifiable.

THEOREM 1 (CONSISTENCY). *Under Conditions* 1–6*, the sieve estimator $\hat{\theta}_n$ is consistent.*

Proof. Our approach to establishing consistency will be to utilize the argmax theorem (Theorem 14.1 in Kosorok (2008). We first need to show that $M_n \rightsquigarrow M$ in $l^\infty(K)$ for all compact $K \subset H = \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^{d'} \times \mathbb{S}^p \times [a, b]$ where $M(\theta) \equiv P m_\theta$ and $\mathbb{S}^p$ is the collection

of vectors in $\mathbb{R}^p$ with unit length. We will also need to show that $\theta \mapsto M(\theta)$ is upper semicontinuous with a unique maximum at θ_0. Near-maximization must then be established, i.e. $M_n(\hat{\theta}_n) \geq M_n(\theta_0) - o_P(1)$. Finally, the argmax theorem will yield that $\hat{\theta}_n$ converges to θ_0 in probability.

Fix a compact $K \subset H$. We now verify that $\mathcal{F}_K \equiv \{m_\theta : \theta \in K\}$ is Glivenko-Cantelli. The latent features X can be partitioned into four mutually exclusive sets: $A_1 \equiv \{\omega^T X \leq \gamma, \omega_0^T X \leq \gamma_0\}, A_2 \equiv \{\omega^T X \leq \gamma, \omega_0^T X > \gamma_0\}, A_3 \equiv \{\omega^T X > \gamma, \omega_0^T X \leq \gamma_0\}, A_4 \equiv \{\omega^T X > \gamma, \omega_0^T X > \gamma_0\}$. We can write

$$\begin{aligned} m_\theta(y,x) = &\{(\beta_1^T Z + \delta^T U)y + b(\beta_1^T Z + \delta^T U)\}1\{x \in A_1\} \\ &+ \{(\beta_1^T Z + \delta^T U)y + b(\beta_1^T Z + \delta^T U)\}1\{x \in A_2\} \\ &+ \{(\beta_2^T Z + \delta^T U)y + b(\beta_2^T Z + \delta^T U)\}1\{x \in A_3\} \\ &+ \{(\beta_2^T Z + \delta^T U)y + b(\beta_2^T Z + \delta^T U)\}1\{x \in A_4\}. \end{aligned} \tag{7}$$

It is easy to see that the classes $\{\beta^T Z : \theta \in K\}$ and $\{\delta^T U : \theta \in K\}$ are separately Glivenko-Cantelli classes. Thus the sum is also Glivenko Cantelli by Corollary 9.27 (i) in Kosorok (2008). The product of the classes $\{\beta^T Z + \delta^T U : \theta \in K\}$ and $\{y\}$ is also Glivenko-Cantelli by Corollary 9.27 (ii) since the product of the two envelopes is integrable. The function b is continuous and by Corollary 9.27 (iii), the class $\{b(\beta_2^T Z + \delta^T U) : \theta \in K\}$ is Glivenko-Cantelli since the envelope is integrable. It was shown in Wei & Kosorok (2013) and the subsequent technical report the class of indicator function $\{1\{\omega^T X - \gamma \geq 0\} : \theta \in K\}$ is Glivenko-Cantelli. Reapplications of Corollary 9.27 (i) and Corollary 9.27 (ii) shows $\mathcal{F}_k$ itself is Glivenko-Cantelli. Thus $M_n \rightsquigarrow M$ in $l^\infty(K)$ for all compact K.

We now establish upper semicontinuity of $\theta \mapsto M(\theta)$. Using the same sets described above, we have

$$\begin{aligned} M(\theta) &= P(\eta y + b(\eta)) \\ &= \{(\beta_1^T Z + \delta^T U)g^{-1}(\beta_{1,0}^T Z + \delta_0^T U) + b(\beta_1^T Z + \delta^T U)\}P(A_1) \\ &+ \{(\beta_1^T Z + \delta^T U)g^{-1}(\beta_{2,0}^T Z + \delta_0^T U) + b(\beta_1^T Z + \delta^T U)\}P(A_2) \\ &+ \{(\beta_2^T Z + \delta^T U)g^{-1}(\beta_{1,0}^T Z + \delta_0^T U) + b(\beta_2^T Z + \delta^T U)\}P(A_3) \\ &+ \{(\beta_2^T Z + \delta^T U)g^{-1}(\beta_{2,0}^T Z + \delta_0^T U) + b(\beta_2^T Z + \delta^T U)\}P(A_4). \end{aligned} \tag{8}$$

The function g is differentiable and thus continuous, hence g^{-1} is continuous. Next, the function $b(\eta)$ is differentiable and thus continuous. Finally by Condition 3, $\omega^T X$ and $\omega_0^T X$ have bounded densities. Thus $M(\theta)$ is continuous and therefore upper semicontinuous.

Identifiability plus the Kullback-Leibler discrepancy will show that M has a unique maximum at θ_0. Condition 5 ensures the regression coefficients are identifiable. The normal vector to the separating hyperplane ω and the cutpoint γ are identifiable up to sign. Condition 6 guarantees that $\omega = \omega'$ whenever $1\{\omega^T X - \gamma\} = 1\{\omega'^T X - \gamma_0\}$.

Finally, we establish near maximizability. Lemma 1 in the Appendix establishes the sieve $\hat{\Omega}_n$ is dense, i.e. there exists a sequence $\omega_n \in \hat{\Omega}_n$ that converges to ω_0. Let γ_n and ψ_n be the cutpoint and regression estimates corresponding to ω_n, respectively. Denote this sequence by $\theta_n = (\psi_n, \omega_n, \gamma_n)$. Since $\hat{\theta}_n$ maximizes $M_n(\theta)$ over $\hat{\Omega}_n$, we have $M_n(\hat{\theta}_n) \geq M_n(\theta_n)$. By the continuity of $M(\theta)$, we have $M_n(\theta_n) - M_n(\theta_0) = o_P(1)$ and thus

$$M_n(\hat{\theta}_n) \geq M_n(\theta_0) - o_P(1).$$

Now the conditions of the argmax theorem are met, and the desired consistency follows. □

3. EXTENSION TO RIGHT-CENSORED SURVIVAL DATA

In this section we consider the case when Y is a right-censored survival time. Let T denote the true lifetime and C the censoring time. The observed data consists of $Y = \min(T, C)$, the censoring indicator $1\{T \leq C\}$, along with the covariates Z and U and the feature vector X. Censoring is assumed to be independent of survival, conditional on X.

For simplicity assume there is no overlap between X and Z. The linear predictor η takes on a similar form as in Equation (1) except Z no longer contains an intercept term. The intercept term can only be added to one subgroup, otherwise the model would not be identifiable. We have the following form for the linear predictor:

$$\eta_i = (\beta_1 + (\beta_2 - \beta_1)1\{\omega^T x_i - \gamma \geq 0\})^T z_i + \text{intercept} * 1\{\omega^T x_i - \gamma \geq 0\} + \delta^T u_i. \quad (9)$$

Our model is a Cox model with the linear predictor as in (9). The conditional hazard function is

$$h(t|z, u, x) = \exp(\eta) h_0(t)$$

where $h_0(t)$ is the baseline hazard function.

The survival model above generalizes the survival classification problem studied in the technical report by Wei and Kosorok which was

$$h(t|x) = \exp(\beta) h_0(t) \text{ when } \omega^T X - \gamma \geq 0$$

and

$$h(t|x) = h_0(t) \text{ when } \omega^T X - \gamma \geq 0.$$

To see this, set $Z = 1$ and U to be empty.

We follow the general strategy in the technical report by Wei and Kosorok of accounting for censoring. Let $0 = t_1 < t_2 < \cdots < t_H < \infty = t_{H+1}$ be a partition of the observed survival times and let I_h be the h-th slice from slicing on the pair $(T, \omega^T X)$. Because T is not always observed, we must be careful in estimating $m_h(\omega) = E(X|T, \omega^T X \in I_h)$. As in the technical report by Wei and Kosorok, we use a method called Recursively Imputed Survival Trees, introduced in Zhu & Kosorok (2012), to estimate the weight function

$$w(Y, t, X) = P(T \geq t|X)/P(T \geq Y|X).$$

Let the resulting estimate be denoted by $\hat{w}$. In the technical report by Wei and Kosorok, unbiased estimates of m_h and p_h were derived using $\hat{w}$ to adjust for censoring. The expressions are

$$\hat{m}_h(\omega) = \frac{1}{n\hat{p}_h(\omega)} \sum z_i \{(y_i, \omega^T x_i) \in I_h\} + \hat{w}(y_i, t_h, x_i) 1\{y_i < t_h, \delta_i = 0, \omega^T x_i \in I_h\}$$
$$- \hat{w}(y_i, t_{h+1}, x_i) 1\{y_i < t_{h+1}, \delta_i = 0, \omega^T x_i \in I_h\}\}, \quad (10)$$

where

$$\hat{p}_h(\omega) = n^{-1} \sum 1\{(y_i, \omega^T x_i) \in I_h\} + \hat{w}(y_i, t_h, x_i) 1\{y_i < t_h, \alpha_i = 0, \omega^T x_i \in I_h\}$$
$$- \hat{w}(y_i, t_{h+1}, x_i) 1\{y_i < t_{h+1}, \alpha_i = 0, \omega^T x_i \in I_h\}. \quad (11)$$

The remaining steps are the same as the methodology outlined in Section 2·1. Namely, calculate the weighted sample covariance matrix in Equation (2) based on Equations (10) and (11). Follow by taking the largest eigenvector of the weighted sample covariance matrix which will become a candidate direction in the sieve. Finally, the estimate of ω is the direction in the sieve that maximizes the likelihood of the data.

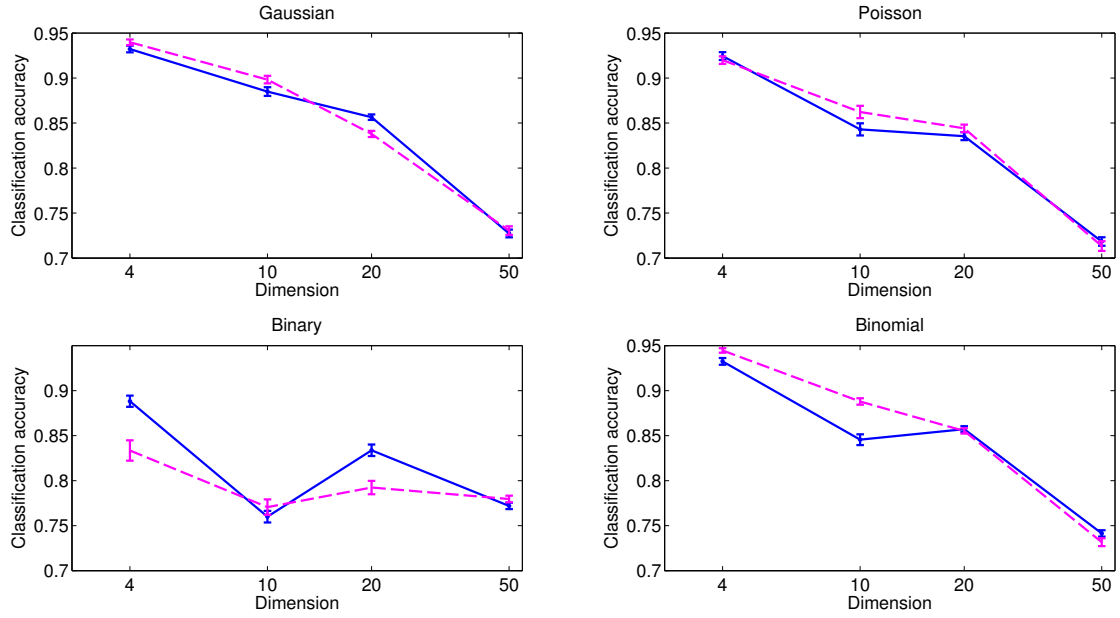

Fig. 1: Exponential family distributions: the classification accuracy of the proposed methodology when X is not in U (solid) and X is in U (dashes) for various distributions of Y. Error bars are also indicated.

We briefly given an outline for the consistency of the sieve maximum likelihood estimator in the survival setting. Instead of a true likelihood, we work with a profile likelihood in the survival model proposed here. That the profile likelihood is continuous with a unique maximum follows from the consistency proof given in the technical report by Wei and Kosorok. The Glivenko-Cantelli component of the Argmax Theorem can be established similar to the argument given in Section 2·2. Finally, demonstrating near-maximizability is quite straightforward given the continuity.

4. SIMULATIONS

Simulations are conducted to examine the performance of the proposed methodology under various settings. In particular, the following factors are investigated: 1) the dimension of the feature vector X, 2) overlap between X and U, and 3) the distribution of the response variable Y.

The dimensions $p = 4, 10, 20, 50$ are considered for X. When there is no overlap between X and U, the dimension of U is set to 5. Otherwise the first two marginals of X are added, bringing the dimension of U to 7. We also consider several different distributions for the outcome variable Y – Gaussian, Poisson, Bernoulli, and Binomial with 10 trials.

The sample size is fixed at 100 observations. The direction ω is set to the unit vector in the direction of $(1, \ldots, 1, -1, \ldots, -1)$ where the number of positive 1's is roughly half of p. The cutpoint γ is fixed at 0. The individual regression coefficients are set to $\beta_1 = (1/2, -\log p, \ldots, -\log p)$ and $\beta_2 = (1/4, \log p, \ldots, \log p)$. The value of the joint regression coefficient δ is drawn from a d-variate random variable with uniform$(-1, 1)$ independent marginals.

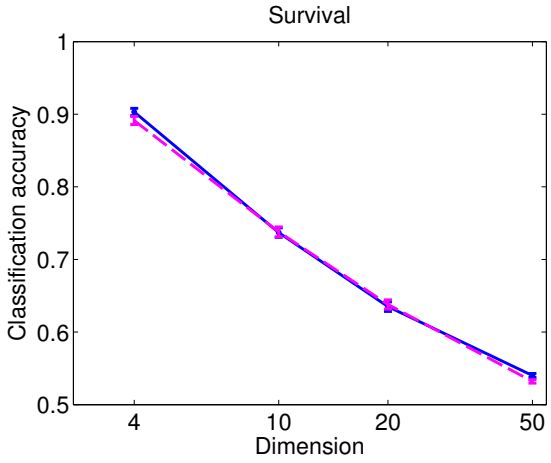

Fig. 2: Survival time: the classification accuracy of the proposed methodology when X is not in U (solid) and X is in U (dashes). Error bars are also indicated.

We perform 100 Monte Carlo realizations for each simulation setting. The same δ is used for each dimension of X. In contrast, the joint and individual variables U and Z and the feature vector X are randomly drawn from the standard multivariate Gaussian distribution for each Monte Carlo realization.

Accuracy of the methodology is measured by the percentage of subjects classified correctly on a large independent test set. This only involves generating realizations of the feature vector X. Figure 1 displays the average classification accuracy over 100 Monte Carlo simulations for different distributions of Y in the exponential family. A general trend across all four panels is a decrease in classification accuracy as dimension of X increases. However even up to dimension 50, the classification accuracy is reasonably high considering the sample size is fixed at $n = 100$. The classification performance is comparable for the Gaussian, Poisson and Binomial distribution. The Bernoulli setting proves to be a bit more challenging than the Binomial since there is less data here. Finally, the performance of the method does not seem very sensitive to whether there is overlap between X and U.

Next we examine the performance of the methodology for survival outcome. Let η be as in Equation (9) with the individual regression coefficients set to $\beta_1 = (-\log p, \ldots, -\log p)$ and $\beta_2 = (\log p, \ldots, \log p)$, and the intercept to 1/2. The true survival time T is exponential distributed with mean $2/\exp(\eta)$. We observe $Y = \min(T, C)$ where the censoring time C is distributed uniform$(0, 10)$. The percentage of censoring is approximately 30%. In all other regards the parameters for the survival simulation is as above.

Figure 2 shows the classification performance of the methodology for survival outcome. Again, we see the accuracy decreasing as dimension increases. The classification accuracy is just above 50% when the dimension of X is 50. The dimensionality of X proves to be more challenging in the survival setting perhaps due to the presence of censoring. Still, the classification at moderately high dimensions is reasonably accurate.

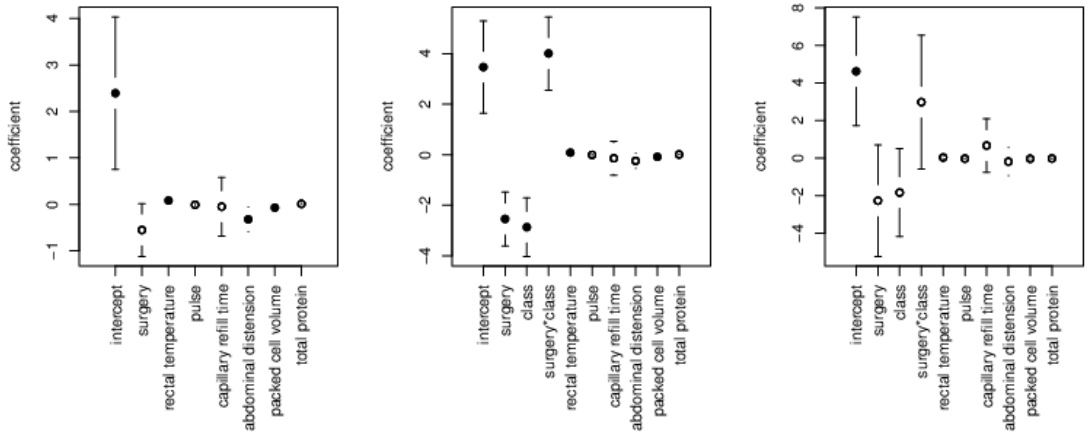

Fig. 3: Horse colic dataset: estimated coefficients (and error bars) in the logistic regression models (i) $Y \sim Z + U$ on the training set, (ii) $Y \sim Z + Z1\{\hat{\omega}_n^T X - \hat{\gamma}_n \geq 0\} + U$ on the training set, and (iii) $Y \sim Z + Z1\{\hat{\omega}_n^T X - \hat{\gamma}_n \geq 0\} + U$ on the test set. Filled circles indicate significance at the 0.05 level.

5. DATA EXAMPLES

5·1. *Horse colic disease*

In this section, we study a dataset containing information on 368 horses suffering from colic. We are interested in determining whether it is beneficial to perform surgery on horses with colic. Let Y be the binary endpoint dead versus alive, coded 1 and 0, respectively. Let Z be the treatment indicator, 1 for surgery and 0 for traditional treatment. The joint variable U consists of the confounding variables rectal temperature, pulse, respiratory rate, capillary refill time, abdominal distension, packed cell volume, and abdomcentesis total protein.

The dataset is split into a training set ($n = 300$) and a test set ($n = 68$). We first fit a standard logistic regression model $Y \sim Z + U$. The estimated coefficients are displayed in the first panel of Figure 3. Surgery is significantly associated with mortality but it is a negative effect, i.e. the horse has a higher chance of dying if it is treated with surgery.

The proposed methodology is applied to discover subgroups with possibly different treatment responses. The X variables considered for possible interaction with treatment include age (1 for old, 0 for young) and the surgical possibility of the lesion (1 for surgical, 0 for not surgical). The output of the proposed method is the following separating hyperplane

$$0.9010 * \text{age} + 0.4338 * \text{surgical lesion} \geq 1.3348. \tag{12}$$

Thus class 1 (when the above expression is true) is comprised of old horses with lesions that are indeed surgical. Class 0 (when the above expression is false) is comprised of young horses and horses whose lesions were not surgical.

The second panel of Figure 3 shows the coefficients of the model $Y \sim Z + Z1\{\hat{\omega}_n^T X - \hat{\gamma}_n \geq 0\} + U$ on the training data. We see that surgery has a significantly beneficial effect for class 1 horses, i.e. old horses with surgical lesions. On the other had, surgery is seen to be significantly harmful in young horses and horses who do not have surgical lesions.

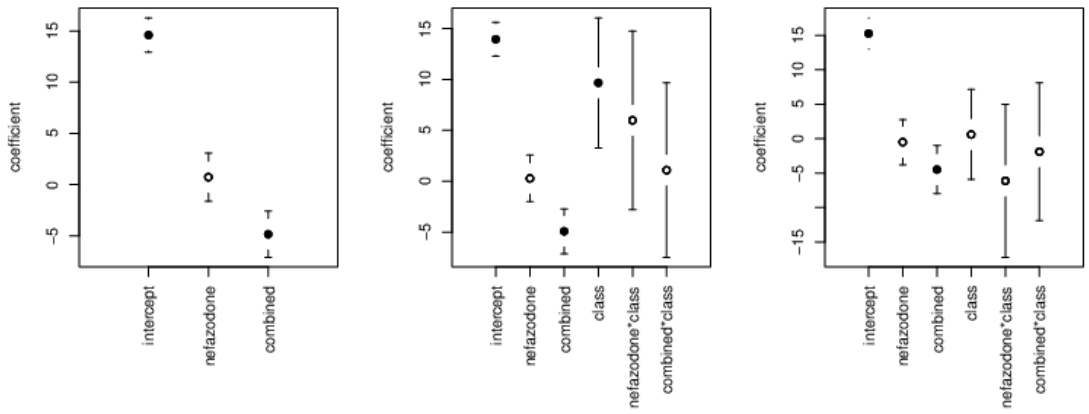

Fig. 4: Depression dataset: estimated coefficients (and error bars) in the linear regression models (i) $Y \sim Z$ on the training set, (ii) $Y \sim Z + Z1\{\hat{\omega}_n^T X - \hat{\gamma}_n \geq 0\}$ on the training set, and (iii) $Y \sim Z + Z1\{\hat{\omega}_n^T X - \hat{\gamma}_n \geq 0\}$ on the test set. Filled circles indicate significance at the 0.05 level.

We next fit the same model on the test set to assess the generalizability of the estimated hyperplane in Equation (12). The third panel of Figure 3 shows the coefficients of the model $Y \sim Z + Z1\{\hat{\omega}_n^T X - \hat{\gamma}_n \geq 0\} + U$ on the test set. The sign of the coefficients in the test set agree with that of the coefficients fitted on the training set. The p-values are not as significant as in the training set but still quite low. Thus we can comfortably conclude that we have found two subgroups whose treatment responses are different.

5·2. *Nefazodone-CBASP trial*

The Nefazodone-CBASP trial compared three different treatments for patients suffering chronic depression. Patients with non-psychotic chronic major depressive disorder (MDD) were randomized to either 1) the Nefazodone drug, 2) cognitive behavioral-analysis system of psychotherapy (CBASP), or 3) a combination of the two. The primary outcome measurement used in assessing the efficacy of the treatments is the score on the 24-item Hamilton Rating Scale for Depression (HRSD). Lower HRSD is desirable. For the detailed study design, Keller et al. (2000) can be consulted.

The data (courtesy of John Rush) consists of 570 patients which we split into a training set ($n = 399$) and a test set ($n = 171$). Let Y in Model (1) be the HRSD score, assumed to be Gaussian distributed. Psychotherapy is taken to be the baseline treatment and coded $Z = (0, 0)$, Nefazodone is coded $Z = (1, 0)$ and the combined treatment $Z = (0, 1)$. Since the patients were enrolled in a randomized controlled trial, we take U in Model (1) to be empty. We considered 25 pretreatment variables for the feature vector X.

The original analysis in Keller et al. (2000) indicates that the combination of the drug and psychotherapy is significantly more efficacious than either treatment alone. Fitting the linear model $Y \sim Z$, we confirm this is indeed true, the result of the fit is displayed in the first panel of Figure 4. Next, we apply the proposed methodology to discover subgroups of patients with possibly different responses to treatment. The coefficients of the linear model $Y \sim Z + Z1\{\hat{\omega}_n^T X - \hat{\gamma}_n \geq 0\}$ on the training data are displayed in the second panel of Figure

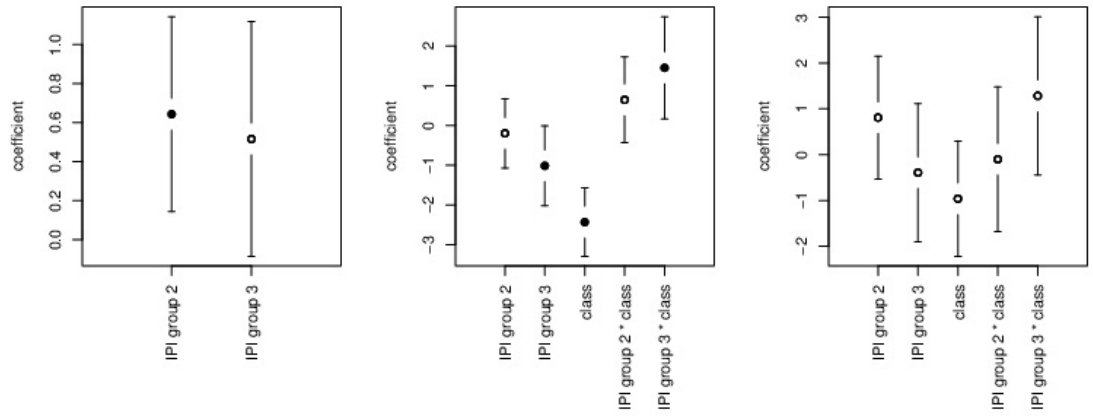

Fig. 5: DLBCL dataset: estimated coefficients (and error bars) in Cox models (i) $Y \sim Z$ on the training set, (ii) $Y \sim Z + Z1\{\hat{\omega}_n^T X - \hat{\gamma}_n \geq 0\}$ on the training set, and (iii) $Y \sim Z + Z1\{\hat{\omega}_n^T X - \gamma \geq 0\}$ on the test set. Filled circles indicate significance at the 0.05 level.

4. We see that neither of the interaction terms are significant. This suggests that the combination treatment may indeed be superior to psychotherapy for all patients.

It also turns out that the estimated hyperplane divides the training data such that the overwhelming majority (93%) fall in one subgroup. This is strong evidence that there is only one subgroup. Indeed, we see from the last panel of Figure 4 that the class variable and the interaction terms are not significant in the test set.

We also performed the analysis using Nefazodone as the baseline which also showed the combination treatment to be the superior treatment for all subjects. Other independent analysis have also drawn similar conclusions, see Qian & Murphy (2011), Gunter et al. (2011), and Zhao et al. (2012).

5·3. *Diffuse large B-cell lymphoma*

Diffuse large B-cell lymphoma (DLBCL) is a cancer of white blood cells and the most common type of non-Hodgkin lymphoma among adults. Here we analyze a survival dataset collected on 240 patients with DLBCL of which there are 138 patient deaths at follow-up (42% censoring). The outcome of interest Y is survival. Our goal is to measure the association between survival and the International Prognostic Index (IPI), a well-established predictor of the survival of DLBCL patients. It has been noted in the literature that the survival outcome in patients who have identical IPI values can vary considerably. As such, we expect there to be subgroups of patients whose IPI is differentially connected to survival.

We split the data into a training set $(n = 149)$ and a test set (n=73). The censoring percentages are 47% and 34%, respectively. We first fit the Cox model $Y \sim Z$ on the training data. The first panel of Figure 5 shows there is a significant difference between IPI group 2 and IPI group 1 as well between IPI group 3 and IPI group 1.

The proposed methodology is then applied with U set to be empty and X to include four gene expression signatures 1) Germinal center B cell, 2) Lymph node, 3) Proliferation, and 4) MHC class II, and the gene expression of the BMP6 gene. Let $Z = (0, 0)$ for IPI group 1, $Z = (1, 0)$

for IPI group 2, and $Z = (0, 1)$ for IPI group 3. The estimated hyperplane splits the subjects in the training set into two groups at roughly a 75–25 split. The coefficients of the model $Y \sim Z + Z1\{\hat{\omega}_n^T X - \hat{\gamma}_n \geq 0\}$ on the training set is displayed in the second panel of Figure 5. We see that there is no significant difference between IPI group 2 and IPI group 1. Within class 0, IPI group 3 experiences significantly better survival than IPI group 1. In class 1, the opposite is true.

We now fit the same model on the test set. The coefficients of the model $Y \sim Z + Z1\{\hat{\omega}_n^T X - \hat{\gamma}_n \geq 0\}$ on the test set are summarized in the third panel of Figure 5. The magnitude and sign of the estimated coefficients in the training set and the test set are similar, with the exception of the IPI group 2 variable and its interaction with the class variable. This is not surprising since the p-values for these variables were not significant in the training set. Thus we can conclude that patients in IPI subgroup 3 have different survival depending on which class they are in. In addition, there is no significant survival difference between IPI group 2 and IPI group 1.

6. DISCUSSION

In this article we introduced a model to study treatment effect heterogeneity and an effective methodology for its estimation. Several improvements and extensions are important to consider. An obvious extension of the method is to ultra-high dimensional X features. One idea is to add a penalty term on the coefficients of ω to the log likelihood. Alternatively it is also feasible to construct the sieve in such a way that the candidate directions therein are already sparse.

Performing inference for ω and γ is also important and challenging. To guard against overfitting in the data analysis examples, we applied the separating hyperplane to an independently held out test set and assessed the significance of the interaction effects. This in some sense wastes part of the data, as we have to set aside a test set to perform model diagnostics. A theoretical basis for performing inference would be preferable. Establishing the weak convergence of the estimator will lead to confidence intervals for instance.

Another interesting future research area is to build variable selection into the methodology. Currently as each variable is weighted in the final classifier, these weights can be loosely interpreted to indicate a variable's importance. This should be assessed in a more rigorous way however and work in this direction could help identify predictive biomarkers, i.e. a marker which can be used to identify subpopulations of patients who are most likely to respond to a given therapy.

ACKNOWLEDGEMENTS

We thank Professor Andrew Nobel for suggestions given during the first author's dissertation proposal. The first author was supported by the National Science Foundation Graduate Research Fellowship. Part of the paper was completed at the University of Tromso, Norway where the first author was supported by the National Science Foundation and the Norwegian Research Council for a research visit in summer 2013. The second author was funded in part by US NIH grant P01 CA142538

APPENDIX 1

In this section, we treat the case where the variables X and Z overlap. Let $\hat{v}_n^1(\omega)$ and $\hat{v}_n^2(\omega)$ be the eigenvectors corresponding to the two largest eigenvalues of the weighted covariance matrix in Equation (2).

Let L_n' be the log likelihood of the data under the following modification to the linear predictor component

$$\eta_i = (\beta_{1,0} + (\beta_{2,0} - \beta_{1,0})\frac{\exp\{\omega_0^T x_i - \gamma_0 \geq 0\}}{1 + \exp\{\omega_0^T x_i - \gamma_0 \geq 0\}})^T z_i + \delta_0^T u_i. \tag{A1}$$

As with the log likelihood L_n, the modified log likelihood L_n' can also be parametrized solely in terms of the direction vector. Now consider the following optimization problem

$$\arg\max_{c_1,c_2} -L_n'(c_1\hat{v}_n^1(\omega) + c_2\hat{v}_n^2(\omega)). \tag{A2}$$

For each ω in the simple sieve, let the solution to the above minimization problem be the boosted direction. This optimization problem can be solved using the fminsearch function in Matlab which implements the Nelder-Mead algorithm for multidimensional unconstrained nonlinear minimization.

APPENDIX 2

LEMMA 1. *Under Conditions* 1–6*, there exists a sequence* ω_n *in* $\hat{\Omega}_n$ *that converges to* ω_0.

Proof. For simplicity, let us consider the case where there is no overlap between X and Z. Recall the definition of $\hat{\nu}_n(\omega)$. It is the largest eigenvector of the weighted covariance matrix $\hat{V}_n(\omega)$ where ω is a direction in the simple sieve.

Let $p_h(\omega) = E1\{Y, \omega^T X \in I_h\}$ be the theoretical proportions in each slice. Let $Z = \Sigma_{xx}^{-1}[X - EX]$ be the standardized covariate and $m_h(\omega) = E[E(Z|Y)|Y, \omega^T X \in I_h]$ be the theoretical mean in each slice. Define the matrix

$$V(\omega) = \sum_{h=1}^{H} p_h(\omega) m_h m_h(\omega)'.$$

It is easy to see $\hat{V}_n(\omega)$ is uniformly consistent for $V(\omega, \gamma)$ over $(\omega, \gamma) \in \mathbb{S}^d \times [a, b]$. By Corollary 3.1 in Li (1991) which uses Condition 4, the largest eigenvector of $V(\omega)$ falls in the linear space generated by $\omega_0 \Sigma_{xx}^{1/2}$. Since $\hat{\nu}_n(\omega)$ is consistent for the largest eigenvector of $V(\omega)$ and $\hat{\Sigma}_{xx}$ is consistent for Σ_{xx}, we have $\hat{\nu}_n(\omega)^T \hat{\Sigma}_{xx}^{-1/2} \to \omega_0$ uniformly over $\omega \in \mathbb{S}^d$.

REFERENCES

CHIPMAN, H. A., GEORGE, E. I. & MCCULLOCH, R. E. (2010). Bart: Bayesian additive regression trees. *Annals of Applied Statistics* **4**, 266–298.

GRENANDER, U. (1981). *Abstract Inference*. New York: Wiley.

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

HALL, P., MARRON, J. S. & NEEMAN, A. (2005). Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society Series B* **67**, 427–444.

HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, J. (2001). *The Elements of Statistical Learning*. New York: Springer-Verlag.

IMAI, K. & STRAUSS, A. (2011). Estimation of heterogeneous treatment effects from randomized experiments, with application to the optimal planning of the get-out-the-vote campaign. *Political Analysis* **19**, 1–19.

KELLER, M. B., MCCULLOUGH, J. P., KLEIN, D. N., ARNOW, B., DUNNER, D. L., GELENBERG, A. J., MARKOWITZ, J. C., NEMEROFF, C. B., RUSSELL, J. M., THASE, M. E., TRIVEDI, M. H. & ZAJECKA, J. (2000). A comparison of nefazodone, the cognitive behavioral-analysis system of psychotherapy, and their combination for the treatment of chronic depression. *New England journal of medicine* **342**, 1462–70.

KOSOROK, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. New York: Springer, 1st ed.

LAGAKOS, S. W. (2006). The challenge of subgroup analyses: Reporting without distorting. *New England Journal of Medicine* **354**, 1667–1669.

LEBLANC, M. & KOOPERBERG, C. (2010). Boosting predictions of treatment success. *Proceedings of the National Academy of Science* **107**, 13559–13560.

LI, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* **86**, 316– 327.

NELDER, J. A. & WEDDERBURN, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A* **135**, 370–384.

QIAN, M. & MURPHY, S. A. (2011). Performance guarantee for individualized treatment rules. *Annals of Statistics* **39**, 1180–1210.

ROTHWELL, P. M. (2005). Subgroup analysis in randomized controlled trials: importance, indications, and interpretation. *The Lancet* **365**, 176–186.

SU, X., TSAI, C.-L., WANG, H., NICKERSON, D. M. & LI, B. (2009). Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research* **10**, 141–158.

WEI, S. & KOSOROK, M. R. (2013). Latent supervised learning. *Journal of The American Statistical Association* **In press**.

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

ZHU, R. & KOSOROK, M. R. (2012). Recursively imputed survival trees. *Journal of The American Statistical Association* **107**, 331–340.

[*Received. Revised*]