Published in final edited form as:

J Biopharm Stat. 2015 ; 25(4): 781–794. doi:10.1080/10543406.2014.920861.

Nonparametric Variable Selection for Predictive Models and Subpopulations in Clinical Trials

Jingyi Zhu and Jun Xie*

Department of Statistics, Purdue University, 250 N. University Street, West Lafayette, IN 47907

Abstract

Most clinical trials have heterogeneous treatment effect among patient individuals. It is desirable to identify a patient subpopulation, which has a stronger treatment effect than the rest of patients, so that researchers will be able to determine who will benefit the most or the least from the treatment and design treatment strategies accordingly. This paper develops a nonparametric method for predicting clinical response and identifying subpopulations. The method first selects predictors using kernel-based local regression and a forward procedure via F -tests. It then defines subpopulations with enhanced treatment effects based on the selected predictors and the nonparametric model of the clinical response. Simulation examples and a pharmacogenomics study of bortezomib in multiple myeloma demonstrate the proposed method and show favorable performances compared to other existing methods. The proposed method provides an alternative way to define subpopulations and is not limited by parametric models and their possible misspecification for the clinical response.

Keywords

Clinical trial; Nonparametric local regression; Variable selection; Subpopulation

1 Introduction

In most clinical trials, there is much heterogeneity among individual outcomes and the treatment effect may not be the same on all of the patients. If we could determine which patients will respond better to the treatment, ideally ahead of time but also possibly soon after the treatment is administered, the development and subsequent utilization of the therapy would be dramatically improved. It is substantively interesting but challenging to identify such patient subpopulations that will derive a more pronounced benefit from the active treatment than the rest of patients. More specifically, consider a clinical trial with patients' drug response and a large number of potential predictors, such as genetic information, clinical features, and demographic information. Our objective is to develop statistical methods to select predictive covariates and consequently use them to define subpopulations with enhanced treatment effects.

*To whom correspondence should be addressed..

With a large number of potential predictors, variable selection is often the first step in developing predictive models. There are many reasons for focusing on a subset: the desire to glean important biological insight, operational considerations for how this information can be utilized in subsequent development of the novel therapy, and the fact that a simple model has a better chance to hold in a new trial and lends itself more easily to validation efforts. Suppose that we have a clinical response variable Y , and a set of p predictors $X_1, \dots, X_p$. The problem of variable selection arises when we want to model the relationship between Y and a subset of $X_1, \dots, X_p$. Variable selection methods have mainly been developed for linear models. Heuristic variable selection procedures are often employed, for example, forward selection, backward elimination, and stepwise selection. A number of well-known selection criteria have been used in heuristic procedures for linear models, including AIC (Akaike, 1973), BIC (Schwarz, 1978), and Mallows' C_p (Mallows, 1973). More recently, regularization methods have also been used as variable selection approaches, for example, LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001), and elastic net (Zou and Hastie, 2005). The regularization methods conduct variable selection mainly in linear or parametric models. However, the linear relationship between Y and one or more predictors $\mathbf{X}$ is often too simple to be proper in complicated clinical data analysis. The traditional linear model does not work for detecting nonlinear patterns of the relationship between Y and $\mathbf{X}$. The limitations of the linear model and other parametric statistical approaches motivate our use of nonparametric methods to model the relationship between Y and $\mathbf{X}$.

The identification of subpopulation with enhanced treatment effect is recently a popular topic in clinical practice and medical research. It relates to the efforts in discovering patient-specific treatment strategy, or personalized medicine. Crump *et al.* (2006) conducted statistical tests for the heterogeneity of treatment effects across pre-specified patient subpopulations. Moineddin *et al.* (2008) proposed a multi-level random-effect model to identify subpopulations from patient baseline characteristics. Ruberg *et al.* (2010) and Foster *et al.* (2011) proposed to use a CART (Classification and Regression Tree) approach to select predictors and consequently to use their cut-off values to define subpopulations of patients. The tree splitting idea in CART was further explored by Lipkovich *et al.* (2011). Zhang *et al.* (2012) used a regression model for the expected clinical response conditional on treatment and covariates. The parametric regression model defines an optimal treatment regime, or equivalently, subpopulations. In this paper we propose a nonparametric method to model the expected clinical response conditional on a small set of selected covariates. We intend to relax any parametric model assumptions, hence the method is not limited by misspecification of the regression model for the clinical response.

Our method is a combination of forward variable selection and nonparametric local regression. Forward variable selection is merely a heuristic procedure, but it is easy to implement and can obtain a subset of predictors with reasonably small prediction errors. Meanwhile, the use of nonparametric local regression has the major advantage that we make fewer assumptions about the functional form of the model for the clinical response. After we developed the method, we found that a similar idea had been suggested by Storlie and Helton (2007) but under the context of reliability analysis. Our focuses are on clinical predictive models and subpopulations. We demonstrate our variable selection approach

using data simulated from a simplified yet realistic clinical trial. Our method has high accuracy in test data sets and performs comparably to the CART method. We also implement our method in a pharmacogenomics study of bortezomib in multiple myeloma and compare it with another existing method of linear predictive model. In the bortezomib example, our nonparametric model with three predictors achieves the same prediction power as a linear model with a large number of predictors (over 100).

The structure of the paper is organized as follows: In Section 2, we define the framework of subpopulation. Our idea is in analogy with that of Zhang *et al.* (2012) but generalizes their parametric regression model to a nonparametric model. In Section 3, we present the detailed method of nonparametric variable selection. In Section 4, we show that the nonparametric predictive models provide consistent estimators of treatment effects when the smoothing parameters are appropriately chosen. In Section 5, we show results of a simulation study and the application to the bortezomib data set. We conclude with discussions in Section 6.

2 Subpopulation with enhanced treatment effect

Consider a randomized clinical trial. Each patient receives either an active treatment or placebo at random. Let $\mathbf{X} = (X_1, \dots, X_p)$ denote a vector of p predictors (genetic biomarkers, demographics, etc.), and Y denote a clinical response. In principle, the clinical response Y has two components, Y_{trt} and $Y_{control}$, where Y_{trt} is the clinical response if a patient receives the active treatment, and $Y_{control}$ is the clinical response if a patient receives the placebo. We consider two types of treatment effects: the global treatment effect and the conditional treatment effect. The global treatment effect is $E(Y_{trt} - Y_{control})$, where $E(\cdot)$ denotes the expectation of Y . The conditional treatment effect is $E(Y_{trt} - Y_{control}|\mathbf{X})$, where $E(\cdot|\mathbf{X})$ denotes the conditional expectation of Y given $\mathbf{X}$.

Denote the sample space of $\mathbf{X}$ as $\mathcal{X}$. A partition of $\mathcal{X}$ defines subpopulations of patients. A subpopulation with an enhanced treatment effect is defined as a patient group with covariates values in a subset of $\mathcal{X}$ that has a larger conditional treatment effect than the global treatment effect. Formally, the subpopulation can be represented as

$$S = \left\{ \mathbf{X} \in \mathcal{X} : E(Y_{trt} - Y_{control}|\mathbf{X}) > E(Y_{trt} - Y_{control}) \right\}.$$

This subpopulation definition is similar to the idea of optimal treatment regime used in Zhang *et al.* (2012) except that Zhang *et al.* (2012) considered $E(Y_{trt} - Y_{control}|\mathbf{X})$ versus 0. We prefer to compare $E(Y_{trt} - Y_{control}|\mathbf{X})$ to the global treatment effect $E(Y_{trt} - Y_{control})$. In fact, without loss of generality, we assume $E(Y_{trt} - Y_{control}) = 0$ throughout the paper. Therefore, our definition of subpopulation is more rigorous. If we indeed have a clinical trial with $E(Y_{trt} - Y_{control}) < 0$, we will replace the inequality by $E(Y_{trt} - Y_{control}|\mathbf{X}) > 0$ in the above definition and modify our implementation procedure accordingly.

Given a data set, i.e., a randomized clinical trial with n patients, we estimate $E(Y_{trt} - Y_{control})$ by the difference of the sample means between the treatment and the control groups. We estimate the conditional expectations $E(Y_{trt} - Y_{control}|\mathbf{X})$ by the difference of two

nonparametric functions of $\mathbf{X}$, one for $E(Y_{trt}|\mathbf{X})$ and the other for $E(Y_{control}|\mathbf{X})$. More specifically, we have

$$S = \left\{ \mathbf{X} \in \mathbb{R}^d : \hat{g}_{trt}(\mathbf{X}) - \hat{g}_{control}(\mathbf{X}) > \bar{Y}_{trt} - \bar{Y}_{control} \right\},$$

where $\hat{g}_{trt}(\cdot)$ denotes the nonparametric estimate of $E(Y_{trt}|\mathbf{X})$ in the treatment group, and $\hat{g}_{control}(\cdot)$ denotes the nonparametric estimate of $E(Y_{control}|\mathbf{X})$ in the control group, $\bar{Y}_{trt}$ and $\bar{Y}_{control}$ are the sample means of Y in the treatment and the control groups, respectively. Note that we are modeling the treatment and the control groups separately, instead of considering a combined response model with covariates $\mathbf{X}$, a treatment variable, and their interactions. In fact, once we relax the parametric model assumption, the model for the treatment group and that for the control group are arbitrary functions and in different functional forms. Therefore, the interactions of treatment and covariates on treatment effects are automatically incorporated. In the following, we develop a nonparametric variable selection approach to estimate $E(Y_{trt}|\mathbf{X})$ and $E(Y_{control}|\mathbf{X})$, which provide predictive models for the response and also derive the subpopulation S .

3 Variable selection via LOESS

We first describe the method of variable selection in the context of nonparametric models. Let y_i , $i = 1, \dots, n$, denote n measurements of the response variable Y . Let x_{ij} , $i = 1, \dots, n$ and $j = 1, \dots, p$, denote n observations of p potential predictors. Without loss of generality we assume that the data of each predictor, $x_{1j}, \dots, x_{nj}$, have been normalized so that all predictors have the same scale, for example, with mean 0 and standard deviation 1.

3.1 Estimation via LOESS given a fixed set of predictors

Assume a given subset of multiple predictors, $(X_1, X_2, \dots, X_d)$, where d is the number of predictors in the model and is often limited to four. Let $\mathbf{x}_i = (x_{i1}, \dots, x_{id})$, $i = 1, \dots, n$, be n measurements of the selected predictors. Assume a model of the form

$$y_i = g(\mathbf{x}_i) + \epsilon_i,$$

where $g(\cdot)$ is an unknown smooth function and ϵ_i 's are i.i.d. error terms with mean 0 and finite variance σ^2 and independent of $\mathbf{x}_i$. The model assumption implies that $E(Y|\mathbf{X}) = g(\mathbf{X})$. In our previous subpopulation discussions we have two such nonparametric models, one for Y_{trt} and the other for $Y_{control}$. Cleveland (1979) proposed the locally weighted scatterplot smoothing (LOWESS), a local regression method for a response variable Y on a single predictor X . It was further generalized to multivariate predictors, known as LOESS (Cleveland and Devlin, 1988), to model the relationship between a response variable Y and multiple predictors. With LOESS we can estimate a large class of smooth functions without being restricted to a specific class of parametric functions. The estimate of g at a single point $\mathbf{x}$ uses all neighbors around $\mathbf{x}$, where the neighbors are decided by a span parameter. Each neighbor is then weighted according to its distance from $\mathbf{x}$ by a kernel function. A linear or quadratic function of $\mathbf{x}$ is fitted to Y using weighted least squares.

3.2 Forward selection criterion

We want to conduct forward variable selection using LOESS. Predictors are added to the model one by one if they are statistically significant. The local regression method implies that

$$\hat{g}(\mathbf{x}) = \sum_{i=1}^n l_i(\mathbf{x}) y_i.$$

That is, the LOESS estimate $\hat{g}(\mathbf{x})$ is a linear combination of the observed response y_i where $l_i(\mathbf{x})$ depends on the observed predictor values in a neighborhood of $\mathbf{x}$, but not on y_i . Let $\hat{y}_i = \hat{g}(\mathbf{x}_i)$ be the fitted values, $\hat{\epsilon}_i = y_i - \hat{y}_i$ be the residuals, and denote $y = (y_1, \dots, y_n)^t$, $\hat{y} = (\hat{y}_1, \dots, \hat{y}_n)^t$, $\hat{\epsilon} = (\hat{\epsilon}_1, \dots, \hat{\epsilon}_n)^t$. Since each $\hat{y}_i$ is a linear combination of y we have $\hat{y} = Ly$ where L is an $n \times n$ matrix and $\hat{\epsilon} = (I - L)y$ where I is the $n \times n$ identity matrix. Suppose that one predictor has been selected and $L_1 y$ is the vector of its fitted values. We consider adding a second predictor into the model and let $L_2 y$ be the fitted values using the two predictors. We want to test a null hypothesis H_0 of one predictor for y against an alternative hypothesis H_a of two predictors using a nonparametric F -test. More specifically, let $y^t R_1 y = y^t (I - L_1)(I - L_1)^t y$ and $y^t R_2 y = y^t (I - L_2)(I - L_2)^t y$ be the residual sum of squares of the two fits. Under H_0 , we have the following test statistic

$$\hat{F} = \frac{(y^t R_1 y - y^t R_2 y) / v_1}{(y^t R_2 y) / \delta_1},$$

which approximately follows an F distribution with the degrees of freedom v^2/v_2 and δ^2/δ_2 , where $v_1 = \text{trace}(R_1 - R_2)$, $v_2 = \text{trace}(R_1 - R_2)^2$, $\delta_1 = \text{trace}(R_2)$, and $\delta_2 = \text{trace}(R_2)^2$ (Cleveland and Devlin, 1988). The degrees of freedom are obtained by generalizing the F -test statistic of parametric models to a nonparametric case. Another definition of the test statistic was used in Storlie and Helton (2007), with different degrees of freedom defined by v_1 and δ_1 for the F distribution under H_0 . However, our simulations show that Storlie and Helton (2007) test statistic tends to over-select variables, including more predictors in the model than needed.

We select the first predictor by comparing a null hypothesis of a constant model to an alternative hypothesis of one predictor for Y . If the most significant single predictor has a p -value less than a cutoff, e.g., 0.01, it is added into the model. Next we consider adding a second predictor into the model. We conduct F -tests for all possible second predictors. If none of the second predictors are significant, we end with a model with only one predictor. Otherwise, we select the most significant one and extend the model to two predictors. We continue using the nonparametric F -test as the criterion to select significant predictors and this procedure stops if no predictor is found to be significant.

3.3 Smoothing parameter selection

Nonparametric methods, including LOESS, use a smoothing parameter to control potential over-fitting of local regression. The smoothing parameter here is the proportion of the

neighbor points out of all data points that are used to fit $g(\mathbf{x})$ at $\mathbf{x}$. It is referred to as the span parameter α . If α is too small, insufficient data fall within the neighborhood resulting in an over-fitting with large variance. On the other hand, if α is too large the local regression may not fit data well resulting in a fit with large bias. Thus the span parameter must be chosen to compromise the bias-variance trade-off.

Commonly used criteria of selecting span parameters in general nonparametric techniques, e.g., smoothing splines, include AIC and Generalized Cross Validation. In our method, we use an extended version of AIC, known as AIC_C . It was first introduced by Hurvich and Tsai (1989) for a linear model,

$$\begin{aligned} AIC_C &= \log(\hat{\sigma}^2) + \frac{1+d/n}{1-(d+2)/n} \\ &= \log(\hat{\sigma}^2) + 1 + \frac{2(d+1)}{n-d-2}, \end{aligned}$$

where $\hat{\sigma}^2$ is the average of residual sum of squares and d is the number of variables included in the model. Hurvich *et al.* (1998) further generalized AIC_C to the context of nonparametric regression for span parameter selection. The AIC_C score for a local estimate with the span parameter α is

$$\begin{aligned} AIC_C(\alpha) &= \log(\hat{\sigma}^2) + \frac{1+tr(L_\alpha)/n}{1-\{tr(L_\alpha)+2\}/n} \\ &= \log(\hat{\sigma}^2) + 1 + \frac{2\{tr(L_\alpha)+1\}}{n-tr(L_\alpha)-2}, \end{aligned}$$

where $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{g}_\alpha(x_i))^2$ is the estimated error variance and L_α is the estimation matrix L as defined in Section 3.2 but depends on α . It was shown in the simulation study by Hurvich *et al.* (1998) that compared to the criterion of Generalized Cross Validation or AIC, the use of AIC_C avoided the large variability and the tendency to undersmooth. AIC_C is also easy to apply in practice since it is a function of L only through its trace.

Ideally an optimal span parameter is chosen where the AIC_C score is minimized. However, it is found from our study that AIC_C scores have several different patterns. For example, the reduction of AIC_C scores becomes negligible near the upper boundary of the span parameter. Therefore minimizing AIC_C tends to choose larger span parameters which may not be necessary. Here we define a modified criterion that selects an optimal span parameter $\hat{\alpha}$ which is at least 0.2 and satisfies

$$\hat{\alpha} = \min_{\alpha} \left\{ \argmin_{\alpha} AIC_C, \argmax_{\alpha} \Delta AIC_C, \arg_{\alpha} \{\Delta AIC_C = 0\} \right\}. \quad (1)$$

Looking for α that is the root of $AIC_C = 0$ is an alternative criterion to minimize AIC_C . We also study the change of AIC_C , i.e., ΔAIC_C , with large changes being favorable. In implementation, we consider a grid of the span parameter between 0.2 and 0.8, as suggested in Cleveland and Devlin (1988), and with an increment of 0.01. We select the smallest α value in the range of 0.2 to 0.8, which either corresponds to the minimum of AIC_C or the maximum change of AIC_C .

In our variable selection procedure, we first use (1) to select an optimal span parameter for each candidate predictor X_i , $i = 1, \dots, p$. Let $\hat{\alpha}_1$ be the optimal span parameter for the most significant predictor at the first step. In the following steps, the span parameter is chosen to be the maximum value between 0.2 and the power of $\hat{\alpha}_1$, i.e., $\hat{\alpha}_1^d$, where d is the number of predictors selected into the model and is up to 4. This is a simple rule to consider cubic neighborhoods from multi-dimensional predictors and use the same $\hat{\alpha}_1$ value for all predictors added into the model. For instance, assume we have $\hat{\alpha}_1 = 0.8$ and $d = 3$. We will use $\hat{\alpha}_1^3 = 0.512$ to define a neighborhood when we estimate $g(\mathbf{x})$ for any given value $\mathbf{x}$ in the 3-dimensional space $\mathbf{R}^3$. More specifically, we use the amount of $0.512n$ observations around $\mathbf{x}$ to estimate $g(\mathbf{x})$, where Eclidean distance defines the neighbor points around $\mathbf{x}$. This amount of neighbor points will further be weighted by a multivariate kernel function, e.g., a tricube kernel in $\mathbf{R}^3$, according to their Eclidean distance to $\mathbf{x}$. An appropriately chosen span parameter and a small number of predictors, i.e. $d \leq 4$, help prevent overfitting of the nonparametric model.

4 Predictive model for subpopulation

The local regression via LOESS is a kernel-type nonparametric estimator, whose properties have been extensively studied, for example, by Ruppert and Wand (1994) and Fan and Gijbels (1996). Suppose $\mathbf{x} = (x_1, \dots, x_d)^t$ is a point in $\mathbb{R}^d$, where we want to estimate the predictive model $g(\mathbf{x})$. Given the observed data $\mathbf{X}_i = (X_{i1}, \dots, X_{id})^t$, $i = 1, \dots, n$, we obtain a local linear estimate of $g(\mathbf{x})$ with the intercept and slope coefficients $\{\beta_0, \beta_1\}$ as the minimum of

$$\sum_{i=1}^n \left\{ Y_i - \beta_0 - \beta_1^t (\mathbf{X}_i - \mathbf{x}) \right\}^2 K_\alpha(\mathbf{X}_i - \mathbf{x}),$$

where $K_\alpha(\cdot)$ is a kernel function to weight the neighbor data around $\mathbf{x}$. Analogous to LOESS, we use the tricube kernel (Cleveland and Devlin, 1988) in a neighborhood defined by the span parameter α . Following the general theorems of nonparametric kernel estimators (Fan and Gijbels, 1996; Ruppert and Wand, 1994) the LOESS estimator such obtained has a large sample property under regularity conditions. More specifically, with appropriately chosen α 's, the LOESS estimator has both the bias and variance go to zero as the sample size $n \rightarrow \infty$. In other words, the LOESS method provides a consistent estimate.

We fit predictive models for the treatment and the control groups separately after selecting significant predictors for these two groups respectively. Let $\mathbf{X}_{trt}$ be the significant predictors for the treatment group and $\mathbf{X}_{control}$ be the significant predictors for the control group. Since it is possible that either $\mathbf{X}_{trt}$ or $\mathbf{X}_{control}$ is $\emptyset$, we define a subset of the sample space $\mathcal{X}$ in three different ways.

1. $\mathbf{X}_{trt} = \emptyset$, but $\mathbf{X}_{control} \neq \emptyset$.

The subpopulation is defined as

$$S = \left\{ \mathbf{X}_{trt} \in \partial \hat{\Pi} : \hat{g}_{trt}(\mathbf{X}_{trt}) > \bar{Y}_{trt} \right\}.$$

2. $\mathbf{X}_{trt} = \emptyset$, but $\mathbf{X}_{control} \neq \emptyset$.

The subpopulation is defined as

$$S = \left\{ \mathbf{X}_{control} \in \partial \hat{\Pi} : \hat{g}_{control}(\mathbf{X}_{control}) < \bar{Y}_{control} \right\}.$$

3. Both of $\mathbf{X}_{trt}$ and $\mathbf{X}_{control}$ are $\neq \emptyset$.

Let $\mathbf{X} = (\mathbf{X}_{trt}, \mathbf{X}_{control})$. The subpopulation is defined as

$$S = \left\{ \mathbf{X} \in \partial \hat{\Pi} : \hat{g}_{trt}(\mathbf{X}_{trt}) - \hat{g}(\mathbf{X}_{control}) > \bar{Y}_{trt} - \bar{Y}_{control} \right\}.$$

Patients, whose covariates values are within the subset S such defined, correspond to a subpopulation with enhanced treatment effects. These patients will benefit the most from the treatment and we should design a treatment regime to specifically assign them the treatment.

5 Results

5.1 Simulation examples

We use data simulated from a simplified yet realistic clinical trial to demonstrate our procedure. The data is from an open challenge of data analysis posted online by Eli Lilly & Company's statistical group. It consists of $n = 322$ patients, each with a response Y for the clinical outcome and 99 continuous predictors $X_1, X_2, \dots, X_{99}$. There is another treatment index variable indicating whether a patient received an active treatment or placebo, as randomly assigned at the beginning of the trial. There are two data sets generated by two different response models. In the first data set Y only depends on the predictor X_{19} and the ideal subpopulation with enhanced treatment effect is $X_{19} > -0.22$. In the second data set, Y depends on two predictors X_{30} and X_{43} and the ideal subpopulation with enhanced treatment effect is $X_{30} > -0.42$ and $X_{43} > -0.29$. We will use the truth of subpopulations from these simulations to assess our proposed method and compare it with an existing method.

In our exploratory data analysis we find that some predictors have extreme values which may dramatically affect the model fitting. Therefore, we apply a *5IQR* rule (5 times interquartile range) to detect possible outliers and exclude them from the following analysis. For the first data set we identify X_{19} as the most significant predictor in the treatment group but no significant predictor in the control group. Figure 1 shows that X_{19} can be clearly identified from the LOESS fit with a very small p-value, less than 10^{-5} , but it becomes insignificant in a linear model for Y . In addition, if we consider both the treatment and the control arms and fit a combined linear model with X_{19} and the treatment variable and their interaction, then only a main treatment effect is significant but not X_{19} nor the interaction.

These results indicate that a linear model is not able to identify any significant predictors for the clinical response.

Applying our method in the first data set, we select one predictor, X_{19} , for the treatment group. The span parameter is chosen as $\alpha = 0.2$ according to the AIC_C criterion (1). No significant predictor is selected for the model of the control group. Figure 2 shows a plot of nonparametric predictive model for the response in the treatment group. The curve represents the predicted LOESS function versus X_{19} , with the dashed horizontal line as the sample mean of Y in the treatment group. Whenever the LOESS curve crosses the horizontal line with large values, it defines a subpopulation with enhanced treatment effect.

As shown in Figure 2, we identify subpopulations with enhanced treatment effects according

to $S = \left\{ \mathbf{X}_{trt} \in \mathbb{R}^p : \hat{g}_{trt}(\mathbf{X}_{trt}) > \bar{Y}_{trt} \right\}$. To stabilize the prediction results, we further apply a refinement procedure to validate or discard the identified subpopulations. Specifically, in this example we first construct a number of nonoverlapping subpopulation intervals. For each nonoverlapping interval, if it contains a small number of observations (less than 4) or the length of the interval is small (less than 0.1), we discard the corresponding interval. Otherwise we retain this interval. Consequently, the subpopulation with enhanced treatment effect identified by our procedure is $X_{19} \in [-0.23, 0.49]$ and $X_{19} \in [0.63, 10.85]$ for the first data set. A careful examination of Figure 2 actually indicates that the fitted curve is above the reference line for almost all values of $X_{19} > -0.23$. Therefore, we may use the whole set $X_{19} > -0.23$ as the subpopulation.

For the second data set we identify X_{30} as the most significant predictor with a p-value 3.67×10^{-5} for the treatment group at the first step of variable selection. The span parameter is chosen as $\alpha = 0.5$ according to the AIC_C criterion (1). The predictor X_{43} has a p-value 0.00212 at the first step of variable selection but it becomes insignificant with a p-value 0.23 at the second step of variable selection, after X_{30} is already in the model. Therefore, we only select one predictor X_{30} for the treatment group. No significant predictor is selected for the model of the control group. We identify a subpopulation as $X_{30} > -0.37$.

For these simulated studies, we know the ideal subpopulations hence use the truth to assess our method and compare it with the CART approach. Table 1 shows the prediction errors for the first data set, where the identified subpopulation of $X_{19} \in [-0.23, 0.49]$ and $X_{19} \in [0.63, 10.85]$ is compared to the ideal subpopulation defined by $X_{19} > -0.22$. We have an overall accuracy of 95% (Sensitivity=94%, Specificity=95%). If we use the whole set $X_{19} > -0.23$ to define the subpopulation, as indicated in Figure 2, then we obtain an overall accuracy of 98% (Sensitivity=100%, Specificity=95%). These results are comparable to that of CART, as provided online with the simulation data. Table 2 shows the prediction errors for the second data set. The identified subpopulation by our procedure is compared to the ideal subpopulation defined by $X_{30} > -0.42$ and $X_{43} > -0.29$. Table 2 also reports the prediction errors of the CART method. Although we only select one predictor X_{30} for the second data set, the performance of our procedure is promising with an overall accuracy of 82% (Sensitivity=94%, Specificity=70%), whereas the CART model has an overall accuracy of 83% (Sens=100%, Spec=66%). Our nonparametric method with few selected variables

provides good results of subpopulations in these examples, compared to both the truth and a competing method.

5.2 Application in a pharmacogenomics example

We implement our method in a pharmacogenomics study of bortezomib in multiple myeloma (Mulligan *et al.*, 2007). Multiple myeloma is an incurable malignancy and bortezomib is the first therapeutic proteasome inhibitor tested in humans for treating relapsed multiple myeloma. As the new active agent bortezomib is a therapeutic choice in addition to the standard chemotherapy, there is a need to reliably identify the patient population that will mostly benefit from the therapy. The data set of this study is available at Gene Expression Omnibus (GSE9782). There are four clinical trials conducted in multiple centers in the United States, Canada, Europe, and Israel from June 2002 to October 2003, denoted as trial 024, trial 025, trial 039, and trial 040, with a total of 264 patients. The clinical outcome is a five-level variable denoting a patient response after a therapy, ranging from progressive disease, no change, minimal response, partial response to complete response. The potential predictors include 44,928 gene expression values.

In the clinical trials, most patients (169) received the new treatment bortezomib but only 70 patients in trial 039 received chemotherapy as controls. Therefore, in this study we focus on analysing only the treatment group, i.e., the novel therapy bortezomib. The goal here is to construct predictive models for the clinical response in the treatment group and identify patient subpopulations who respond the best to the treatment, instead of comparing the treatment and the control groups.

In the exploratory study of Mulligan *et al.* (2007), the five-category clinical response was simplified to two levels: progressive disease (PD) and response (R), excluding no change (NC) patients. Since the data is high dimensional with 44,928 gene expression values but 264 patients, they applied a two-stage gene filtering method in which only the 9200 gene probe sets with the strongest between-sample variance relative to their in-sample replicate variance were retained. Among the 9200 genes only the top 100 differentially expressed genes with respect to clinical response (PD vs. R) by *t*-tests were used as predictors in the predictive model. The data were also divided into a training data set for trials 025 and 040, and a test data set for trial 039. Trial 024 was not used in the analysis due to a very small number of patients (7) with evaluable responses. A linear predictor classifier (Wright *et al.*, 2003) was developed on the training data and was used to classify each patient to be either PD or R in the test data. The classifier was based on a linear combination of the 100 predictors with each being weighted by its *t*-test score.

In our procedure we use the original five-level clinical responses as the response variable Y and encode the ordinal categories by values 1-5 such that progressive disease (PD) is coded as 1, no change (NC) is coded as 2, minimal response (MR) is coded as 3, partial response (PR) is coded as 4, and complete response (CR) is coded as 5. We use the same gene filtering method of Mulligan *et al.* (2007) to keep only the top 100 differentially expressed genes for further analysis. We also divide the data into a training data for trials 025 and 040 and a test data for trial 039. There are 91 observations in the training set and 71 in the test set.

We first conduct the nonparametric forward variable selection on the training data to identify a small number of significant genes. Three genes are selected by the forward variable selection when we set the significance level at 0.05. At the first step, gene 219233_s_at is selected with p-value 8.83×10^{-6} from the F -test. At the second step, gene 212240_s_at is selected with p-value 0.0076 from the F -test. At the third step, gene 200017_at is selected with the corresponding p-value of the F -test to be 0.042. Gene 219233_s_at, known as gasdermin B, may play a role as secretory or metabolic product involved in secretory pathway. Gene 212240_s_at, known as phosphoinositide 3-kinase regulatory subunit 1, has several important biological functions and is necessary for the insulin-stimulated increase in glucose uptake and glycogen synthesis in insulin-sensitive tissues. Gene 200017_at, known as ribosomal protein S27a, is a component of the 40S subunit of the ribosome. Identification of these genes provides a hint for further studies in myeloma and the treatment effect of bortezomib. They will also be used to define subpopulations that benefit from bortezomib.

A predictive model with the three genes is fitted by LOESS. The span parameter is chosen to be 0.8 at the first step of the variable selection according to AIC_C . The span parameter is about 0.6, as the square of 0.8, at the second step and 0.5 ($0.8^3 = 0.512$) at the third step. The predictive model is used to predict patient response in the test data. If the predicted clinical response value is less than or equal to 2 it is classified as nonresponse (NR), which includes both PD and NC, otherwise it is classified as response (R), which includes MR, PR and CR.

A subpopulation can be defined as $S = \{\mathbf{X}_{trt} \in \mathcal{X} : \hat{g}_{trt}(\mathbf{X}_{trt}) > 2\}$, which corresponds to the patient subgroup that responds to the bortezomib treatment.

As demonstration, Figure 3 shows two plots of the nonparametric LOESS fit for patient response in the test data, one projected on the two-dimensional space consisting of the first predictor 219233_s_at and the response Y , the other projected on the three-dimensional space consisting of the first two predictors, 219233_s_at and 212240_s_at, and Y . In the two-dimensional plot, the patients who are responders are labelled by circles and the patients who are non-responders (NC and PD) are labelled by crosses. The dashed horizontal line displays the cutoff value 2. The patient subpopulation is defined to be $\{\mathbf{X} : \hat{g}(\mathbf{X}) > 2\}$ where $\hat{g}(\cdot)$ is the predictive LOESS function. We compare the prediction result of PD vs. R based on our predictive model on the test data with the result from Mulligan *et al.* (2007) as shown in Table 3. In summary, we identify 37 out of the 38 patients who are responders to the bortezomib treatment and only one patient, who is a responder to the treatment, is incorrectly classified as progressive disease (PD). Two out of the 15 patients who have PD to the treatment are correctly classified, but the other 13 patients are incorrectly classified as responders to the treatment. The overall accuracy of our prediction result is 74%, which is comparable to 75% of Mulligan *et al.* (2007)'s result. Thus, with a nonparametric model a small number of predictors can achieve the same prediction power as a linear model with a much larger number of predictors (100). Applying to a new patient our model will predict whether or not it has a more pronounced benefit from the bortezomib treatment.

6 Discussion

In this paper we perform a nonparametric forward variable selection procedure to identify significant predictive covariates among a large set of potential predictors which may include standard clinical and laboratory features and whole genome gene expression measurements. Forward variable selection is merely a heuristic procedure but it has the advantage of easy implementation in practice. It is not limited by high dimensional predictors because it starts with a constant model with only an intercept and will end up with selecting a small number of significant predictors. The local regression method, LOESS, is computationally efficient, especially when we consider a small number of significant predictors (4). In practice we recommend checking for outliers before doing any analysis. We also need to pay attention to fitting a local regression in the boundary of data. While local linear regression has been shown to provide a simple and effective way of modeling slopes in the boundary region and reduce bias compared with other kernel estimates (e.g., Nadaraya-Watson estimate) (Hastie and Loader, 1993), the mean squared error of the local linear regression may still be big when data are sparse and of high curvature in the boundary. Fortunately, the boundary effect is not severe in our examples. In general, visualization of the data and their LOESS fitting in the boundary is always recommended.

While variable selection helps to build a predictive model based on a small number of significant predictors, an alternative approach is to consider that interesting features of high-dimensional data are retrievable from low-dimension projections. Exploring dimension reduction techniques in analysis of high dimensional but low sample size data is one of our future works.

Acknowledgement

The authors thank the Editor and the anonymous referees for their comments and suggestions that helped enhance the paper.

References

- Akaike, H. Information theory and an extension of the maximum likelihood principle; 2nd International Symposium on Information Theory; 1973. p. 267-281.
- Cleveland WS. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*. 1979; 74:829–836.
- Cleveland WS, Devlin SJ. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*. 1988; 83:596–610.
- Crump, RK.; Hotz, VJ.; Imbens, GW.; Mitnik, OA. Nonparametric tests for treatment effect heterogeneity. Vol. 324. National Bureau of Economic Research; 2006. Technical Working Paper
- Fan, J.; Gijbels, I. Local polynomial modelling and its applications. 1st. Chapman & Hall; 1996.
- Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*. 2001; 96:1348–1360.
- Foster JC, Taylor JMG, Ruberg SJ. Subgroup identification from randomized clinical trial data. *Statist. Med*. 2011; 30:2867–2880.
- Hastie T, Loader C. Local regression: automatic kernel carpentry. *Statistical Science*. 1993; 8(2):120–129.
- Hurvich CM, Tsai C. Regression and time series model selection in small samples. *Biometrika*. 1989; 76:297–307.

- Hurvich CM, Simonoff JS, Tsai C. Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *J. R. Statist. Soc.* 1998; 60:271–293.
- Lipkovich I, Dmitrienko A, Denne J, Enas G. Subgroup identification based on differential effect search – A recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in Medicine*. 2011; 30:2601–2621. [PubMed: 21786278]
- Mallows CL. Some comments on C_p . *Technometrics*. 1973; 15:661–676.
- Moineddin R, et al. Identifying subpopulations for subgroup analysis in a longitudinal clinical trial. *Contemporary Clinical Trials*. 2008; 29:817–822. [PubMed: 18718556]
- Mulligan G, et al. Gene expression profiling and correlation with outcome in clinical trials of the proteasome inhibitor bortezomib. *Blood*. 2007; 109:3177–3188. [PubMed: 17185464]
- Ruberg SJ, Chen L, Wang Y. The mean does not mean as much anymore: finding sub-groups for tailored therapeutics. *Clinical Trials*. 2010; 7:574–583. [PubMed: 20667935]
- Ruppert D, Wand MP. Multivariate locally weighted least squares regression. *Ann. Statist.* 1994; 22:1346–1370.
- Schwarz G. Estimating the dimension of a model. *Ann. Statist.* 1978; 6:461–464.
- Storlie CB, Helton JC. Multiple predictor smoothing methods for sensitivity analysis: description of techniques. *Reliability Engineering and System Safety*. 2007; 93:28–54.
- Tibshirani R. Regression shrinkage and selection via the LASSO. *J. R. Statist. Soc.* 1996; 58:267–288.
- Wright G, Tan B, Rosenwald A, Hurt EH, Wiestner A, Staudt LM. A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma. *PNAS*. 2003; 100:9991–9996. [PubMed: 12900505]
- Zhang B, Tsiatis AA, Laber EB, Davidian M. A robust method for estimating optimal treatment regimes. *Biomtrics*. 2012; 68:1010–1018.
- Zuo H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, B*. 2005; 67:301–320.

Figure 1.

Comparison of variable selection between simple linear regression (SLR) and LOESS. The p-values are obtained by SLR (left panel) and LOESS (right panel) from the treatment group of the first data set. The predictive covariate X_{19} is marked in both plots but is only significant according to LOESS.

Figure 2. Subpopulation identification plot of the first data set in a zoomed-in region not including large values of the predictor. The curve represents the predicted LOESS curve, with the dashed horizontal line as the sample mean of Y in the treatment group.

Figure 3.

Predictive model and subpopulation identification of the bortezomib data set. Left: Only the first significant predictor is used in the plot, although three predictors are selected. The patients who are responders are labelled by circles and the patients who are non-responders are labelled by crosses. The curve represents the LOESS curve, but projected on the first predictor, with the dashed horizontal line as the cutoff value 2. Right: The first two significant predictors are used to show the LOESS curve projected on the three-dimensional space.

Table 1

Subpopulation prediction table of the first data set.

		Ideal		Total
		In S	Not in S	
Identified	In S	148	8	156
	Not in S	9	157	166
Total		157	165	322

Sensitivity: 94%

Specificity: 95%

Positive Predictive Value: 95%

Negative Predictive Value: 95%

Accuracy: 95%

Table 2

Comparison of subpopulation prediction table of the second data set. The upper panel is the result from our method, and the lower panel is the result from Eli Lilly's regression tree model.

		Ideal		Total
		In S	Not in S	
Identified	In S	149	49	198
	Not in S	10	114	124
Total		159	163	322

Sensitivity: 94%

Specificity: 70%

Positive Predictive Value: 75%

Negative Predictive Value: 92%

Accuracy: 82%

		Ideal		Total
		In S	Not in S	
Identified	In S	159	55	214
	Not in S	0	108	108
Total		159	163	322

Sensitivity: 100%

Specificity: 66%

Positive Predictive Value: 74%

Negative Predictive Value: 100%

Accuracy: 83%

Table 3

Comparison of prediction table of PD vs. R for the bortezomib data in the pharmacogenomics example. The upper panel is the result from our method and the lower panel is the result from Mulligan *et al.* (2007).

		Actual		Total
		R	PD	
Predicted	R	37	13	50
	PD	1	2	3
Total		38	15	53

Sensitivity: 97%

Specificity: 13%

Positive Predictive Value: 74%

Negative Predictive Value: 67%

Accuracy: 74%

		Actual		Total
		R	PD	
Predicted	R	35	10	45
	PD	3	5	8
Total		38	15	53

Sensitivity: 92%

Specificity: 33%

Positive Predictive Value: 78%

Negative Predictive Value: 63%

Accuracy: 75%