Nonparametric Variable Selection and Classification: The CATCH Algorithm

Shijie Tang[a], Lisha Chen[b,*], Kam-Wah Tsui[c,1], Kjell Doksum[d,1]

[a] *Boehringer Ingelheim Pharmaceuticals, Danbury, CT*
[b] *Department of Statistics, Yale University, New Haven, CT*
[c] *Department of Statistics, University of Wisconsin, Madison, WI*
[d] *Department of Statistics, University of Wisconsin, Madison, WI, and Department of Statistics, Columbia University, New York City, NY*

[*] Corresponding and joint first author.

[1] Partially supported by NSF grant DMS-0604931.

Abstract

In a nonparametric framework, we consider the problem of classifying a categorical response Y whose distribution depends on a vector of predictors X, where the coordinates X_j of X may be continuous, discrete, or categorical. To select the variables to be used for classification, we construct an algorithm which for each variable X_j computes an importance score s_j to measure the strength of association of X_j with Y. The algorithm deletes X_j if s_j falls below a certain threshold. It is shown in Monte Carlo simulations that the algorithm has a high probability of only selecting variables associated with Y. Moreover when this variable selection rule is used for dimension reduction prior to applying classification procedures, it improves the performance of these procedures. Our approach for computing importance scores is based on root Chi-square type statistics computed for randomly selected regions (tubes) of the sample space. The size and shape of the regions are adjusted iteratively and adaptively using the data to enhance the ability of the importance score to detect local relationships between the response and the predictors. These local scores are then averaged over the tubes to form a global importance score s_j for variable X_j. When confounding and spurious associations are issues, the nonparametric importance score for variable X_j is computed conditionally by using tubes to restrict the other variables . We call this variable selection procedure CATCH (Categorical Adaptive Tube Covariate Hunting). We establish asymptotic properties, including consistency.

Keywords: Adaptive variable selection, Importance score, Chi-square statistic

1. Introduction

We consider classification problems with a large number of predictors that can be numerical or categorical. We are interested in the case in which many of the predictors may be irrelevant for classification, thus the variables useful for classification need to be selected. In genomic research, the classes considered are often cases and controls. Thus variable selection is the same as the important problem of deciding which variables are associated with disease. With modern technology, data of large dimension arise in many scientific disciplines including biology, genomics, astronomy, economics and computer science. In particular, the number of variables can be greater than the sample size, which poses a considerable challenge to the statistical analysis. If only some of the variables are useful for classification, over-fitting is a problem for methods that use all variables, and therefore variable selection becomes critical for statistical analysis.

Methods for variable selection in the classification context include methods that incorporate variable selection as part of the classification procedure. This class includes random forest (Breiman [1]), CART (Breiman et al. [2]) and GUIDE (Loh [12]). Random forest assigns an importance score for each of the predictors and one can drop those variables whose importance score fall below a certain threshold. CART, after pruning, will choose a subset of optimal splitting variables to be the most significant variables. GUIDE is a tree-based method particularly powerful in unbiased variable selection and interaction detection. Other research includes methods that incorporate variable selection by applying shrinkage methods with L_1 norm constraints on the parameters (Tibshirani [17]) that generate sparse vectors of parameter estimates. Wang and Shen [18] and Zhang et al. [19] incorporate variable selection with classification based on support vector machine (SVM) methods. Qiao et al. [14] consider variable selection based on linear discriminant analysis. One limitation of these variable selection methods is that they are not based on nonparametric methods and therefore they may not work well when there is a complex relationship between the prediction variables and the variables to be classified.

We propose a nonparametric method for variable selection called Categorical Adaptive Tube Covariate Hunting (CATCH) that performs well as a variable selection algorithm and can be used to improve the performance of available classification procedures. The idea is to construct a nonparametric measure of the relational strength between each predictor and the categorical response, and to retain those predictors whose relationship to the response is above a certain threshold. The nonparametric measure of importance for each predictor is obtained by first measuring the importance of the predictor using local information, and then combing such local importance scores to obtain an overall importance score. The local importance scores are based on root chi-square type statistics for local contingency tables.

In addition to the aforementioned nonparametric feature, the CATCH procedure has another property: it measures the importance of each variable conditioning on all other variables thereby reducing the confounding that may lead to selection of variables spuriously related to the categorical variable Y. This is accomplished by constraining all predictors but the one we are focusing on. For the case where the number of predictors is huge ($d \gg n$), this can be done by restricting principal components for some types of studies. See Remark 3.4.

Our approach to nonparametric variable selection is related to the EARTH algorithm (Doksum et al. [3]) which applies to nonparametric regression problems with a continuous response variable and continuous predictors. It measures the conditional association between a predictor i and the response variable conditional on all the other predictors $\{j\}_{j \neq i}$, by constraining the $\{j\}_{j \neq i}$ variables to regions called tubes. The local importance score is based on a local linear or a local polynomial regression. The contribution of the current paper is to develop variable selection methods for the classification problem with a categorical response variable and predictors that can be continuous, discrete or categorical.

Our CATCH algorithm can be used as a variable selection step before classification. Any classification method, preferably nonparametric, can be used after we single out the important variables. In particular, SVM and random forest are statistical classification methods that can be used with CATCH. We show in simulation studies that when the true model is highly nonlinear and there are many irrelevant predictors, using CATCH to screen out irrelevant or weak predictors greatly improves the performance of SVM and random forest.

The CATCH algorithm works with general classification data, with both numerical and categorical predictors. Moreover, the CATCH algorithm is robust when the predictors interact with each other, especially when numerical and categorical predictors interact, e.g., for hierarchical interactive association between the numerical and categorical predictors. We present a model with a reasonable sample size and predictor dimension for which random forest has a relatively low chance of detecting the significance of the categorical predictor. The CART and GUIDE algorithms, with pruning, can find the correct splitting variables including the categorical one, but yield relatively high classification errors. Moreover CART and GUIDE have trouble choosing the splitting predictors in the correct order. The CATCH algorithm achieves higher accuracy in the task of variable selection. This is due to the importance scores being conditional as illustrated in the simulation example in section 4.1.

The rest of the paper will proceed as follows. In section 2 we introduce importance scores for univariate predictors. In sections 3.1 - 3.4 we extend these scores to multivariate predictors, and in section 3.5, we introduce the CATCH algorithm. In section 4, we use simulation studies to show the effectiveness of the CATCH algorithm. A real example is provided in section 5. In section 6, we provide some theoretical properties to justify the definition of local contingency efficacy, and in section 7, we show asymptotic consistency of the algorithm.

2. Importance Scores for Univariate Classification

Let $(X^{(i)}, Y^{(i)}), i = 1, \cdots, n$, be independent and identically distributed (i.i.d.) as $(X, Y) \sim P$. We first consider the case of a univariate X, then use the methods constructed for the univariate case to construct the multivariate versions.

2.1. Numerical Predictor: Local Contingency Efficacy

Consider the classification problem with a numerical predictor. Let

$$Pr(Y = c|x) = p_c(x), \quad c = 1, \cdots, C, \quad \sum_{c=1}^{C} p_c(x) \equiv 1. \tag{2.1}$$

We introduce a measure of how strongly X is associated with Y in a neighborhood of a "fixed" point $x^{(0)}$. Points $x^{(0)}$ will be selected at random by our algorithm. Local importance scores will be computed for each point, then averaged. For a fixed $h > 0$, define $N_h(x^{(0)}) = \{x : 0 < |x - x^{(0)}| \leq h\}$ to be a neighborhood of $x^{(0)}$, and let $n(x^{(0)}, h) = \sum_{i=1}^{n} I(X^{(i)} \in N_h(x^{(0)}))$ denote the number of data points in the neighborhood. For $c = 1, \cdots, C$, let $n_c^-(x^{(0)}, h)$ be the number of observations $(X^{(i)}, Y^{(i)})$ satisfying $x^{(0)} - h \leq X^{(i)} < x^{(0)}$ and $Y^{(i)} = c$, let $n_c^+(x^{(0)}, h)$ be the number of $(X^{(i)}, Y^{(i)})$ satisfying $x^{(0)} < X^{(i)} \leq x^{(0)} + h$ and $Y^{(i)} = c$. Let $n_c(x^{(0)}, h) = n_c^+(x^{(0)}, h) + n_c^-(x^{(0)}, h)$, $n^-(x^{(0)}, h) = \sum_{c=1}^{C} n_c^-(x^{(0)}, h)$, and $n^+(x^{(0)}, h) = \sum_{c=1}^{C} n_c^+(x^{(0)}, h)$. Table 1 gives the resulting local contingency table:

Table 1: Local contingency table

	$Y = 1$	$Y = 2$	$\cdots\cdots$	$Y = C$	Total
$x^{(0)} - h \leq X < x^{(0)}$	$n_1^-(x^{(0)}, h)$	$n_2^-(x^{(0)}, h)$	$\cdots\cdots$	$n_C^-(x^{(0)}, h)$	$n^-(x^{(0)}, h)$
$x^{(0)} < X \leq x^{(0)} + h$	$n_1^+(x^{(0)}, h)$	$n_2^+(x^{(0)}, h)$	$\cdots\cdots$	$n_C^+(x^{(0)}, h)$	$n^+(x^{(0)}, h)$
Total	$n_1(x^{(0)}, h)$	$n_2(x^{(0)}, h)$	$\cdots\cdots$	$n_C(x^{(0)}, h)$	$n(x^{(0)}, h)$

To measure how strongly Y relates to local restrictions on X, we consider the chi-square statistic:

$$\mathcal{X}^2(x^{(0)}, h) = \sum_{c=1}^{C} \left(\frac{(n_c^-(x^{(0)}, h) - E_c^-(x^{(0)}, h))^2}{E_c^-(x^{(0)}, h)} + \frac{(n_c^+(x^{(0)}, h) - E_c^+(x^{(0)}, h))^2}{E_c^+(x^{(0)}, h)} \right), \tag{2.2}$$

where $0/0 \equiv 0$ and

$$E_c^-(x^{(0)},h) = \frac{n^-(x^{(0)},h)n_c(x^{(0)},h)}{n(x^{(0)},h)}; \quad E_c^+(x^{(0)},h) = \frac{n^+(x^{(0)},h)n_c(x^{(0)},h)}{n(x^{(0)},h)}. \tag{2.3}$$

Due to the local restriction on X, the local contingency table might have some zero cells. However this will not be an issue for the χ^2 statistic. The χ^2 statistic can detect local dependence between Y and X as long as there exist observations from at least two categories of Y in the neighborhood of $x^{(0)}$. When all observations belong to the same category of Y, intuitively no dependence can be detected. In this case χ^2 statistic is equal to zero which coincides with our intuition.

We call the neighborhood $N_h(x^{(0)})$ a section and maximize $\mathcal{X}^2(x^{(0)},h)$ (2.2) with respect to the section size h. Let plim denote limit in probability. Our local measure of association is $\zeta(x^{(0)})$ as given in:

Definition 1. *Local Contingency Efficacy (For Numerical Predictor).* The *local contingency section efficacy* and the *Local Contingency Efficacy (LCE)* of Y on a numerical predictor X at the point $x^{(0)}$ are:

$$\zeta(x^{(0)},h) = \text{plim}_{n\to\infty} n^{-1/2}\sqrt{\mathcal{X}^2(x^{(0)},h)}, \qquad \zeta(x^{(0)}) = \sup_{h>0}\{\zeta(x^{(0)},h)\}. \tag{2.4}$$

□

If $p_c(x)$ is constant in x for all $c \in \{1,\cdots,C\}$ for x near $x^{(0)}$, then, as $n \to \infty$, $\mathcal{X}^2(x^{(0)},h)$ converges in distribution to a χ^2_{C-1} variable and in this case, $\zeta(x^{(0)},h) = 0$. On the other hand, if $p_c(x)$ is not constant near $x^{(0)}$, $\zeta(x^{(0)},h)$ determines the asymptotic power for Pitman's alternatives of the test based on $\mathcal{X}^2(x^{(0)},h)$ for testing whether X is locally independent of Y and is called the *efficacy* of this test. Moreover, $\zeta^2(x^{(0)},h)$ is the asymptotic noncentrality parameter of the statistic $n^{-1}\mathcal{X}^2(x^{(0)},h)$.

The estimators of $\zeta(x^{(0)},h)$ and $\zeta(x^{(0)})$ are:

$$\hat{\zeta}(x^{(0)},h) = n^{-1/2}\sqrt{\mathcal{X}^2(x^{(0)},h)}; \quad \hat{\zeta}(x^{(0)}) = \sup_{h>0}\hat{\zeta}(x^{(0)},h). \tag{2.5}$$

We selected the section size h as

$$\hat{h} = \arg max\{\hat{\zeta}(x^{(0)},h) : h \in \{h_1,\ldots,h_g\}\}$$

for some grid $\{h_1,\ldots,h_g\}$. This corresponds to selecting h by maximizing estimated power. See Doksum and Schafer [5], Gao and Gijbels [7], Doksum et al. [3], Schafer and Doksum [16].

2.2. Categorical Predictor: Contingency Efficacy

Let $(X^{(i)}, Y^{(i)}), i = 1, \cdots, n$, be as before except $X \in \{1, \cdots, C'\}$ is a categorical predictor. Let $n(c, c')$ be the number of observations satisfying $Y = c$ and $X = c'$, and let

$$\mathcal{X}^2 = \sum_{c=1}^{C} \sum_{c'=1}^{C'} \frac{(n(c, c') - E(c, c'))^2}{E(c, c')^2}, \tag{2.6}$$

where

$$E(c, c') = \frac{\sum_{c=1}^{C} n(c, c') \sum_{c'=1}^{C'} n(c, c')}{n}. \tag{2.7}$$

Definition 2. *Contingency Efficacy (For Categorical Predictor).* The *Contingency Efficacy* (CE) of Y on a categorical predictor X is.

$$\zeta = \text{plim}_{n\to\infty} n^{-1/2} \sqrt{\mathcal{X}^2}. \tag{2.8}$$

Under the null hypothesis that Y and X are independent, $\mathcal{X}^2$ converges in distribution to a $\chi^2_{(C-1)(C'-1)}$ variable. In this case, $\zeta = 0$. In general, ζ^2 is the asymptotic noncentrality parameter of the statistic $n^{-1}\mathcal{X}^2$. The contingency efficacy ζ determines the asymptotic power of the test based on $\mathcal{X}^2$ for testing whether Y and X are independent. We use ζ as an importance score that measures how strongly Y depends on X. The estimator of (2.8) is:

$$\hat{\zeta} = n^{-1/2} \sqrt{\mathcal{X}^2}. \tag{2.9}$$

3. Variable Selection for Classification Problems: The CATCH Algorithm

Now we consider a multivariate $X = (X_1, ..., X_d)$, $d > 1$. To examine whether a predictor X_l is related to Y in the presence of all other variables, we calculate efficacies conditional on these variables as well as unconditional or marginal efficacies.

3.1. Numerical Predictors

To examine the effect of X_l in the local region of $x^{(0)} = (x_1^{(0)}, \cdots, x_d^{(0)})$, we build a "tube" around the center point $x^{(0)}$, in which data points are within a certain radius δ of $x^{(0)}$ with respect to all variables except X_l. Define $X_{-l} = (X_1, ..., X_{l-1}, X_{l+1}, ..., X_d)^T$ to be the complementary vector to X_l. Let $D(\cdot, \cdot)$ be a distance in R^{d-1} which we will define later.

Definition 3. A *tube*, T_l, of size δ ($\delta > 0$) for the variable X_l at $x^{(0)}$ is the set

$$T_l \equiv T_l(x^{(0)}, \delta, D) \equiv \{x : D(x_{-l}, x_{-l}^{(0)}) \leq \delta\} \tag{3.1}$$

We adjust the size of the tube so that the number of points in the tube is a preassigned number k. We find that $k \geq 50$ is in general a good choice. Note that $k = n$ corresponds to unconditional or marginal nonparametric regression.

Within the tube $T_l(x^{(0)}, \delta, D)$, X_l is unconstrained. To measure the dependence of Y on X_l locally, we consider neighborhoods of $x^{(0)}$ along the direction of X_l within the tube. Let

$$N_{l,h,\delta,D}(x^{(0)}) = \{x = (x_1, \cdots, x_d)^T \in \mathbf{R}^d : x \in T_l(x^{(0)}, \delta, D), |x_l - x_l^{(0)}| \leq h\} \qquad (3.2)$$

be a section of the tube $T_l(x^{(0)}, \delta, D)$. To measure how strongly X_l relates to Y in $N_{l,h,\delta,D}(x^{(0)})$, we consider

$$\mathcal{X}^2_{l,\delta,D}(x^{(0)}, h). \qquad (3.3)$$

where (3.3) is defined by (2.2) using only the observations in the tube $T_l(x^{(0)}, \delta, D)$. Analogous to the definitions of local contingency efficacy (2.4), we define

Definition 4. The *local contingency tube section efficacy* and the *local contingency tube efficacy* of Y on X_l at the point $x^{(0)}$ and their estimates are:

$$\zeta(x^{(0)}, h, l) = \text{plim}_{n\to\infty}\, n^{-1/2}\sqrt{\mathcal{X}^2_{l,\delta,D}(x^{(0)}, h)}; \zeta_l(x^{(0)}) = \sup_{h>0}\{\zeta(x^{(0)}, h, l)\}. \qquad (3.4)$$

$$\hat{\zeta}(x^{(0)}, h, l) = n^{-1/2}\sqrt{\mathcal{X}^2_{l,\delta,D}(x^{(0)}, h)}; \quad \hat{\zeta}_l(x^{(0)}) = \sup_{h>0}\{\hat{\zeta}(x^{(0)}, h, l)\}. \qquad (3.5)$$

For an illustration of local contingency tube efficacy consider $Y \in \{1,\ 2\}$ and suppose logit $P\,(Y = 2|x) = \left(x - \frac{1}{2}\right)^2$, $x \in [0,\ 1]$. Then the local efficacy with h small will be large, while if $h \geq 1$, the efficacy will be zero.

3.2. Numerical and Categorical Predictors

If X_l is categorical but some of the other predictors are numerical, we construct tubes $T_l(x^{(0)}, \delta, D)$ based on the numerical predictors leaving the categorical variables unconstrained, and we use the statistic as defined by (2.6) to measure the strength of the association between Y and X_l in the tube $T_l(x^{(0)}, \delta, D)$, by using only the observations in the tube $T_l(x^{(0)}, \delta, D)$. We denote this version of (2.6) by

$$\mathcal{X}^2_{l,\delta,D}(x^{(0)}). \qquad (3.6)$$

Analogous to the definition of contingency efficacy given in (2.8), we define

Definition 5. The *local contingency tube efficacy* of Y on X_l at $x^{(0)}$ and its estimate are

$$\zeta_l(x^{(0)}) = \text{plim}_{n\to\infty}\, n^{-1/2}\sqrt{\mathcal{X}^2_{l,\delta,D}(x^{(0)})}; \quad \hat{\zeta}_l(x^{(0)}) = n^{-1/2}\sqrt{\mathcal{X}^2_{l,\delta,D}(x^{(0)})}. \qquad (3.7)$$

3.3. The Empirical Importance Scores

The empirical efficacies (3.5) and (3.7) measure how strongly Y depends on X_l near one point $x^{(0)}$. To measure the overall dependency of Y on X_l, we randomly sample M bootstrap points x_a^*, $a = 1, ..., M$, with replacement from $\{x^{(i)} : 1 \leq i \leq n\}$ and calculate the average of local tube efficacies estimates at these points.

Definition 6. The CATCH empirical importance score for variable l is

$$s_l = M^{-1} \sum_{a=1}^{M} \hat{\zeta}_l(x_a^*) \tag{3.8}$$

where $\hat{\zeta}_l(x^{(i_a)})$is defined in (3.5) and (3.7) for numerical and categorical predictors, respectively.

3.4. Adaptive Tube Distances

Doksum et al. [3] used an example to demonstrate that the distance D needs to be adaptive to keep variables strongly associated with Y from obscuring the effect of weaker variables. Suppose Y depends on three variable X_1, X_2 and X_3 among the larger set of predictors. The strength of the dependency ranges from weak to strong, which will be estimated by the empirical importance scores. As we examine the effect of X_2 on Y, we want to define the tube so that there is relatively less variation in X_3 than in X_1 in the tube because X_3 is more likely to obscure the effect of X_2 than X_1. To this end we introduce a tube distance for the variable X_l of the form

$$D_l(x_{-l}, x_{-l}^{(0)}) = \sum_{j=1}^{d} w_j |x_j - x_j^{(0)}| I(j \neq l) \tag{3.9}$$

where the weights w_j adjust the contribution of strong variables to the distance so that the strong variables are more constrained.

The weights w_j are proportional to $s_j - s_j'$ which are determined iteratively and adaptively as follows: in the first iteration, set $s_j = s_j^{(1)} = 1$ in (3.9) and compute the importance score $s_j^{(2)}$ using (3.8). For the second iteration, we use the weights $s_j = s_j^{(2)}$ and the distance

$$D_l(x_{-l}, x_{-l}^{(0)}) = \frac{\sum_{j=1}^{d} (s_j - s_j')_+ |x_j - x_j^{(0)}| I(j \neq l)}{\sum_{j=1}^{d} (s_j - s_j')_+ I(j \neq l)} \tag{3.10}$$

where $0/0 \equiv 1$ and s_j' is a threshold value for s_j under the assumption of no conditional association of X_l with Y computed by a simple Monte Carlo technique. This adjustment is important because some predictors by definition have larger importance scores than others when they are all independent of Y. For example a categorical predictor with more levels has a larger importance score than that with fewer levels. Therefore the unadjusted scores s_j can not accurately quantify the relative importance of the predictors in the tube distance calculation. Next, we use (3.8) to produce $s_j = s_j^{(3)}$ and use $s_j^{(3)}$ in the next iteration, and so on.

In the definition above, we need to define $|x_j - x_j^{(0)}|$ appropriately for a categorical variable X_j. Because X_j is categorical, we need to assume that $|x_j - x_j^{(0)}|$ is constant when x_j and $x_j^{(0)}$ take any pair of different categories. One approach is to define $|x_j - x_j^{(0)}| = \infty I(x_j \neq x_j^{(0)})$ in which case all points in the tube are given the same X_j value as the tube center. The approach is a very sensible one as it strictly implements the idea of conditioning on X_j when evaluating the effect of another predictor X_l, and X_j does not contribute to any variation in Y. The problem with this definition, though, is that when the number of categories of X_j is relatively large as compared to sample size, there will not be enough points in the tube if we restrict X_j to a single category. In such a case, we do not restrict X_j, which means that X_j can take any category in the tube. We define $|x_j - x_j^{(0)}| = k_0 I(x_j \neq x_j^{(0)})$, where k_0 is a normalizing constant. The value of k_0 is chosen to achieve a balance between numerical variables and categorical variables. Suppose that X_1 is a numerical variable with standard deviation one and X_2 is a categorical variable. For X_2, we set $k_0 \equiv E(|X_1^{(1)} - X_1^{(2)}|)$, where $X_1^{(1)}$ and $X_1^{(2)}$ are two independent realizations of X_1. Note that $k_0 = \sqrt{2/\pi}$ if X_1 follows a standard normal distribution. Also, k_0 can be based on the empirical distributions of the numerical variables. Based on the proceeding discussion of the choice of k_0, we use $k_0 = \infty$ in the simulations and $k_0 = \sqrt{2/\pi}$ in the real data example.

Remark 3.1. *(choice of k_0) $k_0 = \sqrt{2/\pi}$ is used all the time unless the sample size is large enough in the sense to be explained below. Suppose $k_0 = \infty$, then $D_l\left(x_{-l}, x_{-l}^{(0)}\right)$ is finite only for those observations with $x_j = x_j^{(0)}$. When there are multiple categorical variables, D_l is finite only for those points, denoted by set $\Omega^{(0)}$, that satisfy $x_j = x_j^{(0)}$ for all categorical x_j. When there are many categorical variables, or when the total sample size is small, the size of this set can be very small or close to zero, in which case the tube cannot be well defined, therefore $k_0 = \sqrt{2/\pi}$ is used, as in the real data example. On the other hand, when the size of $\Omega^{(0)}$ is larger than the specified tube size, we can use $k_0 = \infty$, as in the simulation examples.*

3.5. The CATCH Algorithm

The variable selection algorithm for a classification problem is based on importance scores and an iteration procedure.

The Algorithm

(0) Standardizing: Standardize all the numerical predictors to have sample mean 0, and sample standard deviation 1 using linear transformations.

(1) Initializing Importance Scores: Let $S = (s_1, \cdots, s_d)$ be importance scores for the predictors. Let $S' = (s'_1, \cdots, s'_d)$ such that s'_l is a threshold importance score which corresponds to the model where X_l is independent of Y given X_{-l}. Initialize S to be $(1, \cdots, 1)$ and S' to be $(0, \cdots, 0)$. Let M be the number of tube centers and m the number of observations within each tube.

(2) Tube Hunting Loop: For $l = 1, \cdots, d$, do $(a), (b), (c)$ and (d) below:

(a) **Selecting tube centers:** For $0 < b \leq 1$, set $M = [nb]$ where [ ] is the greatest integer function, and randomly sample M bootstrap points $X_1^*, \ldots, X_M^*$ with replacement from the n observations. Set $k = 1$, do (b).

(b) **Constructing tubes:** For $0 < a < 1$, set $m = [na]$. Using the tube distance D_l defined in (3.10), select the tube size a so that there are exactly $m \geq 10$ observations inside the tube for X_l at X_k^*.

(c) **Calculating efficacy:** If X_l is a numerical variable, let $\hat{\zeta}_l(X_k^*)$ be the estimated local contingency tube efficacy of Y on X_l at X_k^* as defined in (3.5). If X_l is a categorical variable, let $\hat{\zeta}_l(X_k^*)$ be the estimated contingency tube efficacy of Y on X_l at X_k^* as defined in (3.7).

(d) **Thresholding:** Let Y^* denote a random variable which is identically distributed as Y, but independent of X. Let $(Y_0^{(1)}, \cdots, Y_0^{(n)})$ be a random permutation of $(Y^{(1)}, \cdots, Y^{(n)})$, then $Y_0 \equiv (Y_0^{(1)}, \cdots, Y_0^{(n)})$ can be viewed as a realization of Y^*. Let $\hat{\zeta}_l^*(X_k^*)$ be the estimated local contingency tube efficacy of Y_0 on X_l at X_k^* calculated as in (c) with Y_0 in place of Y.

If $k < M$, set $k = k + 1$, go to (b), otherwise, go to (e).

(e) **Updating importance scores:** Set

$$s_l^{new} = \frac{1}{M}\sum_{k=1}^{M}\hat{\zeta}_l(X_k^*), \quad s_l^{\prime new} = \frac{1}{M}\sum_{k=1}^{M}\hat{\zeta}_l^*(X_k^*).$$

Let $S^{new} = (s_1^{new}, \cdots, s_d^{new})$ and $S^{\prime new} = (s_1^{\prime new}, \cdots, s_d^{\prime new})$ denote the updates of S and S'.

(3) Iterations and Stopping Rule: Repeat (2) I times. Stop when the change in S^{new} is small. Record the last S and S', as S^{stop} and S'^{stop}, respectively.

(4) Deletion step: Using S^{stop} and S'^{stop}, delete X_l if $s_l^{stop} - s_l'^{stop} < \sqrt{2}SD(s'^{stop})$. If d is small, we can generate several sets of S'^{stop} and then use the sample standard deviation of all the values in these sets to be $SD(s'^{stop})$. See Remark 3.3.

End of the algorithm □

Remark 3.2. *Step (d) needs to be replaced by (d') below when a global permutation of Y does not result in appropriate "local null distribution". For example when sample sizes* *of different classes are extremely unequal, which is known as unbalanced classification problems,it is very likely to have a pure class of Y in local region after the global permutation and the threshold S' will be spuriously low. And therefore irrelevant variables will be falsely selected into the model. This approach is more nonparametric, but more time consuming.*

(d') **Local Thresholding:** *Let Y^* denote a random variable which is distributed as Y but independent of X_l given X_{-l}. Let $(Y_0^{(1)}, \cdots, Y_0^{(m)})$ be a random permutation of the Y's inside the tube, then $(Y_0^{(1)}, \cdots, Y_0^{(m)})$ can be viewed as a realization of Y^*. If X_l is a numerical variable, let $\hat{\zeta}_l^{'}(X_k^*)$ be the estimated local contingency tube efficacy of Y^* on X_l at X_k^* as defined in (3.5). If X_l is a categorical variable, let $\hat{\zeta}_l^{'}(X_k^*)$ be the estimated contingency tube efficacy of Y^* on X_l at X_k^* as defined in (3.7).*

Remark 3.3. *(The threshold). Under the assumption $H_0^{(l)}$ that Y is independent of X_l given X_{-l} in T_l, s_l and s_l' are nearly independent and $SD(s_l - s_l') \approx \sqrt{2}SD(s_l')$. Thus the rule that deletes X_l when $s_l^{stop} - s_l'^{stop} < \sqrt{2}SE(s'^{stop})$ is similar to the one standard deviation rule of CART and GUIDE. The choice of threshold is also discussed in Section 7. Alternatively, we calculate the p-values for each covariate by permutation tests. More specifically, we permute the response variable and obtain the important scores for the permeated data and then calculate how extreme $s_l'^{stop}$ compared to permutated scores. We then identify a covariate as important if its p-value is less than 0.1.*

Remark 3.4. *(Large d, small n) A key idea in this paper is that when determining the importance of the variable X_j we restrict (condition) the vector X_{-j} to a relatively small set. This is done to address the problem of the spurious correlation of great concern in genomics (Price et al. [13], Lin and Zeng [11]). In such genomic studies, the number of predictors d typically greatly exceeds sample size n. For numerical variables, this can be dealt with by replace X_{-j} with a vector of principal components explaining most of the variation in X_{-j}. Although our paper does not deal with the $d \gg n$ case directly, this remark shows that it is also applicable to the genome wide association studies described in Price et al. [13].*

Remark 3.5. *We now provide computational complexity of the algorithm. Recall the notations: I is the number of repetitions; M is the number of tube centers; m is the tube size; d is the dimension and n is the sample size. The number of operations needed for the algorithm is $I \times d \times \{M \times (2b + 2c + 2d) + 2e\}$, where 2b, 2c, 2d, 2e denote the number of operations required for those steps. The computational complexity required for 2b, 2c, 2d and 2e are $O(nd)$, $O(m)$, $O(m)$and $O(M)$. Since $nd \gg m$ and the required computation is dominated by step 2b, the computational complexity of the algorithm is $O(IMnd^2)$. Solving least squares problem requires $O(nd^2)$ operations. So our algorithm requires higher computational complexity by a factor of IM compared to least squares. However, the computations for M tube centers are independent of the other computations and are performed in parallel.*

4. Simulation Studies

4.1. Variable Selection with Categorical and Numerical Predictors

Example 4.1. Consider $d \geq 5$ predictors $X_1, X_2, ..., X_d$, and a categorical response variable Y. The fifth predictor X_5 is a Bernoulli random variable which takes two values 0 and 1 with equal probability. All other predictors are uniformly distributed on [0,1].

When $X_5 = 0$, the dependent variable Y is related to the predictors in the following fashion:

$$Y = \begin{cases} 0 & \text{if } X_1 - X_2 - 0.25\sin(16\pi X_2) \leq -0.5; \\ 1 & \text{if } -0.5 < X_1 - X_2 - 0.25\sin(16\pi X_2) \leq 0; \\ 2 & \text{if } 0 < X_1 - X_2 - 0.25\sin(16\pi X_2) \leq 0.5; \\ 3 & \text{if } X_1 - X_2 - 0.25\sin(16\pi X_2) > 0.5. \end{cases} \tag{4.1}$$

When $X_5 = 1$, Y depends on X_3 and X_4 in the same way as above with X_1 replaced by X_3 and X_2 replaced by X_4. The other variables $X_6, \ldots, X_d$ are irrelevant noise variables. All d predictors are independent.

In addition to the presence of noise variables $X_6, \ldots, X_d$, this example is challenging in two aspects. First there is an interaction effect between the continuous predictors $X_1, \ldots, X_4$ and the categorical predictor X_5. Such interaction effect may arise in practice. For example, suppose Y represents the effect resulting from multiple treatments $X_1, \ldots, X_4$ and that X_5 represents the gender of a patient. Our model allows males and females to respond differently to treatments. Secondly the classification boundaries for fixed X_5 are highly nonlinear, as shown in Figure 1. This design is to demonstrate the our method can detect nonlinear dependence between the response and the predictors.

Table 2 shows the number of simulations where the predictors $X_1, \cdots, X_5$ are correctly kept in the model, and, in column 6 the average number of simulations where $d-5$ irrelevant predictors are falsely selected to be in the model out of 100 simulation trials from model (4.1). In the simulation, $n = 400, d = 10$, 50 and 100 are used. We compare CATCH with other methods, including Marginal CATCH, Random Forest classifier (RFC) (Breiman [1]) and GUIDE (Loh [12]). For Marginal CATCH we test the association between Y and X_l without conditioning on X_{-l}. RFC is well-known as a powerful tool for prediction. Here we use importance scores from RFC for variable selection. We create additional 30 noise variables and use their average importance scores multiplied by a factor of 2 as a threshold (Doksum et al. [3]) to decide whether a variable should be selected into the model. In particular we generate the noise variables using uniform and Bernoulli distributions, respectively, for numerical and categorical predictors.

The main goal of GUIDE is to solve the variable selection bias problem in regression and classification tree methods. As an illustration, if a categorical predictor takes K distinct values, then the splitting test exhausts $2^K - 1$ possibilities, and therefore is

Table 2: Simulation results from Example 4.1 with sample size $n = 400$. For $1 \leq i \leq 5$, the ith Column gives the estimated probability of selection of relevant variable X_i and the standard error based on 100 simulations. Column 6 gives the mean of the estimated probability of selection and the standard error for the irrelevant variables $X_6, \cdots, X_d$

		Number of Selections(100 simulations)					
METHOD	d	X_1	X_2	X_3	X_4	X_5	$X_i, i \geq 6$
CATCH	10	0.82 (0.04)	0.75 (0.04)	0.78 (0.04)	0.82 (0.04)	1 (0)	0.05(0.02)
	50	0.76 (0.04)	0.76 (0.04)	0.87 (0.03)	0.84 (0.04)	0.96 (0.02)	0.08 (0.03)
	100	0.77 (0.04)	0.73 (0.04)	0.82 (0.04)	0.8 (0.04)	0.83 (0.04)	0.09 (0.03)
Marginal CATCH	10	0.77 (0.04)	0.67 (0.05)	0.7 (0.05)	0.8 (0.04)	0.06 (0.02)	0.1 (0.03)
	50	0.69 (0.05)	0.7 (0.05)	0.76 (0.04)	0.79 (0.04)	0.07 (0.03)	0.1 (0.03)
	100	0.78 (0.04)	0.76 (0.04)	0.73 (0.04)	0.73 (0.04)	0.12 (0.03)	0.1 (0.03)
Random Forest	10	0.71 (0.05)	0.51 (0.05)	0.47 (0.05)	0.59 (0.05)	0.37 (0.05)	0.06 (0.02)
	50	0.64 (0.05)	0.61 (0.05)	0.65 (0.05)	0.58 (0.05)	0.28 (0.04)	0.08 (0.03)
	100	0.66 (0.05)	0.58 (0.05)	0.59 (0.05)	0.65 (0.05)	0.17 (0.04)	0.11 (0.03)
GUIDE	10	0.96 (0.02)	0.98 (0.01)	0.93 (0.03)	0.99 (0.01)	0.92 (0.03)	0.33 (0.05)
	50	0.82 (0.04)	0.85 (0.04)	0.9 (0.03)	0.89 (0.03)	0.67 (0.05)	0.12 (0.03)
	100	0.83 (0.04)	0.84 (0.04)	0.86 (0.03)	0.83 (0.04)	0.61 (0.05)	0.08 (0.03)

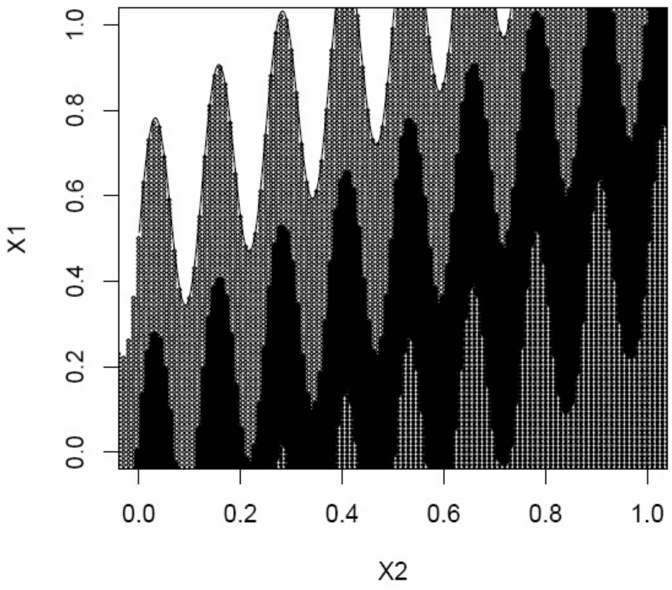

Figure 1: The relationship between Y and X_1, X_2 for model (4.1). The four areas represent the four values of Y

biased toward being selected. The way that GUIDE alleviate this bias is to employ the lack-of-fit tests, i.e. χ^2-test applicable to both numerical and categorical predictors, with the p-values adjusted by the bootstrap procedure. Since GUIDE has negligible selection bias, it can include tests for local pairwise interactions. Furthermore, it achieves sensitivity to local curvature by using linear model at each node.

The results in table 2 show that Marginal CATCH does a very poor job of selecting X_5. This illustrates the importance of local conditioning. RFC does better than Marginal CATCH but also has difficulties, while CATCH and GUIDE can successfully identify X_5 along with X_1 to X_4 with high frequency. When $d = 10$ GUIDE does very well for the relevant variables, but selects too many irrelevant variables as d increases. Surprisingly GUIDE selects fewer irrelevant variables as d increases. It becomes more cautious as the dimension d increases. CATCH performs very well for the high dimensional cases ($d = 50$ and $d = 100$). This indicates that the CATCH algorithm is robust to the intricate interaction structures between the predictors.

In model (4.1) above, the predictors interact in a hierarchical way. When $X_5 = 0$, the response is related to X_1 and X_2 as shown in Figure 1, where the two axes are the two relevant predictors and the four areas partitioned by the curves correspond to different values of Y. When $X_5 = 1$, the response depends on X_3 and X_4 in the same way. This type of hierarchical interaction between the categorical and numerical predictors are difficult to detect even by state of the art classification methods such as support vector machine(SVM) and RFC. In particular, RFC produces an importance score vector that identifies the four numerical variables $X_1, \ldots, X_4$ as very significant, but often leaves the categorical variable X_5 unidentified, as we see in table 2.

The CATCH algorithm, however, performs very well for this complicated model. We now explain why CATCH is able to identify X_5 as an important variable. To explore the association between X_5 and Y, we construct M contingency tables of 2 rows and 4 columns based on M randomly centered tubes, calculate contingency efficacy scores and then average the scores. Figure 2 (a) and (b) illustrate how one of the contingency efficacy scores is calculated. 1000 data points are generated from model (4.1) and they are split into two sets according to x_5. (a) shows approximately half of the points with $x_5 = 0$ in (x_1, x_2) dimensions and (b) shows the rest points with $x_5 = 1$ in (x_3, x_4) dimensions. The data points are colored according to Y. This representation enables us to clearly see the classification boundaries (grey lines). The randomly selected tube center when x_5 is zero is shown as a star in (a). The data points within the tube, close to the tube center in terms of X_{-5}, are highlighted by larger dots. Since the distance defining the tube does not involve X_5, the larger dots have x_5 equal 0 or 1 and are present in both (a) and (b). The counts of the larger dots in different colors in (a) and (b) correspond to the cell counts of the two rows, $x_5 = 0$ and $x_5 = 1$, in the contingency table. As we can see, larger dots in (a) are mostly red and larger dots in (b) are mostly blue and orange, which shows that distribution in Y depends on the value of X_5 and the contingency efficacy score is high. We expect the contingency efficacy score would be low only if the tube center has similar values in (x_1, x_2) and (x_3, x_4). Such tube centers would be very few among the M tube centers since they are randomly chosen. Thus the average contingency efficacy scores is high and X_5 is identified as an important variables.

To contrast this example with a simpler data generating model, suppose the effects of predictors are additive, that is, the categorical variable is related to the response as an added term to the effect of the numerical predictors. In this type of simpler models, the RFC and GUIDE algorithm can efficiently identify relevant predictors but so can the CATCH algorithm. We do not include the simulation results for additive models here.

Remark 4.1. *We observe that the CATCH algorithm is not very sensitive to the choice of the number of tube centers M or the tube size $m = [na]$, where $0 < a \leq 1$ and $[\ ]$ is the greatest integer function. For example 4.1 we perform the simulation using $M = n$ and $M = n/2$, and $a = 0.1, 0.15, 0.2, 0.3, 0.5$. The results are summarized in Table 3. The CATCH algorithm performs similarly for $M = n$ and $M = n/2$ with $M = n$ having a small winning edge. We generally suggest to use $M = n$ if computation resource is of a less concern or n is not large (≤ 200). When computational cost is a concern (Remark 3.5), one can reduce tube size to $M = n/2$ for a compromise between detection power and the computational cost. Table 3 also shows that the tube size works well in the range between 0.1 and 0.3 while the detection power of X_5 decreases for the case $a = 0.5$, where the tube size is too big to achieve local conditioning. This is a consistent* *observation with the comparison to Marginal CATCH in Table 2. On the other hand, the tube size should not be too small to achieve sufficient detection power. We suggest to use* $a = 0.1$ *or 0.15 for sample sizes* $n \geq 400$ *or* $m = 50$ *for smaller sample sizes. In light of these guidelines, we use* $M = n/2 = 200$ *and* $a = 0.15$ *for Examples 4.1-4.3,* $M = n = 200$ *and* $m = 50$ *for Example 4.4.*

Table 3: Sensitivity study of CATCH for Example 4.1 ($n = 400$, $d = 50$) using different number of tube centers M and tube size $m = [na]$. For $1 \leq i \leq 5$, the ith column gives the number of times X_i was selected. Column 6 gives the percentage of times irrelevant variables were selected.

		Number of Selections(100 simulations)					
M	a	X_1	X_2	X_3	X_4	X_5	$X_i, i \geq 6$
$M = 200$	0.10	75	67	77	69	92	7.8
	0.15	76	76	87	84	96	8.2
	0.20	85	84	90	87	93	9.9
	0.30	89	87	95	89	85	9.6
	0.50	89	90	96	93	62	9.9
$M = 400$	0.10	74	76	82	77	92	7.2
	0.15	85	82	89	87	94	8.6
	0.20	87	88	90	86	96	9.1
	0.30	88	89	91	91	92	9.4
	0.50	89	91	94	93	61	10.2

In Example 4.1, all predictors are independent. In a similar setting, we next consider the case where the predictors are dependent.

Example 4.2. We generate standard normal random variables Z_i, for $i = 1, \cdots, d$, in such a way that $cor(Z_1, Z_4) = 0.5$, $cor(Z_3, Z_6) = 0.9$,and the other Z_i's are independent. We set $X_i = \Phi(Z_i)$, for $i = 1, \cdots,\ d$, where $\Phi(\cdot)$ is the CDF of a standard normal, and therefore X_i are uniformly distributed marginally as in Example 4.1. The correlation structure between the X_i's are similar to that of the Z_i's. The dependent variable Y is generated in the same way as in Example 4.1.

The results of Example 4.2 are given in Table 4. By comparing with results in Table 2, both CATCH and GUIDE are doing well in the presence of dependence between the numerical variables. The explanation is two-fold. First, both methods could identify the categorical X_5 as an important predictor most of the time which make the identification of other relevant predictors possible. Second, conditioning on X_5, Y depends on a pair of variables ("X_1 and X_2" or "X_3 and X_4") jointly and in particular this dependence is nonlinear, which makes both methods less affected by the correlation introduced here.

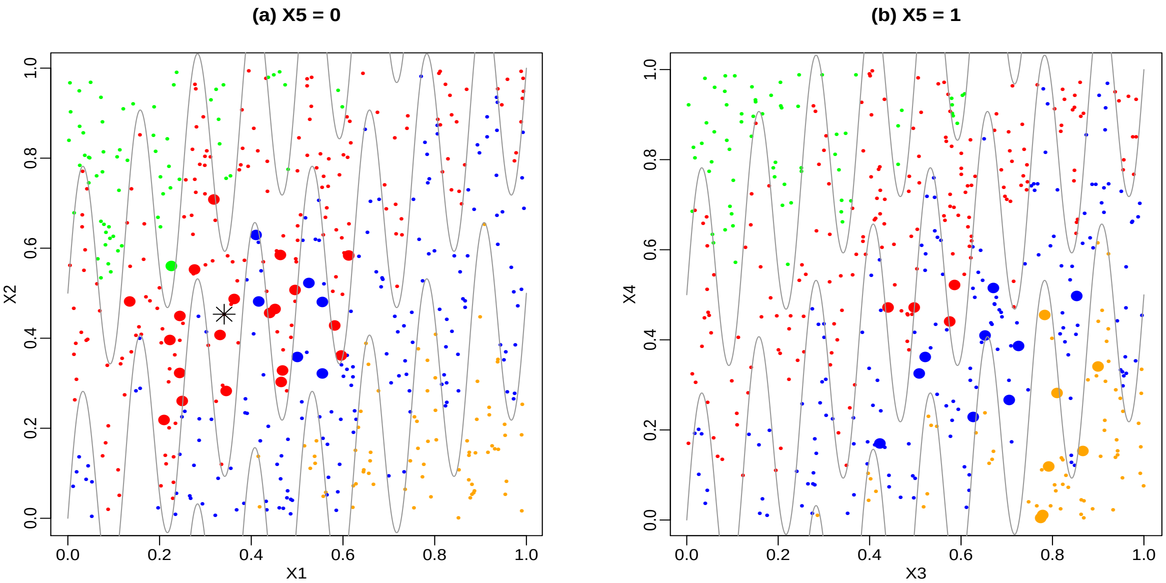

Figure 2: Sample scatter plots of (x_1, x_2) and (x_3, x_4) for simulated data using model (4.1). The left panel includes all the pairs (x_1, x_2) with $x_5 = 0$ while the right panel includes all the pairs (x_3, x_4) with $x_5 = 1$. The larger dots in the two plots are within the tube around a tube center, shown as the star.

We now explain the latter in details for both methods. In CATCH, the algorithm adaptively weighs the effects of predictors by their importance scores in defining the tubes when they are being conditioned on, therefore identifying either of the two will help to identify the other whereas an irrelevant variable, though strongly associated with one of the relevant variables, is down-weighted iteratively and does not strongly affect the selection of the relevant variable. This can explain the slight decreasing selection probability in X_3 in presence of a strongly correlated X_6, and consequently overall less selection of X_3 and X_4 compared to X_1 and X_2 given the correlation between X_1 and X_4. In GUIDE, the difference between the dependent and independent cases is even less because GUIDE is very powerful in detecting local interaction by literally including pairwise interactions when selecting splitting variables.

Marginal CATCH and Random Forest do poorly in this case mainly because X_5 can not be detected as we explained earlier. More specifically, the dependence between Y and X_4 in $X_5 = 1$ case is the same as that between Y and X_2 in $X_5 = 0$. The linear terms of X_1 and X_2 have opposite signs in decision function therefore marginal effects of X_1 and X_4 are obscured when X_5 is not identified as an important variable.

4.2. Improving The Performance of SVM and RF Classifiers

The criterion used in the CATCH algorithm is not based on classification accuracy. But if we use CATCH to screen out the irrelevant variables, and only use the important

Table 4: Simulation results from Example 4.2 with sample size $n = 400$. For $1 \leq i \leq 6$, the ith Column gives the estimated probability of selection of variable X_i and the standard error based on 100 simulations, where X_i, $i = 1, \cdots, 5$ are relevant variables with $cor(X_1, X_4) = .5$. X_6 is an irrelevant variable that is correlated with X_3, with correlation .9. Column 7 gives the mean of the estimated probability of selection and that of the standard error for the irrelevant variables $X_7, \cdots, X_d$

		Number of Selections(100 simulations)						
METHOD	d	X_1	X_2	X_3	X_4	X_5	X_6	$X_i, i \geq 7$
CATCH	10	0.76 (0.04)	0.77 (0.04)	0.73 (0.04)	0.68 (0.05)	0.99 (0.01)	0.34 (0.05)	0.07 (0.02)
	50	0.78 (0.04)	0.77 (0.04)	0.74 (0.04)	0.63 (0.05)	0.92 (0.03)	0.57 (0.05)	0.09 (0.03)
	100	0.76 (0.04)	0.77 (0.04)	0.8 (0.04)	0.74 (0.04)	0.82 (0.04)	0.55 (0.05)	0.09 (0.03)
Marginal CATCH	10	0.35 (0.05)	0.63 (0.05)	0.8 (0.04)	0.32 (0.05)	0.09 (0.03)	0.71 (0.05)	0.12 (0.03)
	50	0.34 (0.05)	0.71 (0.05)	0.7 (0.05)	0.3 (0.05)	0.1 (0.03)	0.6 (0.05)	0.11 (0.03)
	100	0.34 (0.05)	0.68 (0.05)	0.75 (0.04)	0.3 (0.05)	0.1 (0.03)	0.59 (0.05)	0.1 (0.03)
Random Forest	10	0.29 (0.05)	0.48 (0.05)	0.55 (0.05)	0.2 (0.04)	0.3 (0.05)	0.47 (0.05)	0.05 (0.02)
	50	0.26 (0.04)	0.52 (0.05)	0.57 (0.05)	0.3 (0.05)	0.24 (0.04)	0.45 (0.05)	0.08 (0.03)
	100	0.36 (0.05)	0.61 (0.05)	0.66 (0.05)	0.37 (0.05)	0.32 (0.05)	0.46 (0.05)	0.12 (0.03)
GUIDE	10	0.96 (0.02)	0.96 (0.02)	0.94 (0.02)	0.95 (0.02)	0.96 (0.02)	0.84 (0.04)	0.3 (0.05)
	50	0.8 (0.04)	0.92 (0.03)	0.88 (0.03)	0.79 (0.04)	0.72 (0.04)	0.68 (0.05)	0.12 (0.03)
	100	0.81 (0.04)	0.89 (0.03)	0.9 (0.03)	0.76 (0.04)	0.75 (0.04)	0.58 (0.05)	0.08 (0.03)

variables for classification, the performance of other algorithms such as Support Vector Machine (SVM) and Random Forest Classifiers (RFC) can be significantly improved. We compare the prediction performance of (1) SVM, (2) RFC, (3) CATCH followed by SVM, and (4) CATCH followed by RFC in this section. We label (3) and (4) with CATCH+SVM and CATCH+RFC, respectively. In this section we use the "svm" function in the "e1071" package in R with radial kernel to perform SVM analysis, and use the "randomForest" function in the "randomForest" package to perform RFC analysis. We use simulation experiments to illustrate the improvement obtained by using CATCH prior to classifying Y.

Let the true model be

$$(X, Y) \sim P, \tag{4.2}$$

where P will be specified later. In each of 100 Monte Carlo experiments, n pairs of $(X^{(i)}, Y^{(i)}), i = 1, \cdots, n$, are generated from the true model (4.2). Based on the simulated data $(X^{(i)}, Y^{(i)})$, $i = 1, \cdots, n$, the methods (1) SVM, (2) RFC, (3) CATCH+SVM, (4) CATCH+RFC are used to classify a future $Y^{(0)}$ for which we have available a corresponding $X^{(0)}$.

The integrated-misclassification rate (IMR, Friedman [6]) defined below is used to evaluate the performance of the classification algorithms. Let $\mathcal{F}_{\mathbb{X},\mathbb{Y}}(\cdot)$ be the classifier constructed by a classification algorithm based on $\mathbb{X} = (X^{(1)}, \cdots, X^{(n)})$, and $\mathbb{Y} = (Y^{(1)}, \cdots, Y^{(n)})^T$.

Let $(X^{(0)}, Y^{(0)})$ be a future observation with the same distribution as $(X^{(i)}, Y^{(i)})$. We only know $X^{(0)} = x^{(0)}$, and want to classify $Y^{(0)}$. For a possible future observation $(x^{(0)}, y^{(0)})$, define

$$\mathrm{MR}(\mathbb{X}, \mathbb{Y}; x^{(0)}, y^{(0)}) = P(\mathcal{F}_{\mathbb{X},\mathbb{Y}}(x^{(0)}) \neq y^{(0)} | \mathbb{X}, \mathbb{Y}). \tag{4.3}$$

Our criterion is

$$\mathrm{IMR} = E\{E_0[\mathrm{MR}(\mathbb{X}, \mathbb{Y}; x^{(0)}, y^{(0)})]\}. \tag{4.4}$$

where E_0 is with respect to the distribution of $(x^{(0)}, y^{(0)})$ random and E is with respect to the distribution of $(\mathbb{X}, \mathbb{Y})$.

This double expectation is evaluated using Monte Carlo simulation. The result is denoted as IMR*. More precisely, E_0 is estimated using 5000 simulated $(x^{(0)}, y^{(0)})$ observations, then IMR* is computed on the basis of 100 simulated $\{(\mathbb{X}_i, \mathbb{Y}_i), i = 1, \cdots, 100\}$ samples.

Example 4.3. First consider model (4.1) in section 4.1. Figure 3 shows IMR* versus the number of irrelevant variables for this example. The left panel shows that the IMR*'s of the SVM approach increases dramatically as the number of irrelevant variables increases, while the IMR*'s of CATCH+SVM are stable. Thus CATCH stabilizes the performance of SVM for this complex model. The right panel shows that CATCH also stabilizes the performance of RFC. These results indicate that CATCH can improve the prediction accuracy of other classification algorithms.

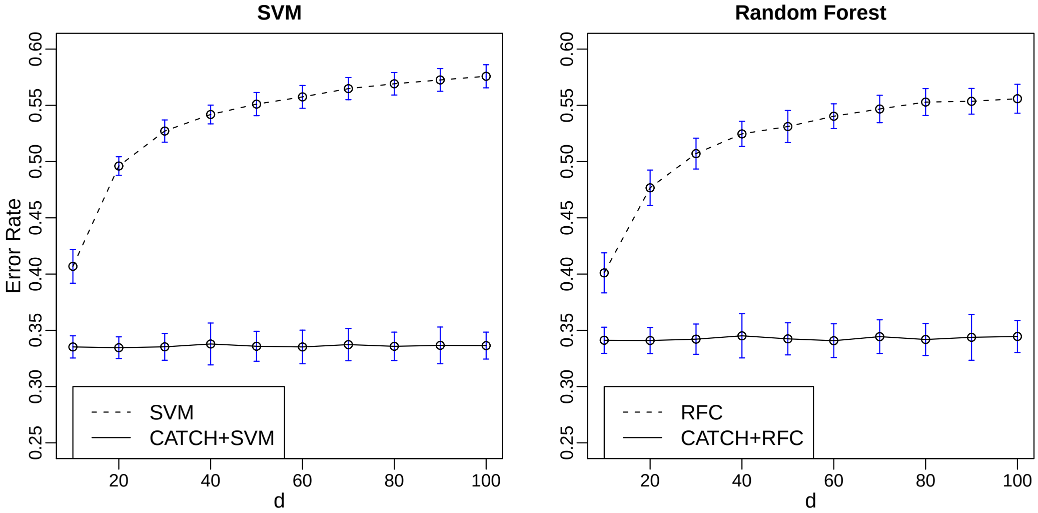

Figure 3: IMR*'s of SVM, RFC, CATCH+SVM, CATCH+RFC versus the number of predictors for the simulation study in model (4.1) in example 4.1. The sample size is $n = 400$.

Example 4.4. Contaminated Logistic Model: Consider a binary response $Y \in \{0, 1\}$ with

$$P(Y = 0|X = x) = \begin{cases} F(x) & \text{if } r = 0 \\ G(x) & \text{if } r = 1 \end{cases} \tag{4.5}$$

where $x = (x_1, \ldots, x_d)$,

$$r \sim Bernoulli(\gamma) \tag{4.6}$$

$$F(x) = (1 + \exp\{-(0.25 + 0.5x_1 + x_2)\})^{-1} \tag{4.7}$$

$$G(x) = \begin{cases} 0.1 & \text{if } 0.25 + 0.5x_1 + x_2 < 0 \\ 0.9 & \text{if } 0.25 + 0.5x_1 + x_2 \geq 0 \end{cases} \tag{4.8}$$

and $x_1, x_2, \ldots, x_d$ are realizations of independent standard normal random variables. Hence, given $X^{(i)} = x^{(i)}, 1 \leq i \leq n$, the joint distribution of $Y^{(1)}, Y^{(2)}, \ldots, Y^{(n)}$ is the contaminated logistic model with contamination parameter γ

$$(1-\gamma)\prod_{i=1}^{n}[F(x^{(i)})]^{Y^{(i)}}[1-F(x^{(i)})]^{1-Y^{(i)}}+\gamma\prod_{i=1}^{n}[G(x^{(i)})]^{Y^{(i)}}[1-G(x^{(i)})]^{1-Y^{(i)}} \qquad (4.9)$$

We perform simulation study for $n = 200$, $d = 50$ and $\gamma = 0$; 0.2; 0.4; 0.6; 0.8 and 1, where $\gamma = 0$ corresponds to no contamination. Figure 4 shows the IMR*'s versus the different values of γ for this example. CATCH improves the performance of the SVM and Random Forest classifiers for this contaminated logistic regression model.

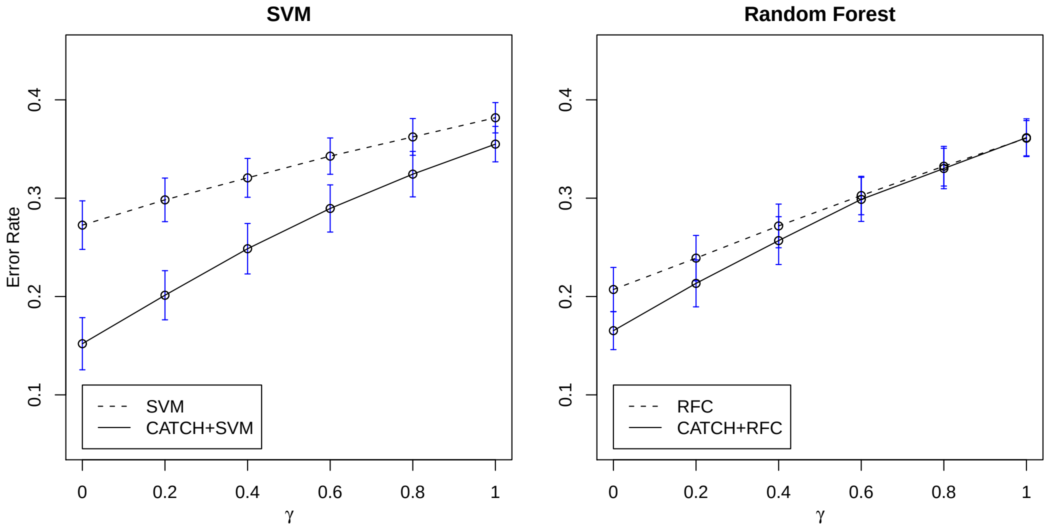

Figure 4: IMR*'s of SVM, RFC, CATCH+SVM, CATCH+RFC versus different contamination parameters γ for simulation model (4.9) in example 4.4.

5. CATCH Analysis of The ANN Data

In this section, CATCH is applied to a data set available at the UCI data base ("`http://ftp.ics.uci.edu/pub/machine-learning-databases/thyroid-disease/`"). The data is from a clinical trial to determine whether a patient is hypothyroid (see e.g., Quinlan [15]). The dataset is split into two parts: a training set with 3772 observations and a testing set with 3428 observations. The response is a categorical variable with three classes: normal (not hypothyroid), hyperfunction, and subnormal functioning. There are 21 predictors, including 15 categorical predictors and 6 numerical predictors. The three classes are very imbalanced with 92.5% of normal patients. A good classifier should produce a significant decrease from an error rate 7.5%, which can be obtained by classifying all observations to the normal class.

CATCH identifies 7 variables as important, including 4 continuous variables and 3 categorical variables. We consider three classification methods: 1. Support Vector Machine with radial kernel, denoted by SVM(r); 2. Support Vector Machine with polynomial kernel, denoted by SVM(p); 3. RFC. We apply these three methods on the training set to build classification models, and use the test set to estimate the misclassification rate of the methods. We apply the methods with CATCH and without CATCH, respectively, where "with CATCH" means that only the 7 important variables identified by CATCH are used to build the model, and "without CATCH" means that all the 21 variables are used in the analysis.

Table 5: Misclassification rates of three classification methods (SVM(r), SVM(p), RFC) with CATCH and without CATCH

	SVM(r)	SVM(p)	RFC
w/o CATCH	0.052	0.063	0.024
with CATCH	0.037	0.055	0.015

Table 5 shows that if CATCH is used to select important variables before the classification methods are applied, the misclassification rates decrease. CATCH reduces the misclassification rates by 29%, 13% and 37% for SVM(r), SVM(p) and RFC respectively. CATCH + RFC has the best performance. Although CATCH is not designed for improving classification accuracy, nonetheless, it can help other classification methods to perform better.

6. Properties of Local Contingency Efficacy

In this section, we investigate the properties of local contingency efficacy. For a univariate continuous predictor, Theorem 6.1 below shows that local contingency efficacy is well-defined and equals to 0 if the response variable is independent of the predictor. Consider an $R \times C$ contingency table. Let n_{rc}, $r = 1, \cdots, R$, $c = 1, \cdots, C$, be the observed frequency in the (r, c)-cell, and let p_{rc} be the probability that an observation belongs to the (r, c)-cell. Let

$$p_r = \sum_{c=1}^{C} p_{rc}; \quad q_c = \sum_{r=1}^{R} p_{rc}; \tag{6.1}$$

and

$$n_{r\cdot} = \sum_{c=1}^{C} n_{rc}; \quad n_{\cdot c} = \sum_{i=1}^{R} n_{rc}; \quad n = \sum_{i=1}^{R}\sum_{c=1}^{C} n_{rc}; \quad E_{rc} = n_{r\cdot} n_{\cdot c}/n. \tag{6.2}$$

Consider testing that the rows and the columns are independent, i.e.,

$$H_0 : \forall r, c \quad p_{rc} = p_r q_c \qquad versus \qquad H_1 : \exists r, c \quad p_{rc} \neq p_r q_c, \tag{6.3}$$

a standard test statistic is:

$$\mathcal{X}^2 = \sum_{r=1}^{R}\sum_{c=1}^{C} \frac{(n_{rc} - E_{rc})^2}{E_{rc}}. \tag{6.4}$$

Under H_0, $\mathcal{X}^2$ converges in distribution to $\chi^2_{(R-1)(C-1)}$. Moreover, define $0/0 = 0$, we have the well known result:

Lemma 6.1. *As n tends to infinity, $\mathcal{X}^2/n$ tends in probability to*

$$\tau^2 \equiv \sum_{r=1}^{R}\sum_{c=1}^{C} \frac{(p_{rc} - p_r q_c)^2}{p_r q_c}$$

Moreover, if $p_r q_c$ is bounded away from 0 and 1 as $n \to \infty$, and if R and C are fixed as $n \to \infty$, then $\tau^2 = 0$ if and only if H_0 is true.

Remark 6.1. *Lemma 6.1 shows that the contingency efficacy ζ in (2.8) is well defined. Moreover, $\zeta = 0$ if and only if X and Y are independent.*

Theorem 6.1. *Consider a categorical response Y and a continuous predictor X with density function $f(x)$ such that for a fixed $h > 0$, $f(x) > 0$ for $x \in (x^{(0)} - h, x^{(0)} + h)$, then*

(1) For fixed h, the estimated local section efficacy $\hat{\zeta}(x^{(0)}, h)$ defined in (2.5) converges almost surly to the section efficacy $\zeta(x^{(0)}, h)$ defined in (2.4).

(2) If X and Y are independent, then the local efficacy $\zeta(x^{(0)})$ defined in (2.4) is zero.

(3) If there exists $c \in \{1, \cdots, C\}$, such that $p_c(x) = P(Y = c|x)$ has a continuous derivative at $x = x^{(0)}$ with $p'_c(x^{(0)}) \neq 0$, then $\zeta(x^{(0)}) > 0$.

The χ^2 statistic (6.4) can be written in terms of multinomial proportions $\hat{p}_{rc}$ with $\{\sqrt{n}(\hat{p}_{rc} - p_{rc}), 1 \leq r \leq R, 1 \leq c \leq C\}$ converging in distribution to a multivariate normal. By Taylor expansion of $T_n \equiv \sqrt{\chi^2/n}$ at $\tau = \text{plim}\sqrt{\chi^2/n}$, we find that

Theorem 6.2.

$$\sqrt{n}(T_n - \tau) \longrightarrow_d N(0, v^2) \tag{6.5}$$

where the asymptotic variance v^2, which can be obtained from the Taylor expansion, is not needed in this paper.

Remark 6.2. *The result (6.5) applies to the multivariate χ^2-based efficacies $\hat{\zeta}$ with n replaced by the number of observations that goes in to the computations of $\hat{\zeta}$. In this case we use the notation χ_j^2 for the chi-square statistic for the jth variable.*

7. Consistency Of CATCH

Let Y denote a variable to be classified into one of the categories $\{1, \ldots, C\}$ by using the X_j's in the vector $X = (X_1, \ldots, X_d)$. Let $X_{-j} \equiv X - X_j$ denote X without the X_j component. Assume that the squared Pearson nonparametric correlation (Doksum and Samarov [4], Huang and Chen [9]) between X_j and X_{-j} is less than 0.5. We call the variable X_j relevant for Y if conditionally given X_{-j}, $\mathcal{L}(Y|X_j = x)$ is not constant in x; and irrelevant otherwise. Our data are $(X^{(1)}, Y^{(1)}), \ldots, (X^{(n)}, Y^{(n)})$, i.i.d. as (X, Y). A variable selection procedure is *consistent* if the probability that all the relevant variables are selected and none of the irrelevant variables are selected tends to one as n tends to infinity.

We give a general nonparametric framework where the variable selection rule based on the CATCH importance scores s_j given in (3.8) is consistent.

When the predictors are not all categorical and CATCH involves tubes, we assume that the number of observations in each tube tends to ∞ as $n \to \infty$, and we assume that for those importance scores that require the selection of a section size h, the selection is from a finite set $\{h_j\}$ with $\min_j\{h_j\} > h_0$ for some $h_0 > 0$. It follows that the number of observations $m_{j,k}$ in the kth tube used to calculate the importance score s_j for the variable X_j tends to infinity as $n \to \infty$.

The key quantities that determine consistency are the limits λ_j of s_j. That is

$$\lambda_j = \tau(\zeta_j, P) = E_P[\zeta_j(X)] = \int \zeta_j(x)dP(x),\ j = 1, \ldots, d,$$

where $\zeta_j(x)$ is the local contingency efficacy defined by (2.4) for numerical X's and by (2.8) for categorical X's, and P is the probability distribution of X. Our estimate s_j of τ_j is

$$s_j = \tau(\hat{\zeta}_j, \hat{P}) = \frac{1}{M}\sum_{k=1}^{M} \hat{\zeta}_j(X_k^*),$$

where $\hat{\zeta}_j$ is defined in section 3 and $\hat{P}$ is the empirical distritution of X.

Let d_0 denote the number of X's that are irrelevant for Y. Set $d_1 = d - d_0$, and without loss of generality, reorder the λ_j so that

$$\begin{array}{ll} \lambda_j > 0 & \text{if } j = 1, \ldots, d_1; \\ \lambda_j = 0 & \text{if } j = d_1 + 1, \ldots, d. \end{array} . \tag{7.1}$$

Our variable selection rule is :

$$\text{“keep variable } X_j \text{ if } s_j \geq t \text{ for some threshold } t\text{”} \tag{7.2}$$

Rule (7.2) is consistent if and only if

$$P(\min_{j \leq d_1}\{s_j\} \geq t \quad \text{and} \quad \max_{j > d_1}\{s_j\} < t) \longrightarrow 1 \tag{7.3}$$

To establish (7.3), it is enough to show that

$$P(\max_{j > d_1}\{s_j\} > t) \longrightarrow 0 \tag{7.4}$$

and

$$P(\min_{j \leq d_1}\{s_j\} \leq t) \longrightarrow 0 \tag{7.5}$$

We call (7.4) and (7.5) type I and type II consistency, respectively. Type I and II error probabilities refer to probabilities of any false positives and any false negatives, respectively.

7.1. Type I consistency

We consider (7.4) first and note that

$$P(\max_{j > d_1}\{s_j\} > t) \leq \sum_{j=d_1+1}^{d} P(s_j > t) \ . \tag{7.6}$$

For the jth irrelevant variable X_j, $j > d_1$, the results in Section 6 apply to the local efficacy $\hat{\zeta}_j(x)$ at a fixed point x. The importance score s_j defined in (3.8) is the average of such efficacies over a sample of M random points $\{x_a^*\}$. We assume that there is a point $x^{(0)}$ such that for irrelevant x_j's and for n large enough, $\hat{\zeta}_j(x^{(0)})$ is stochastically larger in the right tail than the average $M^{-1} \sum_{a=1}^{M} \hat{\zeta}_j(x_a^*)$. That is, we assume that there exist $t > 0$ and n_0 such that for $n \geq n_0$,

$$p(s_j > t) \leq P(s_j^{(0)} > t), \ j > d_1 \tag{7.7}$$

The existence of such a point $x^{(0)}$ is established by considering the probability limit of

$$\arg\max\{\hat{\zeta}_j(x_a^*) : x_a^* \in \{x_1^*, \cdots, x_M^*\}\}$$

Note by Lemma (6.1), $s_j^{(0)} \rightarrow_p 0$ as $n \rightarrow \infty$. It follows that we have established

(7.4) for t and $d_0 = d - d_1$ that are fixed as $n \to \infty$. To examine the interesting case where $d_0 \to \infty$ as $n \to \infty$, we need large deviation results, which we turn to next. In the $d_0 \to \infty$ case, we consider $t \equiv t_n$ that tend to ∞ as $n \to \infty$ and we use the weaker assumption: there exists $x^{(0)}$ and n_0 such that

$$P(s_j > t_n) \le P(s_j^{(0)} > t_n) \tag{7.8}$$

for $n \ge n_0$ and each $\{t_n\}$ with $\lim_{n\to\infty} t_n = \infty$. We also assume that the number of columns C and rows R in the contingency table that define the efficacy χ_j^2/n given by (6.4) stays fixed as $n \to \infty$. In this case, $P(s_j^{(0)} > t_n) \to 0$ provided each term

$$[T_{rc}^{(j)}]^2 \equiv \left[\frac{n_{rc}E_{rc}}{\sqrt{n}\sqrt{E_{rc}}}\right]^2$$

in $s_j^2 = \chi_j^2/n$, defined for X_j by (6.4) satisfies

$$P\left(|T_{rc}^{(j)}| > t_n\right) \longrightarrow 0$$

This is because

$$\begin{aligned} P\left(\sqrt{\sum_r \sum_c [T_{rc}^{(j)}]^2} > t_n\right) &\le P\left(RC \max_{rc}[T_{rc}^{(j)}]^2 > t_n^2\right) \\ &\le RC \max P\left(|T_{rc}^{(j)}| > t_n/\sqrt{RC}\right) \end{aligned}$$

and t_n and $t_n/\sqrt{RC}$ are of the same order as $n \to \infty$. By large deviation theory, Hall [8] and Jing et al. [10], for some constant A,

$$P(|T_{rc}^{(j)}| > t_n) = \frac{2t_n^{-1}}{\sqrt{2\pi}} e^{-t_n^2/2}\left[1 + At_n^3 n^{-\frac{1}{2}} + o\left(t_n^3 n^{-\frac{1}{2}}\right)\right] \tag{7.9}$$

provided $t_n = o\left(n^{\frac{1}{6}}\right)$. If (7.9) is uniform in $j > d_1$, then (7.6) and (7.9) implies that type I consistency holds if

$$d_0 \left(\frac{t_n}{\sqrt{2}}\right)^{-1} e^{-\frac{t_n^2}{2}} \to 0 \tag{7.10}$$

To examine (7.10), write d_0 and t_n as $d_0 = \exp\left(n^b\right)$ and $t_n = \sqrt{2}n^r$. By large deviation theory, (7.10) holds when $r < \frac{1}{6}$. Thus if $0 < \frac{1}{2}b < r < \frac{1}{6}$, then

$$d_0 \left(\frac{t_n}{\sqrt{2}}\right)^{-1} e^{-\frac{t_n^2}{2}} = n^{-r} \exp\left(n^b - n^{2r}\right) \to 0$$

Thus $d_0 = \exp(n^b)$ with $b = \frac{1}{3}\epsilon$ for any $\epsilon > 0$, is the largest d_0 can be to have consistency. For instance, if there are exactly $d_0 = exp\left(n^{\frac{1}{4}}\right)$ irrelevant variables, and we use the threshold $t_n = \sqrt{2}n^{\frac{1}{7}}$, the probability of selecting any of the irrelevant variables tends to zero at the rate

$$n^{-\frac{1}{4}} \exp\left(n^{\frac{1}{4}} - n^{\frac{2}{7}}\right)$$

7.2. Type II consistency

We now assume that for each relevant variable X_j, there is a least favorable point $x^{(1)}$ such that the local efficacy $s_j^{(1)} \equiv \hat{\zeta}(x^{(1)})$ is stochastically smaller than the importance score $s_j = M^{-1} \sum_{a=1}^{M} \hat{\zeta}(x_a^*)$. That is, we assume that for each relevant variable X_j there exists n_1 and $x^{(1)}$ such that for $n \geq n_1$

$$P\left(s_j^{(1)} \leq t_n\right) \geq P\left(s_j \leq t_n\right), \quad j \leq d_1 \tag{7.11}$$

The existence of such a least favorable $x^{(1)}$ is established by considering the probability limit of

$$\arg\min\{\hat{\zeta}_j(x_a^*) : x_a^* \in \{x_1^*, \cdots, x_M^*\}\}$$

We again assume that R and C are fixed, so that to show type II consistency, it is enough to show that for each (r, c) and $j \leq d_1$,

$$P\left(|T_{rc}^{(j)}| \leq t_n\right) \to 0.$$

This is because

$$P\left(\sqrt{\sum_r \sum_c \left[T_{rc}^{(j)}\right]^2} \leq t_n\right)$$
$$\leq P(\min_{r,c} \left[T_{rc}^{(j)}\right]^2 \leq t_n^2)$$
$$\leq RC \max_{r,c} P\left(|T_{rc}^{(j)}| \leq t_n\right)$$

By Theorem 6.2, $\sqrt{n}\left(s_j^{(1)} - \tau_j\right) \to_d N(0, \nu^2)$, $\tau_j > 0$. It follows that if t_n and d_1 are bounded above as $n \to \infty$ and $\tau_j > 0$ then $P\left(s_j^{(1)} \leq t_n\right) \to 0$. Thus type II consistency holds in this case. To examine the more interesting case where t_n and d_1 are not bounded above and τ_j may tend to zero as $n \to \infty$, we will need to center the $T_{rc}^{(j)}$. Thus set

$$\theta_n^{(j)} = \frac{\sqrt{n}\left(p_{rc}^{(j)} - p_r^{(j)} p_c^{(j)}\right)}{\sqrt{p_r^{(j)} p_c^{(j)}}}\ , \tag{7.12}$$

$$T_n(j) = T_{rc}^{(j)} - \theta_n^{(j)} \ ,$$

where for simplicity the dependence of $T_n^{(j)}$ and $\theta_n^{(j)}$ on r and c is not indicated.

Lemma 7.1.

$$P\left(\min_{j \le d_1} |T_{rc}^{(j)}| \le t_n\right) \le \sum_{j=1}^{d_1} P\left(|T_n^{(j)}| \ge \min_{j \le d_1}\{\theta_n^{(j)}\} - t_n\right)$$

Proof.

$$\begin{aligned} P\left(\min_{j \le d_1} |T_{rc}^{(j)}| \le t_n\right) &= P\left(\min_{j \le d_1} |T_n^{(j)} + \theta_n^{(j)}| \le t_n\right) \\ &\le P\left(\min_{j \le d_1} |T_n^{(j)}| - \max_{j \le d_1} |\theta_n^{(j)}| \le t_n\right) \\ &= P\left(\max_{j \le d_1} |T_n^{(j)}| \ge \min_{j \le d_1} |\theta_n^{(j)}| - t_n\right) \\ &\le \textstyle\sum_{j=1}^{d_1} P\left(|T_n^{(j)}| \ge \min_{j \le d_1} |\theta_n^{(j)}| - t_n\right) \end{aligned}$$

□

Set

$$a_n = \min_{j \le d_1} \left|\theta_n^{(j)}\right| - t_n,$$

then by large deviation theory, if $a_n = o\left(n^{\frac{1}{6}}\right)$,

$$P(\left|T_n^{(j)}\right| \ge a_n) \propto \left(\frac{a_n}{\sqrt{2}}\right)^{-1} \exp\{-a_n^2/2\} \tag{7.13}$$

where $\propto$ signifies asymptotic order as $n \to \infty, \ a_n \to \infty$. It follows that if (7.13) holds uniformly in j then by Lemma 7.1,

$$P\left(\min_{j \le d_1} \left|T_{rc}^{(j)}\right| \le t_n\right) \propto d_1 \left(\frac{a_n}{\sqrt{2}}\right)^{-1} \exp\{-a_n^2/2\}$$

Thus type II consistency holds when $a_n = o\left(n^{\frac{1}{6}}\right)$ and

$$d_1 \left(\frac{a_n}{\sqrt{2}}\right)^{-1} \exp\{-a_n^2/2\} \to 0 \tag{7.14}$$

In Section 7.1, t_n is of order n^r, $0 < r < \frac{1}{6}$. For this t_n, type II consistency holds if a_n is of order n^r, $0 < r < \frac{1}{6}$, and

$$d_1 n^{-r} \exp\{-n^{2r}/2\} \to 0,$$

If $a_n = O(n^k)$ with $k \geq \frac{1}{6}$, then $P(\left|T_{(n)}^{(j)}\right| \geq a_n)$ tends to zero at a rate faster than in (7.13) and consistency is ensured. For instance, if all relevant variables X_j have $p_{rc}^{(j)}$ fixed as $n \to \infty$, then $\min_{j \leq d_1} \left|\theta_n^{(j)}\right|$ is of order $\sqrt{n}$ and a_n is of order $n^{\frac{1}{2}} - t_n$.

7.3. Type I and II consistency

We see from Sections 7.1 and 7.2 that consistency holds under a variety of conditions. One set of conditions is

Theorem 7.1. *Under the regularization conditions of Sections 7.1 and 7.2, if*

$$d_0 = c_0 \exp(n^b),\ t_n = c_1 n^r,$$
$$\min_{j \leq d_1} \|\theta^{(j)}\| - t_n = c_2 n^r,\ d_1 = c_3 \exp\left(n^b\right)$$

for positive constants $c_0,\ c_1,\ c_2$ *and* c_3 *and* $0 < \frac{1}{2}b < r < \frac{1}{6}$, *then CATCH is consistent.*

Appendix

Proof of Theorem 6.1

(1) Let $N_h^{\#} = \sum_{i=1}^n I(X^{(i)} \in N_h(x^{(0)}))$. Then $N_h^{\#} \sim Bi(n, \int_{x^{(0)}-h}^{x^{(0)}+h} f(x)dx)$ and $N_h^{\#} \to_{a.s.} \infty$ as $n \to \infty$.

$$\begin{aligned}\hat{\zeta}(x^{(0)}, h) &= n^{-1/2}\sqrt{\mathcal{X}^2(x^{(0)}, h)} \\ &= (N_h^{\#}/n)^{1/2}(N_h^{\#})^{-1/2}\sqrt{\mathcal{X}^2(x^{(0)}, h)}.\end{aligned}$$

Since $N_h^{\#} \sim Bi(n, \int_{x^{(0)}-h}^{x^{(0)}+h} f(x)dx)$, we have

$$N_h^{\#}/n \to \int_{x^{(0)}-h}^{x^{(0)}+h} f(x)dx. \tag{7.15}$$

By Lemma 6.1, and Table (1),

$$(N_h^{\#})^{-1/2}\sqrt{\mathcal{X}^2(x^{(0)}, h)} \to \left(\sum_{r=+,-}\sum_{c=1}^{C} \frac{(p_{rc} - p_r q_c)^2}{p_r q_c}\right)^{\frac{1}{2}}, \tag{7.16}$$

where

$$p_{+c} = Pr(X > x_0, Y = c | X \in N_h(x^{(0)})), \quad p_{-c} = Pr(X \leq x_0, Y = c | X \in N_h(x^{(0)}));$$

for $r = +, -$ and $c = 1, \cdots, C$,

$$p_r = \sum_{c=1}^{C} p_{rc}; \quad q_c = \sum_{r=+,-} p_{rc},$$

Actually $p_+ = Pr(X > x_0 | X \in N_h(x^{(0)}))$, $p_- = Pr(X \leq x_0 | X \in N_h(x^{(0)}))$, $q_c = Pr(Y = c | X \in N_h(x^{(0)}))$.

By (7.15) and (7.16),

$$\hat{\zeta}(x^{(0)}, h) \rightarrow \Big(\int_{x^{(0)}-h}^{x^{(0)}+h} f(x)dx \Big)^{\frac{1}{2}} \Big(\sum_{r=+,-} \sum_{c=1}^{C} \frac{(p_{rc} - p_r q_c)^2}{p_r q_c} \Big)^{\frac{1}{2}} \equiv \zeta(x^{(0)}, h).$$

(2) Since X and Y are independent, $p_{rc} = p_r q_c$ for $r = +, -$ and $c = 1, \cdots, C$, then $\zeta(x^{(0)}, h) = 0$ for any h. Hence:

$$\zeta(x^{(0)}) = \sup_{h>0} \{\zeta(x^{(0)}, h)\} = 0.$$

(3) Recall that $p_c(x^{(0)}) = Pr(Y = c | x^{(0)})$. Without loss of generality, assume $p'_c(x^{(0)}) > 0$. Then there exists h, such that

$$Pr(Y = c | x^{(0)} < X < x^{(0)} + h) > Pr(Y = c | x^{(0)} - h < X < x^{(0)}),$$

which is equivelant to $p_{+c}/p_+ > p_{-c}/p_-$. This shows that $\zeta(x^{(0)}) > 0$ since by Lemma 6.1, $\zeta(x^{(0)}) = 0$ results in $p_{+c}/p_+ = p_{-c}/p_- = p_c$.

References

[1] Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.

[2] Breiman, L., Friedman, J. H., Olshen, R. A., Stone, J., 1984. Classification and regression trees. Wadsworth, Belmont.

[3] Doksum, K., Tang, S., Tsui, K.-W., 2008. Nonparametric variable selection: The earth algorithm. Journal of the American Statistical Association 103:(484), 1609–1620.

[4] Doksum, K. A., Samarov, A., 1995. Nonparametric estimation of global functionals and a measure of explanatory power of covariates in regression. Annals of Statistics 23, 1443–1473.

[5] Doksum, K. A., Schafer, C., 2006. Powerful choices: Tuning parameter selection based on power. Frontiers in Statistics: Dedicated to Peter Bickel, Editors J, Fan and H. L. Koul, Imperial College Press, 113–141.

[6] Friedman, J., 1991. Multivariate adaptive regression splines. The annals of statistics, 1–67.

[7] Gao, J., Gijbels, I., 2008. Bandwidth selection in nonparametric kernel testing. Journal of the American Statistical Association 103 (484), 1584–1594.

[8] Hall, P., 1992. The Bootstrap and Edgeworth Expansions. Springer, New York.

[9] Huang, L.-S., Chen, J., 2008. Analysis of variance, coefficient of determination, and f-test for local polynomial regression. Annals of Statistics 36, 2085–2109.

[10] Jing, B.-Y., Shao, Q.-M., Wang, Q., 2003. Self-normalized cramér-type large deviations for independent random variables. Annals of Probability 31, 2167–2215.

[11] Lin, D. Y., Zeng, D., 2011. Correcting for population stratification in genomewide association studies. Journal of the American Statistical Association 106, 997–1008.

[12] Loh, W.-Y., 2009. Improve the precision of classification trees. The Annals of Applied Statistics 3, 1710–1737.

[13] Price, A., Patterson, N., Plenge, R., Weinblatt, M., Shadick, N., Reich, D., 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics 38 (8), 904–909.

[14] Qiao, Z., Zhou, L., Huang, J., 2008. Linear discriminant analysis for high dimensional. Low Sample Size Data, special issue of World Congress on Engineering 2008.

[15] Quinlan, J. R., 1987. Simplifying decision trees. International Journal of Man-Machine Studies 27(3), 221–234.

[16] Schafer, C., Doksum, K. A., 2009. Selecting local models in multiple regression by maximizing power. Metrika 69, 283–304.

[17] Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. Royal. Statist. Soc. B. 58, 267–288.

[18] Wang, L., Shen, X., 2007. On l1-norm multi-class support vector machines: methodology and theory. Journal of the American Statistical Association 102, 583–594.

[19] Zhang, H. H., Liu, Y., Wu, Y., Zhu, J., 2008. Variable selection for the multicategory svm via adaptive sup-norm regularization. Electronic Journal of Statistics 2, 149–167.