

# Variable Selection in Sparse Semiparametric Single Index Models\*

Jianghao Chu<sup>†</sup>

Tae-Hwy Lee<sup>‡</sup>

Aman Ullah<sup>§</sup>

First Version: May 2018  
This Version: September 2018

## Abstract

In this paper we consider the “Regularization of Derivative Expectation Operator” (Rodeo) of Lafferty and Wasserman (2008) and propose a modified Rodeo algorithm for semiparametric single index models in big data environment with many regressors. The method assumes sparsity that many of the regressors are irrelevant. It uses a greedy algorithm, in that, to estimate the semiparametric single index model (SIM) of Ichimura (1993), all coefficients of the regressors are initially set to start from near zero, then we test iteratively if the derivative of the regression function estimator with respect to each coefficient is significantly different from zero. The basic idea of the modified Rodeo algorithm for SIM (to be called SIM-Rodeo) is to view the local bandwidth selection as a variable selection scheme which amplifies the coefficients for relevant variables while keeping the coefficients of irrelevant variables relatively small or at the initial starting values near zero. For sparse semiparametric single index models, the SIM-Rodeo algorithm is shown to attain consistency in variable selection. In addition, the algorithm is fast to finish the greedy steps. We compare SIM-Rodeo with SIM-Lasso method in Zeng et al. (2012). Our simulation results demonstrate that the proposed SIM-Rodeo method is consistent for variable selection and show that it has smaller integrated mean squared errors than SIM-Lasso.

*Key Words:* Single index model (SIM), Variable selection, Rodeo, SIM-Rodeo, Lasso, SIM-Lasso.

*JEL Classification:* C25, C44, C53, C55.

---

\*The authors are thankful to the co-editors and two anonymous referees for helpful comments, and the seminar participants at the *Advances in Econometrics* conference at the University of California, Irvine.

<sup>†</sup>Department of Economics, University of California, Riverside, CA 92521. E-mail: jianghao.chu@email.ucr.edu

<sup>‡</sup>Department of Economics, University of California, Riverside, CA 92521. E-mail: taelee@ucr.edu

<sup>§</sup>Department of Economics, University of California, Riverside, CA 92521. E-mail: aman.ullah@ucr.edu

# 1 Introduction

In a series of papers, Poirier (1980ab, 1994, 1996), Poirier and Rudd (1988), and Koop and Poirier (1993, 2004), Dale Poirier has made many seminal contributions to the issues of identification and inference of probit and logit models, in Bayesian and classical approaches, for parametric, semiparametric, partially linear models. This paper proposes a new method of variable selection for sparse single index models, that would be useful for parametric and semiparametric probit and logit models with many regressors.

Nadaraya (1964) and Watson (1964) propose the Nadaraya-Watson local constant kernel regression estimator. Kernel regression has been extremely popular for it is free of parametric assumptions. On the other hand, it suffers from computational complexity and the curse of dimensionality. Ichimura (1993) studies the semiparametric single index model (SIM) to overcome the curse of dimensionality by assuming that the true model is a function of an index which is a linear combination of the explanatory variables. Klein and Spady (1993) study a similar semiparametric single index model for binary outcomes and propose to estimate the model by maximum likelihood. However, these SIM methods gain limited improvements computationally over the local constant and local linear kernel regression and are still slow to implement.

Recent statistics and econometrics literature has been focusing on big data issues which are extremely difficult to solve with kernel regressions. To overcome this problem, under the sparsity assumption, several papers propose regularized SIM methods with penalty terms. See e.g. Huang, Horowitz, and Wei (2010). Su and Zhang (2014) provide a comprehensive review on those literature. However, those penalties may induce additional complexity in computation and lead to huge bias and variance when the ratio of information to noise is small. One such method that seems to be a natural way for SIM is a Lasso-type approach by Zeng, He, and Zhu (2012) for estimation and variable selection in SIM, which they termed as “SIM-Lasso”.

Meanwhile, there is a large volume of literature motivated by statistical machine learning, such as AdaBoost, Boosting, Support Vector Machine and Deep Neural Net. In particular, in this paper, we note that the method of Lafferty and Wasserman (2008), called the Regularization of Derivative

Expectation Operator (Rodeo), may be modified for SIM. Rodeo is a greedy algorithm for variable selection and estimation of the nonparametric regression function based on testing of marginal contribution of an additional variable in selecting relevant explanatory variables. A goal of this paper is to modify Rodeo so that it can be applied to semiparametric SIMs under sparsity. We will call the modified Rodeo for SIM as “SIM-Rodeo”.

The SIM-Rodeo method is able to distinguish relevant explanatory variables from irrelevant variables and gives a competitive estimator for the model. In addition, the algorithm finishes in a reasonable period of time. The method assumes sparsity under which most of the explanatory variables are irrelevant. We use a greedy algorithm that starts with a semiparametric SIM estimator (Ichimura 1993) that sets all coefficients  $\left(\theta_j = \frac{\beta_j}{h}\right)$  as zero which are the ratio of slope coefficients  $\beta_j$  to bandwidth  $h$  in the original Ichimura estimator. Then, we iteratively test if the derivative of the regression estimator with respect to each coefficient  $\theta_j$  is zero. The intuition is for a relevant explanatory variable, changing its coefficient would lead to a dramatic change in the value of the estimator. However, for an irrelevant variable, changing its coefficient would lead to ideally no change to the single index estimator. The impact of changing the coefficient to the attained estimator can be measured with the derivative of the estimator with respect to the coefficient. If the derivative with respect to one coefficient is zero, it implies the corresponding explanatory variable does not have a strong explanatory power on the dependent variable. And it will be seen as an irrelevant variable and given coefficient zero. However, if the derivative with respect to one coefficient is significantly different from zero, then we say the corresponding explanatory variable has a strong explanatory power on the dependent variable. Hence, it will be seen as a relevant explanatory variable and given coefficient greater than zero. The proposed procedure attains a solution path similar to the Least Angle Regression (Efron et al. 2004). The new method is superior to the usual Lasso type penalty (Zeng et al. 2012) in the sense that it does not introduce bias into the estimation process, is free of user-specific parameters and computationally more efficient. Simulation results show that the proposed method is consistent for variable selection and has smaller integrated mean squared errors (IMSE) than using Lasso penalty.

The rest of the paper is organized as follows. Section 2 introduces the intuition and algorithm

of the original Rodeo of Lafferty and Wasserman (2008). Section 3 sets up a model for the semi-parametric single index model of Ichimura (1993), introduces the SIM-Rodeo, and discusses the asymptotic properties of SIM-Rodeo in variable selection and estimation of the semiparametric single index model. Section 4 provides Monte Carlo simulation results for SIM-Rodeo in comparison with SIM-Lasso of Zeng et al (2012). Section 5 concludes.

## 2 Rodeo

This section introduces the idea behind the Regularization of Derivative Expectation Operator (Rodeo) proposed by Lafferty and Wasserman (2008). We first provide an illustration of the Rodeo algorithm, and then, a simple numerical example with one relevant explanatory variable and one irrelevant noise variable.

### 2.1 Algorithm

Let  $y_i \in \mathbb{R}$  be the dependent variable,  $X_i \in \mathbb{R}^k$  be an observation of  $k$  variables,  $X = (X_1', \dots, X_n)'$  be a matrix of  $n$  observations and  $x \in \mathbb{R}^k$  be a local estimation point.

The Rodeo algorithm uses the kernel estimator

$$\hat{m}_h(x) = \frac{\sum_{i=1}^n y_i K(X_i, x, h)}{\sum_{i=1}^n K(X_i, x, h)}, \quad (1)$$

where  $h$  is a vector of length  $k$  that is equal to the number of potential explanatory variable,  $h_j$  is the  $j$ th element of  $h$  that is corresponding to variable  $j$  and  $K(X_i, x, h)$  is the standard notation of a product kernel that takes the form

$$K(X_i, x, h) = \prod_{j=1}^k \kappa\left(\frac{X_{ij} - x_j}{h_j}\right), \quad (2)$$

where  $\kappa(\cdot)$  is usually given as a one-variable density function.  $X_{ij}$  is the  $i$ th observation of the  $j$ th variable and  $x_j$  is the  $j$ th variable of a local estimation point  $x$ . In what follows, we keep the same notation except that in the single index model, our kernel becomes a one-variable density function instead of a product kernel.

The Rodeo algorithm takes the derivative of the kernel estimator (1) with respect to each

bandwidth  $h_j$ . Let

$$\iota = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \quad (3)$$

and

$$W_x = \begin{pmatrix} K(X_1, x, h) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & K(X_n, x, h) \end{pmatrix}. \quad (4)$$

With fairly easy derivation, we can get the closed form of an estimate of the derivative

$$\begin{aligned} Z_j &\equiv \frac{\partial \hat{m}_h(x)}{\partial h_j} \\ &= (\iota' W_x \iota)^{-1} \iota' \frac{\partial W_x}{\partial h_j} y - (\iota' W_x \iota)^{-1} \iota' \frac{\partial W_x}{\partial h_j} \iota (\iota' W_x \iota)^{-1} \iota' W_x y \\ &\equiv \sum_{i=1}^n G_j(X_i, x, h) y_i, \end{aligned} \quad (5)$$

where  $y = (y_1, \dots, y_n)$  is a vector of observations on the dependent variable. The conditional variance of  $Z_j$  can be calculated by

$$\begin{aligned} s_j^2 &\equiv \text{Var}(Z_j | X) \\ &= \sigma^2 \sum_{i=1}^n G_j^2(X_i, x, h), \end{aligned} \quad (6)$$

where  $\sigma^2$  is the variance of the error term in the model. A detailed derivation can be found in Section 3 of Lafferty and Wasserman (2008). Here we skip the derivation for the kernel regression. However, we provide a detailed derivation for the single index model (SIM) in Section 4. Now we get all the ingredients of the Rodeo algorithm. The Rodeo algorithm is as follows.

---

**Algorithm 1. Rodeo (Lafferty and Wasserman 2008)**

---

1. Select constant  $0 < \alpha < 1$  and initial bandwidth

$$h_0 = \frac{c_0}{\log \log n}$$

where  $c_0 > 0$  is sufficiently large.

2. Initialize the bandwidths, and activate all covariates:

- (a)  $h_j = h_0, j = 1, \dots, k.$

- (b)  $\mathcal{A} = \{1, \dots, k\}.$

3. While  $\mathcal{A}$  is nonempty, do for each  $j \in \mathcal{A}$ :

- (a) Compute the estimated derivative and its conditional variance:  $Z_j$  and  $s_j$  using (5) and (6).

- (b) Compute the threshold  $\lambda_j = \hat{s}_j \sqrt{2 \log n}.$

- (c) If  $|Z_j| > \lambda_j$ , then set  $h_j \leftarrow \alpha h_j$ ; otherwise remove  $j$  from  $\mathcal{A}$  (i.e.,  $\mathcal{A} \leftarrow \mathcal{A} - \{j\}$ ).

4. Output bandwidths  $h^* = (h_1, \dots, h_k)$  and estimator  $\hat{m}_{h^*}(x)$  where  $\hat{m}_{h^*}(x)$  is the kernel estimator with bandwidth  $h^*$ .
- 

The basic idea of the Rodeo algorithm by Lafferty and Wasserman (2008) is to view the local bandwidth selection as variable selection in sparse nonparametric kernel regression models by shrinking the bandwidths for relevant variables while keeping the bandwidths of irrelevant variables relatively large. The Rodeo algorithm is greedy as it solves for the locally optimal path choice at each iteration and is shown to attain the consistency in mean square error when it is applied to sparse nonparametric local linear model (Lafferty and Wasserman 2008 Corollary 5.2).<sup>1</sup>

## 2.2 A Numerical Example

Now we give a numerical illustration of how Rodeo works. First we generate 100 data points from the DGP

$$y = \frac{1}{1 + e^{-x_1}} + u, \tag{7}$$

where  $x_1$  is a random variable following uniform distribution with range  $[-3, 3]$  and  $u$  is a random variable following the normal distribution with mean 0 and standard deviation 0.02. The generated

---

<sup>1</sup>We can also make Rodeo for local constant kernel regression models as we will demonstrate in this paper later.

data of  $x_1$  and  $y$  are shown in Figure 1.

Next, we generate an irrelevant variable  $x_2$  that follows the same distribution as  $x_1$  but is not included in the model. Thus,  $x_2$  and  $y$  are independent. The generated data of  $x_2$  and  $y$  are shown in Figure 2.

In the algorithm, we start by setting bandwidths  $h_j$  for all  $j$  large enough so that

$$\frac{X_{ij} - x_j}{h_j} \rightarrow 0 \quad \text{as } h_j \rightarrow \infty \quad \text{for all } i. \quad (8)$$

Hence,

$$K(X_i, x, h) \rightarrow \prod_{j=1}^k \kappa(0) \quad \text{as } h_j \rightarrow \infty \quad \text{for all } i. \quad (9)$$

For simplicity of illustration, we assume the kernel function is an indicator function  $\kappa(X_{ij}, x_j, h_j) = 1(|X_{ij} - x_j| < h_j)$ . This makes our estimate a simple average of the observations that satisfy  $|X_{ij} - x_j| < h_j$  for all  $j$ . If for all  $j$ ,  $|X_{ij} - x_j|$  is smaller than the bandwidth  $h_j$ , then we include observation  $i$  in the average. Otherwise, we exclude it. At the beginning, when the bandwidths are large enough, our estimate is the global mean since all observations are included in the estimate. However, if we shrink the bandwidth  $h_j$ , we exclude the observations whose  $X_{ij}$  has a distance greater than  $h_j$  from  $x_j$ . Hence, our estimate changes from the global mean  $\hat{m}_h(x) = \bar{y}$  to a local mean  $\hat{m}_h(x)$  as shown in Figure 3.

However, this holds only for  $x_1$ . Changing the bandwidth of  $x_2$  does not have the same effect. A larger bandwidth of  $x_2$  includes more observations whose  $X_{i2}$  is far away from  $x_2$ . If  $X_{i2}$  does not determine  $y_i$ , then including those observations does not affect  $\hat{m}_h(x)$ . In fact, from Figure 4 we can see that shrinking the bandwidth of  $x_2$  does not affect the value of the estimate. This observation gives us a criteria to distinguish between relevant explanatory variables and irrelevant variables.

### 3 Rodeo for Single Index Model (SIM-Rodeo)

In this section, we show that Rodeo can be modified for the sparse semiparametric linear single index models by considering the bandwidths as the inverse of the parameters which form the linear single index.

First, we give a short introduction to the general set up of the SIM model and the Ichimura (1993) estimator we use for estimation. We also give detailed intuition and description of our proposed greedy estimation procedure.

### 3.1 SIM

We consider a standard single index model,

$$y = m(x'\beta) + u, \quad (10)$$

where  $\beta = (\beta_1, \dots, \beta_k)$  is a vector of coefficients. Under the sparsity condition, we assume that  $\beta_j \neq 0$  for  $j \leq r$  and  $\beta_j = 0$  for  $j > r$ . We also assume that the random errors  $u$  are independent. However, we allow the presence of heteroskedasticity to encompass a large category of models for binary prediction, e.g. Logit and Probit models. The kernel estimator (Ichimura 1993) we use is as shown below

$$\hat{m}(x'\beta; h) = \frac{\sum_{i=1}^n y_i K\left(\frac{X_i'\beta - x'\beta}{h}\right)}{\sum_{i=1}^n K\left(\frac{X_i'\beta - x'\beta}{h}\right)}, \quad (11)$$

where  $K(\cdot)$  is a kernel function. The semiparametric kernel regression looks for the best  $\beta$  and  $h$  to minimize a weighted squared error loss. However, exact identification is not available. If one blows up  $\beta$  and  $h$  simultaneously by multiplying the same constant, the kernel estimator would yield identical estimates and losses. The standard identification approach is to set the first element of  $\beta$  to be 1 (Ichimura 1993).

As recent research pays more attention to high-dimensional data, most literature makes the sparsity assumption that many, if not most, of the elements of  $\beta$  are zero. The previous mentioned identification method appears to be unsuitable unless we have specific information that the true value of the element of  $\beta$  that we set to be 1 is not zero. The most popular regularization method, Lasso (Tibshirani 1996), also fails for the same reason. With  $L_1$  penalty, the algorithm can always achieve a lower loss by shrinking  $\beta$  and  $h$  while keeping the ratio of  $\frac{\beta}{h}$  constant. This would lead to a lower value in the penalty term without changing the value of the squared error term.

In terms of variable selection and prediction, we only need to focus on finding the best  $\theta \equiv \frac{\beta}{h}$ .



Hence, we can simplify the estimator to

$$\hat{m}(x'\theta) = \frac{\sum_{i=1}^n y_i K(X_i'\theta - x'\theta)}{\sum_{i=1}^n K(X_i'\theta - x'\theta)}. \quad (12)$$

Instead of the standard two stage estimation of Ichimura (1993), we introduce a test-based greedy approach similar to Lafferty and Wasserman (2008) where it was used for bandwidth selection in local linear regression. The intuition for the method is that if  $x_j$  is a relevant explanatory variable of  $y$ , then we would expect that increasing the magnitude of  $\theta_j$  would lead to a significant change in  $\hat{m}(x'\theta)$ . This can be seen as giving higher weights to the observations closer to  $x'\theta$  and lower weights to the observations further away from  $x'\theta$ . On the other hand, if  $x_j$  is not a relevant explanatory variable of  $y$ , then increasing the magnitude of  $\theta_j$  can be seen as randomly reassigning weights for the observations and will only result in a random (moderate) change in  $\hat{m}(x'\theta)$ . The influence of changing the magnitude of  $\theta_j$  on  $\hat{m}(x'\theta)$  can be measured as the derivative of  $\frac{\partial \hat{m}(x'\theta)}{\partial \theta_j}$ . Hence, we can test if  $x_j$  is a relevant explanatory variable by testing if  $\frac{\partial \hat{m}(x'\theta)}{\partial \theta_j}$  is statistically different from zero.

### 3.2 SIM-Rodeo

The basic idea of the modified Rodeo algorithm for SIM (SIM-Rodeo) is to view the local bandwidth selection as a variable selection in sparse semiparametric single index model. The SIM-Rodeo algorithm amplifies the inverse of the bandwidths for relevant variables while keeping the inverse of the bandwidths of irrelevant variables relatively small. The SIM-Rodeo algorithm is greedy as it solves for the locally optimal path choice at each iteration. SIM-Rodeo is able to distinguish truly relevant explanatory variables from noisy irrelevant variables. In addition, the algorithm is fast to finish the greedy steps.

Now we derive the Rodeo for Single Index Models. First we introduce some notation. Let

$$W_x = \begin{pmatrix} K(X_1'\theta - x'\theta) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & K(X_n'\theta - x'\theta) \end{pmatrix} \quad (13)$$

where  $K(\cdot)$  is the Gaussian kernel. The standard Ichimura (1993) estimator takes the form

$$\hat{m}(x'\theta) = \frac{\sum_{i=1}^n y_i K(X_i'\theta - x'\theta)}{\sum_{i=1}^n K(X_i'\theta - x'\theta)} = (\iota'W_x\iota)^{-1} \iota'W_x y. \quad (14)$$

The derivative of the estimator  $Z_j$  with respect to  $\theta_j$  is

$$Z_j \equiv \frac{\partial \hat{m}(x'\theta)}{\partial \theta_j} \quad (15)$$

$$\begin{aligned} &= (\iota'W_x\iota)^{-1} \iota' \frac{\partial W_x}{\partial \theta_j} y - (\iota'W_x\iota)^{-1} \iota' \frac{\partial W_x}{\partial \theta_j} \iota (\iota'W_x\iota)^{-1} \iota'W_x y \\ &= (\iota'W_x\iota)^{-1} \iota' \frac{\partial W_x}{\partial \theta_j} (y - \iota \hat{m}(x'\theta)). \end{aligned} \quad (16)$$

For the ease of computation, let

$$L_j = \begin{pmatrix} \frac{\partial \log K(X_1'\theta - x'\theta)}{\partial \theta_j} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{\partial \log K(X_n'\theta - x'\theta)}{\partial \theta_j} \end{pmatrix}. \quad (17)$$

Note that

$$\frac{\partial W_x}{\partial \theta_j} = W_x L_j, \quad (18)$$

which appears in equation (16). With the Gaussian kernel,  $K(t) = e^{-\frac{t^2}{2}}$ , then  $L_j$  becomes

$$\begin{aligned} L_j &= \begin{pmatrix} -\frac{1}{2} \frac{\partial (X_1'\theta - x'\theta)^2}{\partial \theta_j} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & -\frac{1}{2} \frac{\partial (X_n'\theta - x'\theta)^2}{\partial \theta_j} \end{pmatrix} \\ &= \begin{pmatrix} -(X_1'\theta - x'\theta)(X_{1j} - x_j) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & -(X_n'\theta - x'\theta)(X_{nj} - x_j) \end{pmatrix}, \end{aligned}$$

where  $X_{1j}$  and  $X_{nj}$  are the  $j$ th elements of vectors  $X_1$  and  $X_n$ . And  $x_j$  is the  $j$ th element of vector  $x$ . To simplify the notation, let  $B_x = (\iota'W_x\iota)^{-1} \iota'W_x$ . Then, the derivative  $Z_j$  becomes

$$\begin{aligned} Z_j &= (\iota'W_x\iota)^{-1} \iota' \frac{\partial W_x}{\partial \theta_j} (y - \iota \hat{m}(x'\theta)) \\ &= B_x L_j (I - \iota B_x) y \\ &\equiv G_j(x, \theta) y. \end{aligned} \quad (19)$$

Note that now we are using a different notation with  $G_j(\cdot)$ . In Section 2,  $G_j(\cdot)$  is a three-argument function and  $G_j(X_i, x, h)$  is a scalar. However, in this section,  $G_j(\cdot)$  is a two-argument function and  $G_j(x, \theta)$  is a vector of length  $n$ . We are aware that this change of notation may cause confusion. Nevertheless,  $G_j(\cdot)$  in Sections 2 and 3 play the same role as the weights of  $y$  in  $Z_j$ . So we think sticking with  $G_j(\cdot)$  would be easier for the readers to understand and compare Rodeo and SIM-Rodeo as long as the difference is pointed out and noticed by the readers.

Next, we give the conditional expectation and variance of  $Z_j$ .

$$Z_j = G_j(x, \theta) y = G_j(x, \theta) (m(x'\beta) + u), \quad (20)$$

$$E(Z_j|X) = E(G_j(x, \theta) (m(x'\beta) + u) | X) = G_j(x, \theta) m(x'\beta), \quad (21)$$

$$\text{Var}(Z_j|X) = \text{Var}(G_j(x, \theta) (m(x'\beta) + u) | X) = \boldsymbol{\sigma}' G_j(x, \theta)' G_j(x, \theta) \boldsymbol{\sigma}, \quad (22)$$

where  $\boldsymbol{\sigma} = (\sigma(u_1), \dots, \sigma(u_n))'$  is the vector of standard deviations of  $u$ . In the algorithm, it is necessary to insert an estimate of  $\sigma$ . In Algorithm 1, Lafferty and Wasserman (2008) suggest to use a generalized estimator of Rice (1984) under homoskedasticity. In our Algorithm 2, we allow the errors to be heteroskedastic as in Logit and Probit models and estimate  $\sigma(u_i)$  using the estimator  $\hat{\sigma}^2(u_i) = m(x_i'\hat{\theta})(1 - m(x_i'\hat{\theta}))$ .

SIM-Rodeo is described in Algorithm 2, which is a modified algorithm of Rodeo (Lafferty and Wasserman 2008).

---

**Algorithm 2. SIM-Rodeo**

---

1. Select a constant  $0 < \alpha < 1$  and the initial value

$$\theta_0 = c_0 \log \log n$$

where  $c_0$  is sufficiently small. Compute  $Z_j$  with  $\theta_j = \theta_0$  for all  $j$ .

2. Initialize the coefficients  $\theta$ , and activate all covariates:

- (a)  $\theta_j = \begin{cases} \theta_0 & Z_j > 0 \\ -\theta_0 & \text{otherwise,} \end{cases} \quad j = 1, \dots, k.$

- (b)  $\mathcal{A} = \{1, \dots, k\}.$

3. While  $\mathcal{A} \neq \emptyset$  is nonempty, do for each  $j \in \mathcal{A}$ :

- (a) Compute  $Z_j$  and  $s_j = \sqrt{\text{Var}(Z_j|X)}$  using (19) and (22) respectively.

- (b) Compute the threshold  $\lambda_j = s_j \sqrt{2 \log n}.$

- (c) If  $|Z_j| > \lambda_j$ , then set  $\theta_j \leftarrow \frac{\theta_j}{\alpha}$ ; Otherwise, remove  $j$  from  $\mathcal{A}$  (i.e.,  $\mathcal{A} \leftarrow \mathcal{A} - \{j\}$ ).

4. Output  $\hat{\theta} = (\theta_1, \dots, \theta_k)$  and estimator  $\hat{m}(x' \hat{\theta})$ .
- 

Notice that in Algorithm 1, when selecting bandwidth for local linear and local constant regression, the bandwidth is always positive. Hence, we do not have to worry about the sign of the bandwidth. However, in our single index model,  $\theta$  is the ratio of  $\beta$  and the bandwidth. Since  $\beta$  could be either positive or negative,  $\theta$  could also take positive or negative values. In Algorithm 2, we propose to use the sign of the derivative estimate  $Z_j$  as the sign of  $\theta_j$ . Our method is based on the observation that if  $\theta_j$  and  $\theta_{j'}$  have the same sign, then their respective  $Z$  statistic  $Z_j$  and  $Z_{j'}$  will also have the same sign. Hence, SIM-Rodeo will give relatively correct signs to each  $\theta$ , i.e. all the positive  $\theta$  will be given the same sign and all the negative  $\theta$  will be given the same sign. A similar method is applied by Ichimura (1993) where the value positive one is given to the first  $\beta$  to ensure identification. Under the sparsity assumption, it is problematic to arbitrarily assign a magnitude greater than zero to any  $\theta$  since the true value could be zero. However, it is safe to assume the sign of one of the  $\theta$  to be positive or negative since positive zero and negative zero will not affect the relative scale of  $\theta$ . Once again, due to the the identification issue with single index model, exact identification of  $\theta$  is not available. However, signs of  $\theta$  can be obtained relatively.

We start by setting  $\theta_j = \theta_0$  that is close to zero. Hence,  $(X_i'\theta - x'\theta)$  are close to zero and  $K(X_i'\theta - x'\theta)$  are close to  $K(0)$ . This means our estimator starts with the simple average of all observations,  $\bar{y}$ . If the derivative of  $\theta_j$  is statistically different from zero. We amplify  $\theta_j$ . If  $x_j$  is indeed a relevant explanatory variable, then the weights  $K(X_i'\theta - x'\theta)$  change according to  $x_j$ . The estimator will give higher weights to observations close to  $x'\theta$  and lower weights to observations away from  $x'\theta$ .

### 3.3 Asymptotic Properties of SIM-Rodeo

We make the following assumptions.

**Assumptions.** (A1) *The density  $f(x)$  of  $(x_1, \dots, x_k)$  is uniform on the unit cube.*

(A2)  *$\liminf_{n \rightarrow \infty} \min_{1 \leq j \leq r} |m_{jj}(\cdot)| > 0$  where  $m_{jj}(\cdot)$  is the second derivative of  $m(\cdot)$ .*

(A3) *All derivatives of  $m(\cdot)$  up to and including fourth order are bounded.*

Assumption (A1) greatly simplifies the proof of Theorem 1. However, it is not necessary as shown in our Monte Carlo designs where  $x$ 's are not uniform distributed. Assumption (A2) is crucial for SIM-Rodeo. As shown in Lemma 1, the expectation of  $Z_j$  for a relevant variable will be zero if the second derivative of  $m(\cdot)$  is zero. As a result, we will not be able to distinguish relevant variables from irrelevant variables through  $Z_j$  since in both cases, the expectation of  $Z_j$  is zero.

In the statement of Theorem 1, we follow the notation of Lafferty and Wasserman (2008) and write  $Y_n = \tilde{O}(a_n)$  to mean that  $Y_n = O(b_n a_n)$  where  $b_n$  is logarithmic in  $n$ . And we write  $a_n = \Omega(b_n)$  if  $\liminf_n \left| \frac{a_n}{b_n} \right| > 0$  and  $a_n = \tilde{\Omega}(b_n)$  if  $a_n = \Omega(b_n c_n)$  where  $c_n$  is logarithmic in  $n$ .

**Theorem 1.** *Suppose that Assumptions (A1), (A2) and (A3) hold. In addition, suppose that*

$$\min_{j \leq r} |m_{jj}(x'\theta)| = \tilde{\Omega}(1) \tag{23}$$

and

$$\max_{j \leq r} |m_{jj}(x'\theta)| = \tilde{O}(1). \tag{24}$$

Then the SIM-Rodeo outputs  $\hat{\theta}$  satisfying

$$\Pr(\theta_j = \theta_0 \text{ for all } j > r) \rightarrow 1 \text{ as } n \rightarrow \infty \quad (25)$$

and

$$\Pr(\theta_j > \theta_0 \text{ for all } j \leq r) \rightarrow 1 \text{ as } n \rightarrow \infty. \quad (26)$$

*Proof.* See Appendix. □

Theorem 1 shows that under the given assumptions and conditions, the coefficients  $\theta$  for relevant variables will always be amplified while the coefficients  $\theta$  for irrelevant variables will always stay at the initial value. Hence, we are able to consistently select the relevant variables by checking whether the coefficients  $\theta$  is amplified by the SIM-Rodeo.

**Remark 1.** *Theorem 1 shows the consistency of the variable selection by the SIM-Rodeo. However, the consistency of estimating  $m(\cdot)$  is not proved in Theorem 1. We conjecture that the consistency holds as supported by our simulation results. We leave this extension for a future work.*

**Remark 2.** *An alternative consistent estimation procedure is as follows. First, we use the proposed SIM-Rodeo algorithm for variable selection. Then, we use the selected explanatory variables to estimate  $\hat{\beta}$  and  $\hat{m}(x'\hat{\beta}; h)$  by using either Ichimura (1993) or Klein and Spady (1993). Since Theorem 1 shows that the SIM-Rodeo consistently selects the relevant variables, the methods of Ichimura (1993) and Klein and Spady (1993) would yield consistent estimation of  $m(\cdot)$  after the consistent variable selection via the SIM-Rodeo.*

## 4 Monte Carlo

This section examines the performance of SIM-Rodeo using Monte Carlo simulation compared with SIM-Lasso (Zeng, He and Zhu 2012) and Maximum Likelihood (Klein and Spady 1993). We first describe the designs of the DGPs. Then a brief introduction of SIM-Lasso is provided. At the end of this section, we give a comprehensive discussion on the simulation results.

## 4.1 Simulation Designs

We follow the simulation designs of Klein and Spady (1993) where the data generating process (DGP) is given by

$$y_i^* = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_k x_{ik} + u_i \quad \text{for } i = 1, \dots, n \quad (27)$$

where

$$\beta_j = \begin{cases} 1 & \text{if } j = 1, 2; \\ 0 & \text{otherwise.} \end{cases} \quad (28)$$

The observed variable  $y_i$  is generated by

$$y_i = \begin{cases} 1 & \text{if } y_i^* \geq 0; \\ 0 & \text{otherwise.} \end{cases} \quad (29)$$

The  $x$ 's are independently and identically distributed.  $x_1$  is a chi-squared variate with 3 degrees of freedom truncated at 6 and standardized to have zero mean and unit variance;  $x_2$  is a standard normal variate truncated at  $\pm 2$  and similarly standardized. All the other  $x$ 's are irrelevant variables and follow uniform distribution between  $-2$  and  $2$ .

We consider two link functions as shown in Figure 5 (Design 1 and Design 2). In Design 1, the  $u_i$ 's are standard normal. In Design 2, they are normal with mean zero and variance  $0.25(1 + v_i^2)^2$  where  $v_i \equiv \beta_1 x_{i1} + \beta_2 x_{i2}$ . In both designs,  $u_i$ 's are independently distributed.

The probability  $\Pr(y = 1|v)$  of the two designs are shown in Figure 5. Design 1 is the standard Probit model. Design 2 is different from Design 1 in the sense that it is not monotone and is steeper in the tails. Hence, Design 2 has a larger curvature than Design 1 on average. As a result, SIM-Rodeo is expected to perform better under Design 2 since Assumption (A2) and Conditions (23) and (24) require the second derivative of the link function to be greater than zero.

## 4.2 SIM-Lasso

We show results for SIM-Rodeo together with SIM-Lasso (Zeng et al 2012) to check the relative efficiency of SIM-Rodeo. The SIM-Lasso is introduced as an application of the Lasso penalty under the framework of Semiparametric Single Index Models for variable selection and estimation. Zeng,

He and Zhu (2012) propose to solve the following minimization problem

$$\min_{a,b,\beta,\|\beta\|=1} \sum_{j=1}^n \sum_{i=1}^n [y_i - a_j - b_j \beta' (X_i - X_j)]^2 w_{ij} + \lambda \sum_{j=1}^n |b_j| \sum_{p=1}^k |\beta_p| \quad (30)$$

where  $\lambda$  is a hyper parameter as in standard Lasso practices and

$$w_{ij} = \frac{K\left(\frac{X_i' \beta - X_j' \beta}{h}\right)}{\sum_{q=1}^n K\left(\frac{X_q' \beta - X_j' \beta}{h}\right)}. \quad (31)$$

The authors provide their code in the supplemental materials of their paper which is available on the website of *Journal of Computational and Graphical Statistics*.

### 4.3 Results

We report  $\theta \left(= \frac{\beta}{h}\right)$  from SIM-Rodeo and SIM-Lasso both for the estimator of Ichimura (1993). The results of the simulations are presented in Tables 1-4. Notice that for both algorithms, large values of  $\theta$  indicate that the associated variables are relevant explanatory variables while small values of  $\theta$  indicate that the associated variable are irrelevant variables. In both designs, only the first two variables are relevant explanatory variables as in the description of the DGPs. We consider different values for  $n \in \{100, 200\}$  and  $k \in \{5, 20\}$  where  $n$  is the number of observations in the training sample and  $k$  is the total number of relevant explanatory variables and irrelevant variables for each observation. We also present results using the maximum likelihood (ML) of Klein and Spady (1993) for the low dimension case ( $k = 5$ ). We skip the ML for the high dimension case ( $k = 20$ ) since maximum likelihood suffers dramatically from the curse of dimensionality. The maximum likelihood is strictly dominated by the other methods even in the low dimension case. And theoretically, it would only get worse when dimension increases. We report the Monte Carlo average of the value of  $\hat{\theta}$  obtained by the methods and the integrated mean squared error

$$\text{IMSE} = \int \left( \hat{m}(x' \hat{\theta}) - m(x' \beta) \right)^2 f(x) dx \quad (32)$$

of the estimate  $\hat{m}(x' \hat{\theta})$  using the  $\hat{\theta}$  obtained where  $f(x)$  is the probability density function of  $x$  as in Assumption (A1).

From the simulation results, we can see that under the sparsity condition, SIM-Rodeo and



SIM-Lasso both outperform the traditional maximum likelihood method of Klein and Spady (1993) which does not take advantage of the sparsity structure in the DGP. While among the two methods that take into account the sparsity structure, SIM-Rodeo outperforms SIM-Lasso in both variable selection and estimation. In Design 2, SIM-Rodeo dominates SIM-Lasso in small and large samples and various degrees of sparsity. In addition, SIM-Rodeo works better under Design 2 than Design 1. This is consistent with our analytical result since the expectation of the derivative estimate  $Z_j$  is depending on the second derivative of  $m(\cdot)$ . When the second derivative of  $m(\cdot)$  is close to zero, the expectations of  $Z_j$  of relevant explanatory variables are also close to zero which makes the difference between relevant variables and irrelevant variables smaller. Moreover, the conditions (23) and (24) in Theorem 1 states that SIM-Rodeo requires a larger value for the second derivative of  $m(\cdot)$  when the number of observations  $n$  increases. As a result, when  $n$  increases from 100 to 200, the already small second derivative in Design 1 becomes even more problematic. That is why in Table 1, the IMSE of SIM-Rodeo for Design 1 does not benefit from the increase of sample size. In summary, SIM-Rodeo and SIM-Lasso both have excellent performance in terms of variable selection. However, SIM-Rodeo generally has a smaller IMSE than SIM-Lasso. Maximum likelihood should not be used when sparsity is assumed since both SIM-Rodeo and SIM-Lasso have considerably better performance.

## 5 Conclusions

The basic idea of the Rodeo algorithm by Lafferty and Wasserman (2008) is to view the local bandwidth selection as variable selection in sparse nonparametric kernel regression by shrinking the bandwidths for relevant variables while keeping the bandwidths of irrelevant variables relatively large. The Rodeo algorithm is greedy as it solves the locally optimal path choice at each stage which is shown to attain the asymptotic optimality in mean square error for sparse nonparametric local linear or local constant kernel regression models (Lafferty and Wasserman 2008 Corollary 5.2).

In this paper, we propose a new algorithm, based on the Rodeo, for variable selection and estimation for the sparse semiparametric linear single index models by viewing the bandwidths as the inverse of the parameters which form the linear single index. The basic idea of the modified

Rodeo algorithm for SIM (which we call SIM-Rodeo) is to view the local bandwidth selection as a variable selection in sparse semiparametric single index model by amplifying the inverse of the bandwidths for relevant variables while keeping the inverse of the bandwidths of irrelevant variables relatively small. The SIM-Rodeo algorithm is greedy as it solves the locally optimal path choice at each stage which can also be shown to attain the asymptotic optimality in mean square error for sparse semiparametric single index models. The SIM-Rodeo method is able to distinguish truly relevant explanatory variables from noisy irrelevant variables and gives a "competitive" estimator for the model. In addition, the algorithm is fast to finish the greedy steps.

We compare the SIM-Rodeo with a Lasso-type approach by Zeng et al (2012) for estimation and variable selection in SIM, which Zeng et al (2012) call SIM-Lasso. Our Monte Carlo simulation shows that SIM-Rodeo outperforms SIM-Lasso in variable selection and also in estimation. The new method is superior to the usual Lasso type penalty in estimation because SIM-Rodeo does not introduce bias from using the additive Lasso penalty and is computationally more efficient. Simulation results also show that the proposed SIM-Rodeo is consistent for variable selection and has smaller integrated mean squared errors than using SIM-Lasso.

## Appendix: Proof of Theorem 1

We follow the notation of Lafferty and Wasserman (2008) and write  $Y_n = \tilde{O}_P(a_n)$  to mean that  $Y_n = O_P(b_n a_n)$  where  $b_n$  is logarithmic in  $n$ . And we write  $a_n = \Omega(b_n)$  if  $\liminf_n \left| \frac{a_n}{b_n} \right| > 0$  and  $a_n = \tilde{\Omega}(b_n)$  if  $a_n = \Omega(b_n c_n)$  where  $c_n$  is logarithmic in  $n$ .

Define

$$\mu_j(\theta) = \frac{\partial}{\partial \theta_j} E[\hat{m}_\theta(x) - m(x) | X_1, \dots, X_n],$$

which is the derivative of the conditional bias. The first lemma analyzes  $\mu_j(\theta)$  and  $\mathbb{E}(\mu_j(\theta))$  under the assumption that  $f$  is uniform. The second lemma analyzes the variance. The third lemma bounds the probabilities  $\mathbb{P}(|Z_j| \geq \lambda_j)$  in terms of tail inequalities for standard normal variables.

In each of these lemmas, we make the following assumptions. We assume that  $f$  is uniform,  $K$  is a Gaussian kernel, and  $\alpha > 1$ . Moreover, without loss of generality, we make use of the following set  $\mathcal{B}$  of coefficients where  $\theta_0 > 0$

$$\mathcal{B} = \left\{ \theta = (\theta_1, \dots, \theta_k) = \left( \underbrace{\alpha^{t_1} \theta_0, \dots, \alpha^{t_r} \theta_0}_{r \text{ terms}}, \underbrace{\theta_0, \dots, \theta_0}_{k-r \text{ terms}} \right) : 0 \leq t_j \leq T_n, j = 1, \dots, r \right\},$$

where  $T_n \leq c_1 \log n$ . Finally, we assume that

$$\begin{aligned} r &= O(1), \\ k &= O\left(\frac{\log n}{\log \log n}\right), \\ \theta_0 &= c_0 \log \log n. \end{aligned}$$

The proofs of the lemmas can be found in Lafferty and Wasserman (2008).

**Lemma 1.** *For each  $\theta \in \mathcal{B}$ ,*

$$\mathbb{E}(\mu_j(\theta)) = \begin{cases} \frac{\nu_2 m_{jj}(x\theta)}{\theta_j} + \frac{g_j(x_R \theta_R)}{\theta_j}, & j \leq r, \\ 0, & j > r, \end{cases}$$

where  $\nu_2 I = \int uu^T K(u) du$  and  $g_j(x_R \theta_R)$  depends only on the relevant variables and bandwidths, and satisfies

$$|g_j(x_R \theta_R)| = O\left(\sum_{l \leq r} \sup_x \frac{|m_{jll}(x\theta)|}{\theta_l^2}\right).$$

Furthermore, for any  $\delta > 0$ ,

$$\Pr \left( \max_{\substack{\theta \in \mathcal{B} \\ 1 \leq j \leq k}} \frac{|\mu_j(\theta) - \mathbb{E}(\mu_j(\theta))|}{s_j(\theta)} > \frac{\sqrt{\delta \log n}}{\log \log n} \leq \frac{1}{n^{\delta \sigma^2} / (8c_0)} \right)$$

where

$$s_j^2(\theta) = \frac{C\theta_j^2}{n} \prod_{l=1}^k \theta_l,$$

with

$$C = \sigma^2 \frac{\int K^2(u) du}{f(x)}.$$

**Lemma 2.** Let  $\nu_j(\theta) = \text{Var}(Z_j | X_1, \dots, X_n)$ . Then

$$\Pr \left( \max_{\substack{\theta \in \mathcal{B} \\ 1 \leq j \leq k}} \left| \frac{\nu_j(\theta)}{s_j^2(\theta)} - 1 \right| > \varepsilon \right) \rightarrow 0,$$

for all  $\varepsilon > 0$ .

**Lemma 3.** For any  $c > 0$  and each  $j > r$ ,

$$\Pr(|Z_j(\theta_0)| > \lambda_j(\theta_0)) = o\left(\frac{1}{n^c}\right).$$

Uniformly for  $\theta \in \mathcal{B}$ ,  $c > 0$  and  $j \leq r$ ,

$$\Pr(|Z_j(\theta)| < \lambda_j(\theta)) \leq \Pr\left(N(0, 1) > \frac{\nu_j |m_{jj}(x\theta)| + z_n}{s_j(\theta) \theta_j}\right) + o\left(\frac{1}{n^c}\right),$$

where  $z_n = O(\theta_j^{-3})$ .

*Proof of Theorem 1.* Let  $\mathcal{A}_t$  be the active set at step  $t$ . Define  $S_t$  to be the event that  $\mathcal{A}_t = \{1, \dots, r\}$ . We want to show that

$$\Pr(S_1) \rightarrow 1,$$

from which the theorem follows.

Fix  $c > 0$ . In what follows, we let  $\xi_n(c)$  denote a term that is  $o(n^{-c})$ ; we will suppress the dependence on  $c$  and simply write  $\xi_n$ .

At Step  $t = 1$ , define the event

$$B_1 = \{|Z_j| > \lambda_j \text{ for all } j \leq r\} \cap \{|Z_j| < \lambda_j \text{ for all } j > r\}.$$

Thus,  $A_1 = B_1$ . We claim that

$$\Pr(B_1^c) \leq O\left(\frac{1}{n}\right) + \xi_n.$$

From Lemma 3, when  $j > r$ ,

$$\Pr\left(\max_{j>r} |Z_j| > \lambda_j\right) \leq \sum_{j=r+1}^k \Pr(|Z_j| > \lambda_j) \leq d\xi_n = \xi_n.$$

When  $j \leq r$ ,

$$\Pr(|Z_j| < \lambda_j \text{ for some } j \leq r) \leq O\left(\frac{1}{n}\right) + \xi_n.$$

Hence,

$$\Pr(\theta_j = \theta_0 \text{ for all } j > r) \rightarrow 1 \text{ as } n \rightarrow \infty.$$

and

$$\Pr(\theta_j > \theta_0 \text{ for all } j \leq r) \rightarrow 1 \text{ as } n \rightarrow \infty$$

□

## References

- Efron, Bradley, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. 2004. "Least Angle Regression." *The Annals of Statistics* 32 (2): 407–499.
- Huang, J., J.L. Horowitz, and F. Wei. 2010. "Variable Selection in Nonparametric Additive Models." *The Annals of Statistics* 38: 2282–2313.
- Ichimura, Hidehiko. 1993. "Semiparametric Least Squares and Weighted SLS Estimation of Single Index Models." *Journal of Econometrics* 58 (1–2): 71–120.
- Klein, Roger W., and Richard H. Spady. 1993. "An Efficient Semiparametric Estimator for Binary Response Models." *Econometrica* 61 (2): 387–421.
- Koop, Gary and Dale J. Poirier. 1993. "Bayesian Analysis of Logit Models Using Natural Conjugate Priors." *Journal of Econometrics* 56 (3): 323–340.
- Koop, Gary and Dale J. Poirier. 2004. "Bayesian Variants of Some Classical Semiparametric Regression Techniques." *Journal of Econometrics* 123 (2): 259–282.
- Lafferty, John and Larry Wasserman. 2008. "Rodeo: Sparse, Greedy Nonparametric Regression." *The Annals of Statistics* 36 (1): 28–63.
- Nadaraya, E. A. 1964. "On Estimating Regression." *Theory of Probability & Its Applications* 9 (1). Society for Industrial and Applied Mathematics: 141–142.
- Poirier, Dale J. 1980a. "A Lagrange Multiplier Test for Skewness in Binary Logit Models." *Economics Letters* 5 (2): 141–143.
- Poirier, Dale J. 1980b. "Partial Observability in Bivariate Probit Models." *Journal of Econometrics* 12 (2): 209–217.
- Poirier, Dale J. and Paul A. Ruud. 1988. "Probit with Dependent Observations." *The Review of Economic Studies* 55 (4): 593–614.
- Poirier, Dale J. 1994. "Jeffreys' Prior for Logit Models." *Journal of Econometrics* 63 (2): 327–339.
- Poirier, Dale J. 1996. "A Bayesian Analysis of Nested Logit Models." *Journal of Econometrics* 75 (1): 163–181.

- Rice, John. 1984. "Bandwidth Choice for Nonparametric Regression." *The Annals of Statistics* 12 (4): 1215–1230.
- Ruppert, D. and M. P. Wand. 1994. "Multivariate Locally Weighted Least Squares Regression." *The Annals of Statistics* 22 (3). Institute of Mathematical Statistics: 1346–1370.
- Stone, C. J. 1977. "Consistent nonparametric regression." *The Annals of Statistics* 5 (4): 595-620.
- Stone, C. J. 1980. "Optimal rates of convergence for nonparametric estimators." *The Annals of Statistics* 8 (6): 1348-1360.
- Stone, C. J. 1982. "Optimal global rates of convergence for nonparametric regression." *The Annals of Statistics* 10 (4): 1040-1053.
- Su, Liangjun and Yonghui Zhang. 2014. "Variable Selection in Nonparametric and Semiparametric Regression Models." *The Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics*, Chapter 9: 249–307. Edited by Jeffrey Racine, Liangjun Su, and Aman Ullah. Oxford University Press.
- Tibshirani, R. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society Series B* 58(1): 267-288.
- Watson, Geoffrey S. 1964. "Smooth Regression Analysis." *Sankhyā: The Indian Journal of Statistics, Series A* (1961-2002). Indian Statistical Institute.
- Zeng, Peng, Tianhong He, and Yu Zhu. 2012. "A Lasso-Type Approach for Estimation and Variable Selection in Single Index Models." *Journal of Computational and Graphical Statistics* 21 (1): 92–109.

Table 1: Design 1 ( $k = 5$ )

		$\theta_1$	$\theta_2$	$\theta_3$	$\theta_4$	$\theta_5$	IMSE of $m(x\theta)$
$n = 100$	Rodeo	0.5739	0.3422	0.0713	0.0693	0.0724	0.0774
	Lasso	0.6032	0.5822	0.0223	0.0228	0.0235	0.1136
	ML	15.7902	11.3961	2.2156	2.2339	2.2334	0.3103
$n = 200$	Rodeo	0.8063	0.5095	0.1811	0.1894	0.1904	0.0780
	Lasso	0.6572	0.6348	0.0142	0.0141	0.0142	0.0740
	ML	13.6022	11.6887	1.5990	1.6322	1.6093	0.2356

Table 2: Design 2 ( $k = 5$ )

		$\theta_1$	$\theta_2$	$\theta_3$	$\theta_4$	$\theta_5$	IMSE of $m(x\theta)$
$n = 100$	Rodeo	0.2486	0.1452	0.0160	0.0120	0.0057	0.0474
	Lasso	0.5241	0.4919	0.0332	0.0334	0.0357	0.1137
	ML	10.0696	4.5824	1.5773	1.6003	1.5993	0.1858
$n = 200$	Rodeo	0.5022	0.2803	0.0296	0.0393	0.0426	0.0369
	Lasso	0.6547	0.6031	0.0209	0.0192	0.0207	0.0616
	ML	7.8625	4.7588	0.8896	0.8920	0.9034	0.1321



Table 3: Design 1 ( $k = 20$ )

		$\theta_1$	$\theta_2$	$\theta_3$	$\theta_4$	$\theta_5$	$\theta_6$	$\theta_7$	$\theta_8$	$\theta_9$	$\theta_{10}$	IMSE
$n = 100$	Rodeo	0.2349	0.2117	0.0032	0.0048	0.0027	0.0049	0.0013	0.0080	0.0009	0.0023	0.1487
		$\theta_{11}$	$\theta_{12}$	$\theta_{13}$	$\theta_{14}$	$\theta_{15}$	$\theta_{16}$	$\theta_{17}$	$\theta_{18}$	$\theta_{19}$	$\theta_{20}$	
	Rodeo	0.0093	0.0066	0.0037	0.0045	0.0026	0.0036	0.0016	0.0033	0.0030	0.0016	
		$\theta_1$	$\theta_2$	$\theta_3$	$\theta_4$	$\theta_5$	$\theta_6$	$\theta_7$	$\theta_8$	$\theta_9$	$\theta_{10}$	IMSE
	Lasso	0.4140	0.3904	0.0036	0.0036	0.0044	0.0045	0.0046	0.0048	0.0047	0.0047	0.2105
		$\theta_{11}$	$\theta_{12}$	$\theta_{13}$	$\theta_{14}$	$\theta_{15}$	$\theta_{16}$	$\theta_{17}$	$\theta_{18}$	$\theta_{19}$	$\theta_{20}$	
Lasso	0.0063	0.0041	0.0056	0.0047	0.0048	0.0057	0.0034	0.0047	0.0045	0.0048		
$n = 200$		$\theta_1$	$\theta_2$	$\theta_3$	$\theta_4$	$\theta_5$	$\theta_6$	$\theta_7$	$\theta_8$	$\theta_9$	$\theta_{10}$	IMSE
	Rodeo	0.4191	0.3404	0.0036	0.0104	0.0182	0.0120	0.0109	0.0118	0.0115	0.0060	0.1238
		$\theta_{11}$	$\theta_{12}$	$\theta_{13}$	$\theta_{14}$	$\theta_{15}$	$\theta_{16}$	$\theta_{17}$	$\theta_{18}$	$\theta_{19}$	$\theta_{20}$	
	Rodeo	0.0077	0.0123	0.0086	0.0082	0.0074	0.0147	0.0105	0.0156	0.0105	0.0081	
		$\theta_1$	$\theta_2$	$\theta_3$	$\theta_4$	$\theta_5$	$\theta_6$	$\theta_7$	$\theta_8$	$\theta_9$	$\theta_{10}$	IMSE
	Lasso	0.4308	0.4120	0.0021	0.0029	0.0026	0.0024	0.0023	0.0023	0.0018	0.0019	0.1572
	$\theta_{11}$	$\theta_{12}$	$\theta_{13}$	$\theta_{14}$	$\theta_{15}$	$\theta_{16}$	$\theta_{17}$	$\theta_{18}$	$\theta_{19}$	$\theta_{20}$		
Lasso	0.0021	0.0025	0.0026	0.0026	0.0020	0.0027	0.0022	0.0026	0.0025	0.0025		

Table 4: Design 2 ( $k = 20$ )

		$\theta_1$	$\theta_2$	$\theta_3$	$\theta_4$	$\theta_5$	$\theta_6$	$\theta_7$	$\theta_8$	$\theta_9$	$\theta_{10}$	IMSE
$n = 100$	Rodeo	0.0692	0.0479	0.0007	0.0011	0.0000	0.0007	0.0000	0.0001	0.0004	0.0001	0.0611
		$\theta_{11}$	$\theta_{12}$	$\theta_{13}$	$\theta_{14}$	$\theta_{15}$	$\theta_{16}$	$\theta_{17}$	$\theta_{18}$	$\theta_{19}$	$\theta_{20}$	
	Rodeo	0.0004	0.0001	0.0000	0.0000	0.0000	0.0007	0.0000	0.0000	0.0002	0.0000	
		$\theta_1$	$\theta_2$	$\theta_3$	$\theta_4$	$\theta_5$	$\theta_6$	$\theta_7$	$\theta_8$	$\theta_9$	$\theta_{10}$	IMSE
	Lasso	0.3488	0.2933	0.0104	0.0120	0.0121	0.0111	0.0106	0.0117	0.0106	0.0091	0.2078
		$\theta_{11}$	$\theta_{12}$	$\theta_{13}$	$\theta_{14}$	$\theta_{15}$	$\theta_{16}$	$\theta_{17}$	$\theta_{18}$	$\theta_{19}$	$\theta_{20}$	
Lasso	0.0094	0.0109	0.0113	0.0093	0.0117	0.0120	0.0123	0.0112	0.0107	0.0106		
$n = 200$		$\theta_1$	$\theta_2$	$\theta_3$	$\theta_4$	$\theta_5$	$\theta_6$	$\theta_7$	$\theta_8$	$\theta_9$	$\theta_{10}$	IMSE
	Rodeo	0.1958	0.1822	0.0018	0.0010	0.0024	0.0017	0.0008	0.0026	0.0022	0.0025	0.0517
		$\theta_{11}$	$\theta_{12}$	$\theta_{13}$	$\theta_{14}$	$\theta_{15}$	$\theta_{16}$	$\theta_{17}$	$\theta_{18}$	$\theta_{19}$	$\theta_{20}$	
	Rodeo	0.0014	0.0024	0.0024	0.0002	0.0003	0.0041	0.0005	0.0013	0.0036	0.0029	
		$\theta_1$	$\theta_2$	$\theta_3$	$\theta_4$	$\theta_5$	$\theta_6$	$\theta_7$	$\theta_8$	$\theta_9$	$\theta_{10}$	IMSE
	Lasso	0.4324	0.3810	0.0058	0.0069	0.0066	0.0064	0.0059	0.0059	0.0063	0.0055	0.1728
	$\theta_{11}$	$\theta_{12}$	$\theta_{13}$	$\theta_{14}$	$\theta_{15}$	$\theta_{16}$	$\theta_{17}$	$\theta_{18}$	$\theta_{19}$	$\theta_{20}$		
Lasso	0.0058	0.0057	0.0053	0.0054	0.0067	0.0069	0.0063	0.0072	0.0052	0.0049		

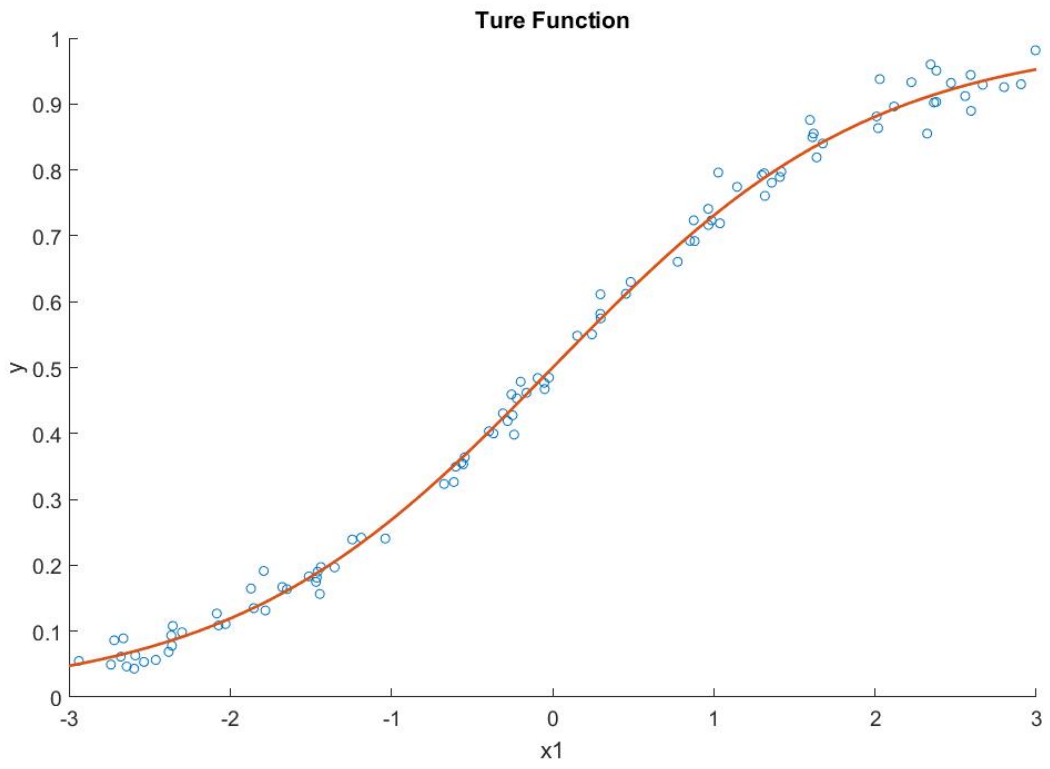


Figure 1:  $y$  with  $x_1$

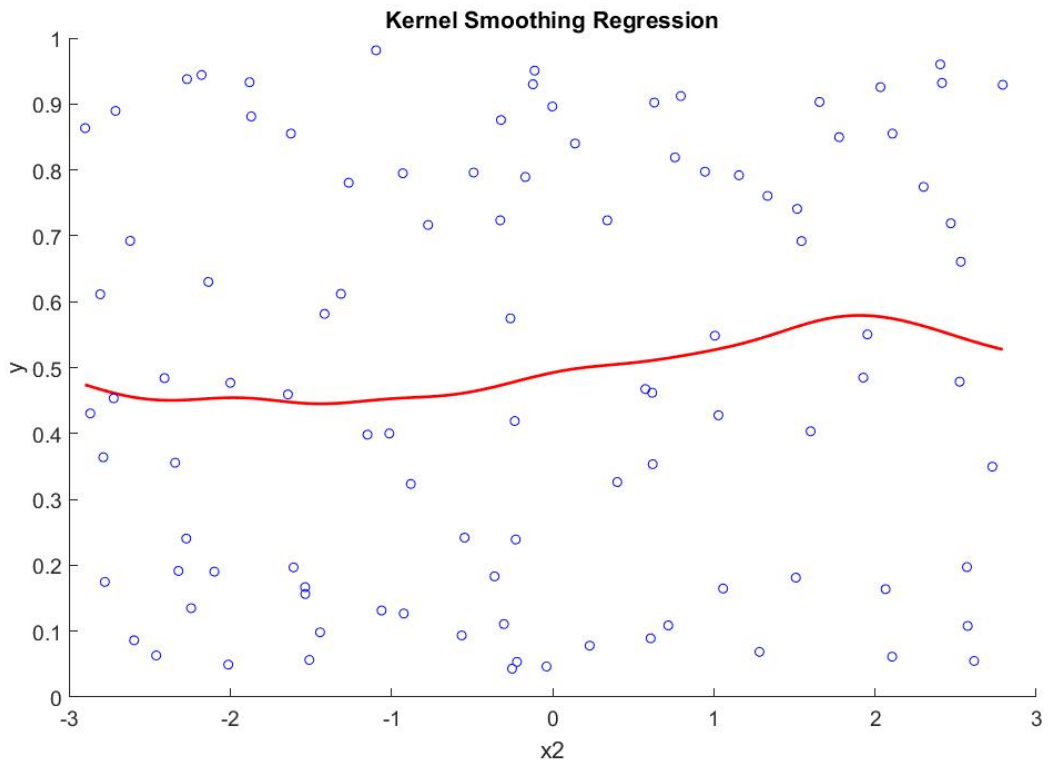
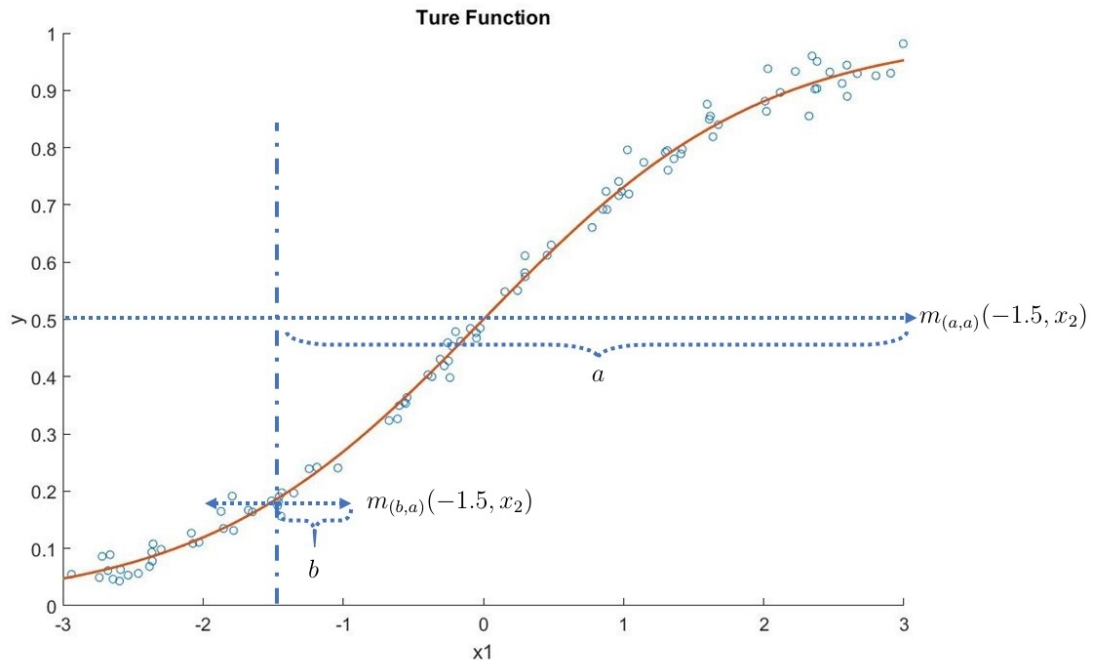
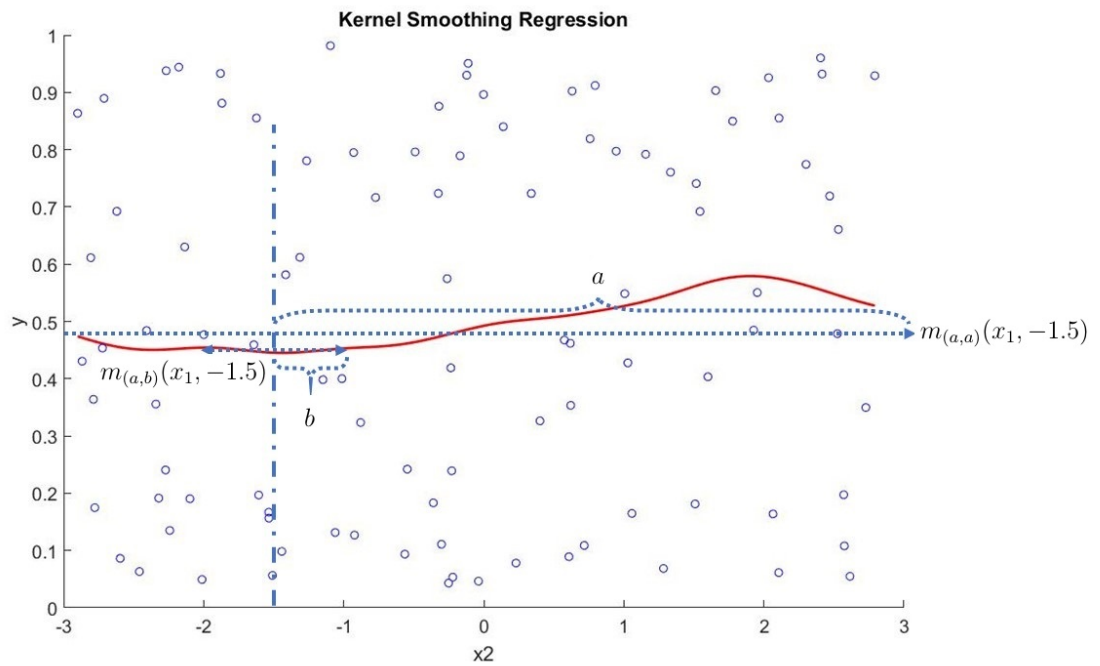


Figure 2:  $y$  with  $x_2$



$m_{(h_1, h_2)}(x_1, x_2)$  is the kernel estimator (1) with bandwidth  $h_1$  for  $x_1$  and bandwidth  $h_2$  for  $x_2$ . We start with a large bandwidth  $a$  for both  $x_1$  and  $x_2$ . Shrinking the bandwidth  $h_1$  from  $a$  to  $b$  leads to a dramatic change in the kernel estimator.

Figure 3: Shrinking bandwidth of  $x_1$  from  $a$  to  $b$



$m_{(h_1, h_2)}(x_1, x_2)$  is the kernel estimator (1) with bandwidth  $h_1$  for  $x_1$  and bandwidth  $h_2$  for  $x_2$ . We start with a large bandwidth  $a$  for both  $x_1$  and  $x_2$ . Shrinking the bandwidth  $h_2$  from  $a$  to  $b$  has no significant effect on the kernel estimator.

Figure 4: Shrinking bandwidth of  $x_2$  from  $a$  to  $b$

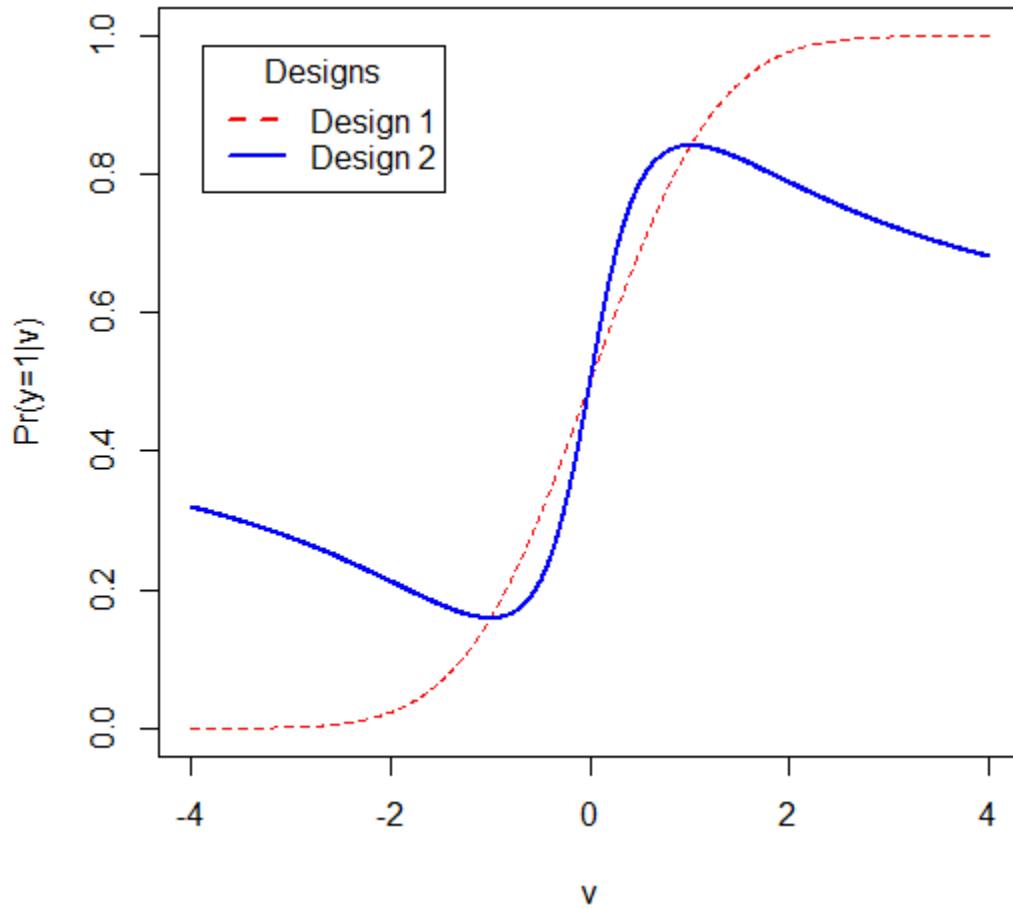


Figure 5: Designs