VARIABLE SELECTION IN SPARSE SEMIPARAMETRIC SINGLE INDEX MODELS

Jianghao Chu, Tae-Hwy Lee and Aman Ullah

University of California, Riverside, USA

ABSTRACT

In this chapter 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 (SIMs) 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 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 SIM, 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 (IMSE) than SIM-Lasso.

Keywords: Single index model (SIM); variable selection; Rodeo; SIM-Rodeo; Lasso; SIM-Lasso

JEL classifications: C25; C44; C53; C55

Topics in Identification, Limited Dependent Variables, Partial Observability, Experimentation, and Flexible Modeling
Advances in Econometrics, Volume 40B, 65–88

ISSN: 0731-9053/doi:10.1108/S0731-90532019000040B005

1. INTRODUCTION

In a series of chapters, (Poirier, 1980a, 1980b, 1994, 1996; Poirier and Ruud, 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, and partially linear models. This chapter proposes a new method of variable selection for sparse single index models (SIMs) 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. However, it suffers from computational complexity and the curse of dimensionality. Ichimura (1993) studied the semiparametric 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) studied a similar semiparametric SIM for binary outcomes and proposed to estimate the model by maximum likelihood (ML). 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 for example 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 chapter, 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 chapter 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 β_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 θ_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, Hastie, Johnstone, & Tibshirani, 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 chapter 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 semiparametric SIM of Ichimura (1993), introduces the SIM-Rodeo, and discusses the asymptotic properties of SIM-Rodeo in variable selection and estimation of the semiparametric SIM. 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 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, $X = (X_1', \ldots, X_n')'$ be a matrix of n observations, and $x \in \mathbb{R}^k$ be a local estimation point.

The Rodeo algorithm uses the kernel estimator:

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

where h is a vector of length k that is equal to the number of potential explanatory variable, h_j is the jth 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

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

where $\kappa(\cdot)$ is usually given as a one-variable density function. X_{ij} is the ith observation of the jth variable and x_j is the jth variable of a local estimation point x. In what follows, we keep the same notation except that in the SIM, 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

$$\iota = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \tag{3}$$

and

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

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

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

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

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

where σ^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 SIM in Section 4. Now we get all the ingredients of the Rodeo algorithm. The Rodeo algorithm is as follows.

Algorithm 1. Rodeo (Lafferty & 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:

- $h_j = h_0, j = 1, \dots, k$.
- $\mathcal{A} = \{1, \dots, k\}$.

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

- compute the estimated derivative and its conditional variance: Z_j and s_j using Eqs. (5) and (6);
- compute the threshold $\lambda_j = \hat{s}_j \sqrt{2 \log n}$; and
- 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 & Wasserman, 2008 Corollary 5.2).[1]

2.2. A Numerical Example

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

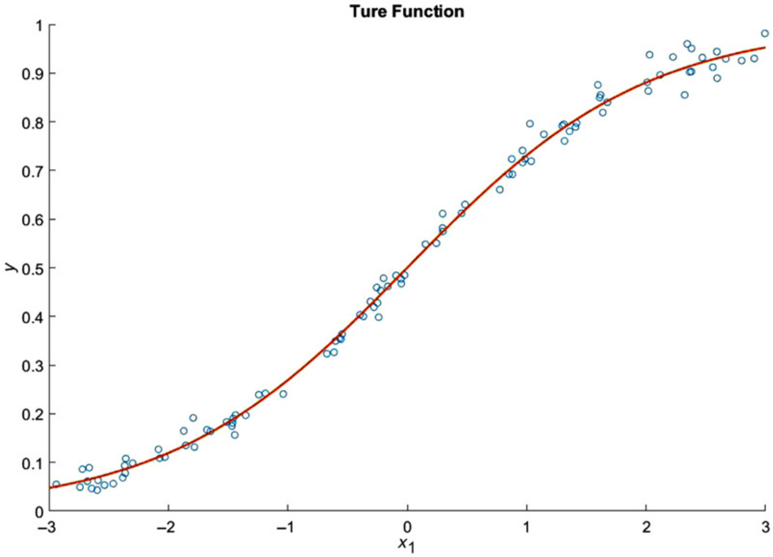

Fig. 1. y with x_1.

0 and standard deviation 0.02. The generated data of x_1 and y are shown in Fig. 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 Fig. 2.

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

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

Hence,

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

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

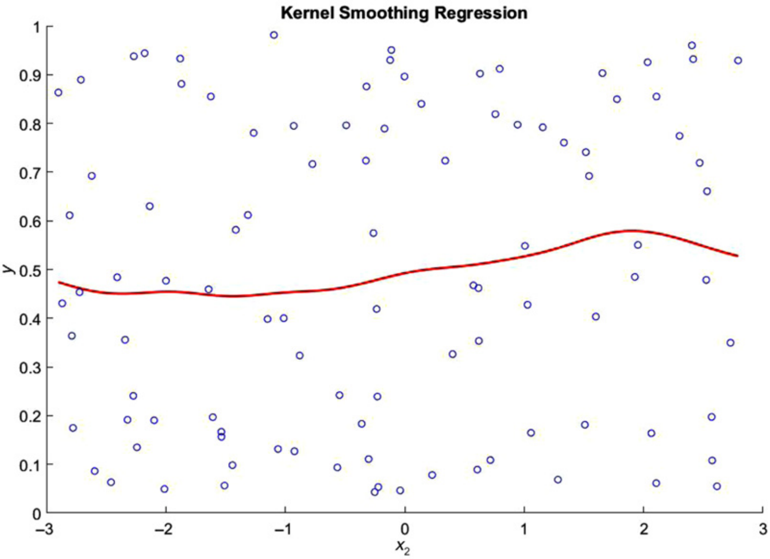

Fig. 2. y with x_2.

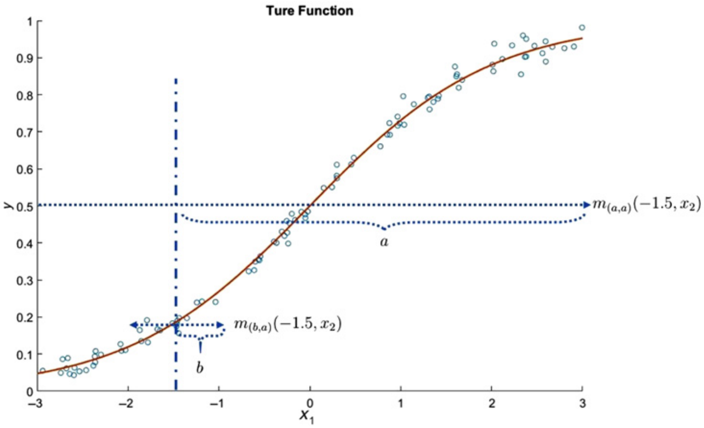

Fig. 3. Shrinking bandwidth of x_1 from a to b. *Notes:* $m_{(h1,\ h2)}$ (x_1, x_2) is the kernel estimator (1) with bandwidth h_1 for x_1 and bandwidth h_2 for x_2. We start with a large bandwidth a for both x_1 and x_2. Shrinking the bandwidth h_1 from a to b leads to a dramatic change in the kernel estimator.

estimate changes from the global mean $\hat{m}_h(x) = \overline{y}$ to a local mean $\hat{m}_h(x)$ as shown in Fig. 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

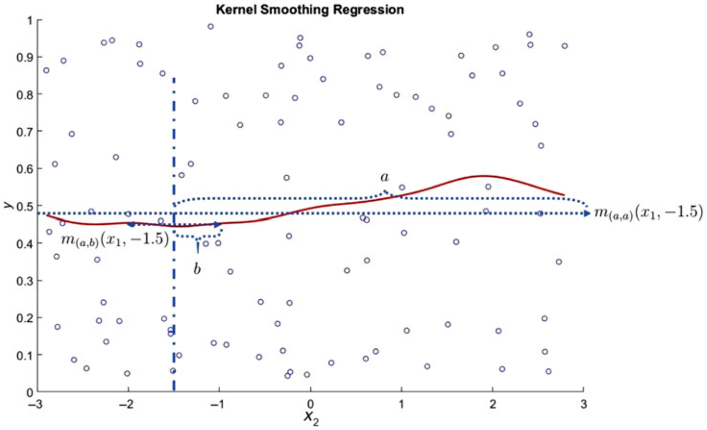

Fig. 4. Shrinking bandwidth of x_2 from a to b. *Notes:* $m_{(h1,\ h2)}$ (x_1, x_2) is the kernel estimator (1) with bandwidth h_1 for x_1 and bandwidth h_2 for x_2. We start with a large bandwidth a for both x_1 and x_2. Shrinking the bandwidth h_2 from a to b has no significant effect on the kernel estimator.

observations does not affect $\hat{m}_h(x)$. In fact, from Fig. 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 SIMs 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 SIM,

$$y = m(x'\beta) + u, \tag{10}$$

where $\beta = (\beta_1, \ldots, \beta_k)$ 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 follows:

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

where $K(\cdot)$ is a kernel function. The semiparametric kernel regression looks for the best β and h to minimize a weighted squared error loss. However, exact identification is not available. If one blows up β 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 β to be one (Ichimura, 1993).

As recent research pays more attention to high-dimensional data, most literature make the sparsity assumption that many, if not most, of the elements of β are zero. The previously mentioned identification method appears to be unsuitable unless we have specific information that the true value of the element of β that we set to be one 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 β 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:

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

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 θ_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 $x'\theta$. However, if x_j is not a relevant explanatory variable of y, then increasing the magnitude of θ_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 θ_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 SIM. 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 SIM. First we introduce some notation. Let

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

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

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

The derivative of the estimator Z_j with respect to θ_j is

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

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

For the ease of computation, let

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

Note that:

$$\frac{\partial W_x}{\partial \theta_j} = W_x L_j, \tag{18}$$

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

$$L_j = \begin{pmatrix} -\frac{1}{2}\frac{\partial\left(X_1'\theta - x'\theta\right)^2}{\partial\theta_j} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & -\frac{1}{2}\frac{\partial(X_n'\theta - x'\theta)^2}{\partial\theta_j} \end{pmatrix}$$

$$= \begin{pmatrix} -\left(X_1'\theta - x'\theta\right)\left(X_{1j} - x_j\right) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & -\left(X_n'\theta - x'\theta\right)\left(X_{nj} - x_j\right) \end{pmatrix},$$

where X_{1j} and X_{nj} are the jth elements of vectors X_1 and X_n. And x_j is the jth 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:

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

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.

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

$$E\left(Z_j | X\right) = E\left(G_j(x, \theta)(m(x'\beta) + u) | X\right) = G_j(x, \theta) m(x'\beta), \tag{21}$$

$$\mathrm{Var}\left(Z_j | X\right) = \mathrm{Var}\left(G_j(x, \theta)(m(x'\beta) + u) | X\right) = \boldsymbol{\sigma}' G_j(x, \theta)' G_j(x, \theta) \boldsymbol{\sigma}, \tag{22}$$

where $\boldsymbol{\sigma} = (\sigma(u_1), \ldots, \sigma(u_n))'$ is the vector of standard deviations of u. In the algorithm, it is necessary to insert an estimate of σ. 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 & 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:
    • $\theta_j = \begin{cases} \theta_0 & Z_j > 0 \\ -\theta_0 & \text{otherwise}, \end{cases} \quad j = 1, \ldots, k.$
    • $\mathcal{A} = \{1, \ldots, k\}$.
(3) While $\mathcal{A} \neq \varnothing$ is nonempty, do for each $j \in \mathcal{A}$:
    • compute $Z_j$ and $s_j = \sqrt{\mathrm{Var}(Z_j|X)}$ using Eqs. (19) and (22) respectively;
    • compute the threshold $\lambda_j = s_j\sqrt{2\log n}$;
    • if $|Z_j| > \lambda_j$, then set $\theta_j \leftarrow \frac{\theta_j}{\alpha}$; Otherwise, remove $j$ from $\mathcal{A}$ (i.e., $\mathcal{A} \leftarrow \mathcal{A} - \{j\}$); and
(4) Output $\hat{\theta} = (\theta_1, \ldots, \theta_k)$ and estimator $\hat{m}\left(x'\hat{\theta}\right)$.
```

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 SIM, θ is the ratio of β and the bandwidth. Since β could be either positive or negative, θ 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 θ_j. Our method is based on the observation that if θ_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 θ, i.e., all the positive θ will be given the same sign and all the negative θ will be given the same sign. A similar method is applied by Ichimura (1993) where the value positive one is given to the first β to ensure identification. Under the sparsity assumption, it is problematic to arbitrarily assign a magnitude greater than zero to any θ since the true value could be zero. However, it is safe to assume the sign of one of the θ to be positive or negative since positive zero and negative zero will not affect the relative scale of θ. Once again, due to the the identification issue with SIM, exact identification of θ is not available. However, signs of θ can be obtained relatively.

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

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

A2. $\lim\inf_{n\to\infty}\min_{1\le j\le r}\left|m_{jj}(\cdot)\right|>0$ 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.

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. *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 $\lim\inf_n\left|\frac{a_n}{b_n}\right|>0$ 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 *A1*, *A2*, and *A3* hold. In addition, suppose that:

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

and

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

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

$$\Pr\left(\theta_j = \theta_0 \, for\, all\, j>r\right)\to 1 \;\; as \;\; n\to\infty \tag{25}$$

and

$$\Pr\left(\theta_j > \theta_0 \, for\, all\, j\le r\right)\to 1 \;\; as \;\; n\to\infty. \tag{26}$$

Proof: see Appendix.

Theorem 1 shows that under the given assumptions and conditions, the coefficients θ for relevant variables will always be amplified while the coefficients θ 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 θ 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 et al., 2012) and ML (Klein & 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 DGP is given by:

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

where

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

The observed variable y_i is generated by:

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

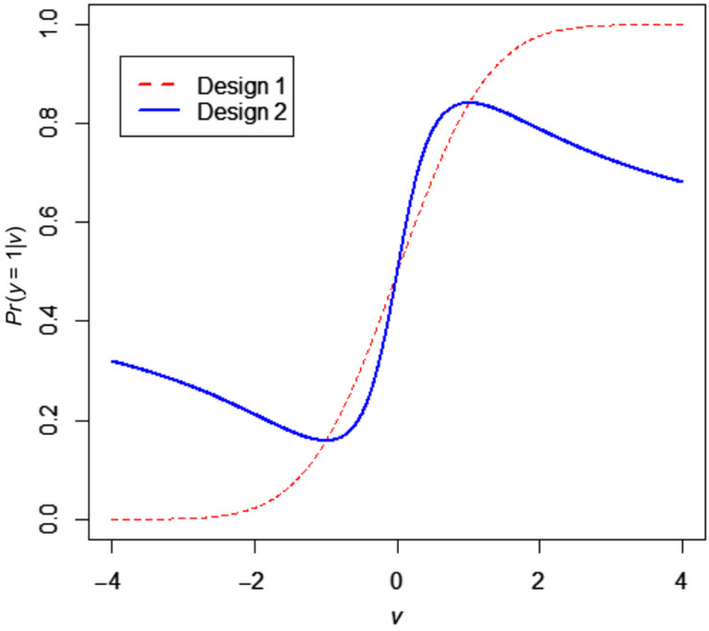

Fig. 5. Designs.

The x's are independently and identically distributed. x_1 is a chi-squared variate with three degrees of freedom truncated at 6 and standardized to have zero mean and unit variance; x_2 is a standard normal variate truncated at ± 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 Fig. 5 (Designs 1 and 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 is shown in Fig. 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 preform better under Design 2 since *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 SIMs for variable selection and estimation. Zeng et al. (2012) propose to solve the following minimization problem:

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

where λ is a hyper-parameter as in standard Lasso practices and

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

The authors provide their code to the supplemental materials of their chapter 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 θ indicate that the associated variables are relevant explanatory variables while small values of θ 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

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

		θ_1	θ_2	θ_3	θ_4	θ_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 ($k = 5$).

		θ_1	θ_2	θ_3	θ_4	θ_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 ($k = 20$).

$n = 100$		θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	θ_{10}	IMSE
	Rodeo	0.2349	0.2117	0.0032	0.0048	0.0027	0.0049	0.0013	0.0080	0.0009	0.0023	0.1487
		θ_{11}	θ_{12}	θ_{13}	θ_{14}	θ_{15}	θ_{16}	θ_{17}	θ_{18}	θ_{19}	θ_{20}	
	Rodeo	0.0093	0.0066	0.0037	0.0045	0.0026	0.0036	0.0016	0.0033	0.0030	0.0016	
		θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	θ_{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
		θ_{11}	θ_{12}	θ_{13}	θ_{14}	θ_{15}	θ_{16}	θ_{17}	θ_{18}	θ_{19}	θ_{20}	
	Lasso	0.0063	0.0041	0.0056	0.0047	0.0048	0.0057	0.0034	0.0047	0.0045	0.0048	
$n = 200$		θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	θ_{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
		θ_{11}	θ_{12}	θ_{13}	θ_{14}	θ_{15}	θ_{16}	θ_{17}	θ_{18}	θ_{19}	θ_{20}	
	Rodeo	0.0077	0.0123	0.0086	0.0082	0.0074	0.0147	0.0105	0.0156	0.0105	0.0081	
		θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	θ_{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
		θ_{11}	θ_{12}	θ_{13}	θ_{14}	θ_{15}	θ_{16}	θ_{17}	θ_{18}	θ_{19}	θ_{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 ($k = 20$).

$n = 100$		θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	θ_{10}	IMSE
	Rodeo	0.0692	0.0479	0.0007	0.0011	0.0000	0.0007	0.0000	0.0001	0.0004	0.0001	0.0611
		θ_{11}	θ_{12}	θ_{13}	θ_{14}	θ_{15}	θ_{16}	θ_{17}	θ_{18}	θ_{19}	θ_{20}	
	Rodeo	0.0004	0.0001	0.0000	0.0000	0.0000	0.0007	0.0000	0.0000	0.0002	0.0000	
		θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	θ_{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
		θ_{11}	θ_{12}	θ_{13}	θ_{14}	θ_{15}	θ_{16}	θ_{17}	θ_{18}	θ_{19}	θ_{20}	
	Lasso	0.0094	0.0109	0.0113	0.0093	0.0117	0.0120	0.0123	0.0112	0.0107	0.0106	
$n = 200$		θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	θ_{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
		θ_{11}	θ_{12}	θ_{13}	θ_{14}	θ_{15}	θ_{16}	θ_{17}	θ_{18}	θ_{19}	θ_{20}	
	Rodeo	0.0014	0.0024	0.0024	0.0002	0.0003	0.0041	0.0005	0.0013	0.0036	0.0029	
		θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	θ_{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
		θ_{11}	θ_{12}	θ_{13}	θ_{14}	θ_{15}	θ_{16}	θ_{17}	θ_{18}	θ_{19}	θ_{20}	
	Lasso	0.0058	0.0057	0.0053	0.0054	0.0067	0.0069	0.0063	0.0072	0.0052	0.0049	

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 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 ML suffers dramatically from the curse of dimensionality. ML 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 IMSE.

$$\text{IMSE} = \int (\hat{m}(x'\hat{\theta}) - m(x'\beta))^2 f(x)\mathrm{d}x \tag{32}$$

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

From the simulation results, we can see that under the sparsity condition, SIM-Rodeo and SIM-Lasso both outperform the traditional ML 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 state 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. ML 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 & Wasserman, 2008 Corollary 5.2).

In this chapter, we propose a new algorithm, based on the Rodeo, for variable selection and estimation for the sparse semiparametric linear SIMs 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 SIM 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 SIMs. 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 IMSEs than using SIM-Lasso.

NOTE

1. We can also make Rodeo for local constant kernel regression models as we will demonstrate in this chapter later.

ACKNOWLEDGMENTS

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.

REFERENCES

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, *32*(2), 407–499.

Huang, J., Horowitz, J. L., & Wei, F. (2010). Variable selection in nonparametric additive models. *The Annals of Statistics*, *38*, 2282–2313.

Ichimura, H. (1993). Semiparametric least squares and weighted SLS estimation of single index models. *Journal of Econometrics*, *58*(1–2), 71–120.

Klein, R. W., & Spady, R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica*, *61*(2), 387–421.

Koop, G., & Poirier, D. J. (1993). Bayesian analysis of logit models using natural conjugate priors. *Journal of Econometrics*, *56*(3), 323–340.

Koop, G., & Poirier, D. J. (2004). Bayesian variants of some classical semiparametric regression techniques. *Journal of Econometrics*, *123*(2), 259–282.

Lafferty, J., & Wasserman, L. (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), 141–142.

Poirier, D. J. (1980a). A Lagrange multiplier test for skewness in binary logit models. *Economics Letters*, *5*(2), 141−143.

Poirier, D. J. (1980b). Partial observability in bivariate probit models. *Journal of Econometrics*, *12*(2), 209−217.

Poirier, D. J. (1994). Jeffreys' prior for logit models. *Journal of Econometrics*, *63*(2), 327−339.

Poirier, D. J. (1996). A Bayesian analysis of nested logit models. *Journal of Econometrics*, *75*(1), 163−181.

Poirier, D. J., & Ruud, P. A. (1988). Probit with dependent observations. *The Review of Economic Studies*, *55*(4), 593−614.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *The Annals of Statistics*, *12*(4), 1215−1230.

Su, L., & Zhang, Y. (2014). Variable selection in nonparametric and semiparametric regression models. In J. Racine L. Su, & A. Ullah (Eds.), *The Oxford handbook of applied nonparametric and semiparametric econometrics and statistics* (Chapter 9, pp. 249−307). Oxford: 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, G. S. (1964). Smooth regression analysis. *Sankhya: The Indian Journal of Statistics, Series A*, *26*(4), 359−372.

Zeng, P., He, T., & Zhu, Y. (2012). A Lasso-type approach for estimation and variable selection in single index models. *Journal of Computational and Graphical Statistics*, *21*(1), 92−109.

AUTHOR BIOGRAPHY

Jianghao Chu is a PhD Candidate at University of California, Riverside. His research interests are nonparametric estimation and prediction with high-dimensional data, machine learning methods for variable selection and classification, finite sample theory for inference with heteroskedasticity, model averaging, financial econometrics, and business analytics. He is expected to receive PhD in Economics in June 2019.

Tae-Hwy Lee is Professor of Economics at University of California Riverside. His current research interests are high-dimensional models, quantiles, nonlinear models, financial econometrics, forecasting, shrinkage methods, and machine learning methods for econometric problems. He received the Econometric Theory Tjalling C. Koopmans Prize and is a Fellow of the *Advances in Econometrics* (AIE). He received PhD in Economics in 1990 from University of California, San Diego (UCSD).

Aman Ullah is Distinguished Professor of Economics at University of California Riverside. His current research interests are nonparametric econometrics, finite sample econometrics, interval data econometrics, information theoretic modeling, and machine learning procedures. He is a Fellow of *Journal of Econometrics* (JOE), *American Association for the Advancement of Science* (AAAS), *Advances in Econometrics* (AIE), *Royal Statistical Society* (RSS), *National Academy of Sciences* (*India*) (NASCI), and *Indian Econometric Society*. His PhD in Economics is from *The Delhi School of Economics* (DSE), Delhi.

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 $\lim\inf_n \left|\frac{a_n}{b_n}\right| > 0$ 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:

$$r = O(1),$$

$$k = O\left(\frac{\log n}{\log\log n}\right),$$

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

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

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

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

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

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

Furthermore, for any $\delta > 0$,

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

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* $\nu_j(\theta) = Var(Z_j|X_1, \dots, X_n)$. *Then*

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

for all $\varepsilon > 0$.

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

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

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

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

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

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:

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

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 ξ_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 $j \leq r$,

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

Hence,

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

and

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