# SPOT: Sparse Optimal Transformations for High Dimensional Variable Selection and Exploratory Regression Analysis

Qiming Huang
Department of Statistics,Purdue University
West Lafayette, Indiana, USA
hqm@purdue.edu

Michael Zhu
Department of Statistics,Purdue University
West Lafayette, Indiana
Center for Statistical Science, Department of Industrial Engineering, Tsinghua University
Beijing, China
yuzhu@purdue.edu

## ABSTRACT

We develop a novel method called SParse Optimal Transformations (SPOT) to simultaneously select important variables and explore relationships between the response and predictor variables in high dimensional nonparametric regression analysis. Not only are the optimal transformations identified by SPOT interpretable, they can also be used for response prediction. We further show that SPOT achieves consistency in both variable selection and parameter estimation. Numerical experiments and real data applications demonstrate that SPOT outperforms other existing methods and can serve as an effective tool in practic e.

## CCS CONCEPTS

•**Mathematics of computing →Exploratory data analysis;** *Spline models; Variable elimination; Regression analysis;* •**Computing methodologies →Feature selection;** *Regularization;*

## KEYWORDS

Monotone transformation, optimal transformation, regression analysis, spline, variable selection

## 1 INTRODUCTION

Regression analysis is arguably one of the most commonly used tools for data analysis in practice. Suppose $Y$ is the response variable of interest and $\mathbf{X} = (X_1, \ldots, X_p)$ is the vector of $p$ predictor variables. Based on a finite sample of $Y$ and $\mathbf{X}$, regression analysis is commonly used to discover how and to which degree the predictor variables $X_j$'s affect $Y$. In its generality, regressing $Y$ against $\mathbf{X}$ is to infer the dependence structure of $Y$ on $\mathbf{X}$. However, most existing regression methods are usually focused on certain characteristics of $Y$ such as the mean (i.e. $\mathbb{E}(Y|\mathbf{X})$), median (i.e. 50th percentile of $Y|\mathbf{X}$), or other quantiles of $Y|\mathbf{X}$. These methods are useful when the chosen characteristics are of primary interests, but may fail to capture the full dependence structure of $Y$ on $\mathbf{X}$. A number of attempts were made in the literature to directly estimate the conditional distribution $P(Y|\mathbf{X})$ [8, 26, 30]. The resulting approaches unfortunately suffer severely from the curse of dimensionality and are thus not practical [6].

Another approach to exploring the dependence of $Y$ on $\mathbf{X}$ is to first apply transformations to $Y$ and $\mathbf{X}$ and then perform regression analysis to the transformed variables. Intuitively, different transformations can lead to the discovery of different aspects of the dependence structure of $Y$ on $\mathbf{X}$. The well-known Box-Cox transformation and additive model can be considered two such approaches. Box and Cox (1964) proposed to apply power transformation to the response variable $Y$ in regression analysis, which aims to make the assumptions of linearity, normality, and homogeneity more appropriate. Different from the Box-Cox transformation, the additive model assumes that $Y$ depends on the transformations of individual predictor variables in an additive fashion, and fitting the additive model is to identify those transformations [10]. Despite the popularity of the Box-Cox transformation and additive model, their effectiveness can be compromised due to their susceptibility to model mis-specification. For example, both will fail in simple cases like $Y = \log(X_1 + X_2^2 + \epsilon)$.

Breiman and Friedman [3] proposed to apply general nonparametric transformations to both $Y$ and $\mathbf{X}$, and further to identify the *optimal transformations* that achieve the maximum correlation between them. The optimal transformations can be equivalently stated as the solution to the following least squares problem.

$$\min_{h \in L^2(P_Y), f_j \in L^2(P_{X_j})} \mathbb{E}\Big[\{h(Y) - \sum_{j=1}^{p} f_j(X_j)\}^2\Big], \tag{1}$$

subject to $\mathbb{E}[h(Y)] = \mathbb{E}[f_j(X_j)] = 0, \mathbb{E}[h^2(Y)] = 1$ and $\mathbb{E}[f_j^2(X_j)] < \infty$. Here, $P_Y$ and $P_{X_j}$ denote the marginal distributions of $Y$ and $X_j$, respectively, and $L^2(P)$ denotes the class of square integrable functions under the measure $P$. We denote the solution to (1) as $h^*$ and $f_j^*(j = 1, \ldots, p)$, which are referred to as the optimal transformations for $Y$ and $\mathbf{X}$, respectively.

Breiman and Friedman further developed the Alternating Conditional Expectation (ACE) algorithm to compute the optimal transformations. Although in general, the optimal transformations are not expected to fully capture the dependence structure of $Y$ on $\mathbf{X}$, they represent in a certain sense the most important features of the dependence structure. Notice that the transformed predictors are additive for the transformed response. This additive structure is important, because it ensures the interpretability of the captured

*KDD '17, August 13-17, 2017, Halifax, NS, Canada*
© 2017 ACM. 978-1-4503-4887-4/17/08…$15.00
DOI: 10.1145/3097983.3098091

dependence, that is, it shows how the predictors jointly affect the transformed response. In order to uncover the remaining dependence, intuitively, the idea of transformation can be iteratively applied. In this article, however, we will focus on the optimal transformations only.

The optimal transformations are subject to two limitations. Firstly, without any shape constraint, the transformation on the response $h^*(Y)$ may not be easily interpretable. In many real life applications such as modeling utility functions in economics, $h^*(Y)$ may not be meaningful if the order of the observations cannot be preserved after transformation. The problem becomes worse if the primary interest after transformation is to predict $Y$ instead of $h^*(Y)$. Secondly, despite the additive structure, the estimation of optimal transformations can suffer from the curse of dimensionality when the number of predictor variables $p$ is large. Even when the optimal transformations can be effectively estimated, the retention of a large number of spurious predictors can compromise their interpretability and prediction capacity.

To overcome those two limitations of the optimal transformations, in this article, we first propose to impose the monotonicity constraint on the transformation of $Y$. This constraint ensures that the transformed response variable is interpretable and invertible, and subsequently the prediction of $Y$ can be performed. Second, in order to eliminate the spurious predictor variables, we regularize the estimation procedure of the optimal transformations by using a special type of penalty called the Smooth Integration of Counting and Absolute deviation (SICA) penalty. We refer to the resulting optimal transformations as the SParse Optimal Transformations or SPOT in short.

Exiting methods that are closely related to SPOT include those developed for sparse additive models. Lin and Zhang [16] proposed the COSSO procedure, which assumes that each component function belongs to a Reproducing Kernel Hilbert space (RKHS). COSSO uses the sum of the RKHS norms of the component functions as a penalty for simultaneous variable selection and model fitting. Ravikumar et al. [24] introduced an approach called SPAM that penalizes the sum of $L_2$ norms of the component functions and is effectively a functional version of the group lasso [33]. Meier et al. [22], Huang et al. [13] and Balakrishnan et al. [1] also developed different methods for sparse high dimensional additive models by using different types of penalty functions.

Our proposed approach SPOT is distinct from the existing methods in two main aspects. Firstly, SPOT considers transformations on both $Y$ and $\mathbf{X}$ with former being subject to the monotonicity constraint. The monotone transformation can be crucial for the cases where the usual additive model for $Y$ does not hold. For such cases, the existing methods may fail to identify the dependence of $Y$ on $\mathbf{X}$, whereas SPOT can still be successful. The monotone transformation clearly includes the identity function as a special case, therefore, SPOT is expected to work well when the additive model for $Y$ indeed holds. Secondly, the SICA penalty used in SPOT enjoys many advantages over other types of penalty existing in the literature. The family of SICA functions proposed by Lv and Fan [19] forms a smooth homotopy between the $L_0$ and $L_1$ types of penalty, and include the $L_0$ and $L_1$ penalty as limiting cases. SICA can avoid the drawbacks of the $L_0$ and $L_1$ penalties while combining their strengths and lead to more stable estimates of model parameters and less stringent conditions under which variable selection consistency can be established (See Section 3.1 for more details).

Due to the use of monotone transformation on $Y$ and the SICA penalty, SPOT produces sparse optimal transformations that are interpretable and can be further used for prediction. We extended the ACE algorithm to compute the estimates of the sparse optimal transformations. Furthermore, we establish the consistency results for SPOT under various regularity conditions and assumptions. Our simulation study and real data application provide convincing evidence of SPOT's effectiveness in performing variable selection, exploring complex dependence structures, and performing prediction for the response. We believe SPOT can become an effective tool for high dimensional exploratory regression analysis in practice.

The rest of the article is organized as follows. Section 2 introduces basic notations used in the article. In Section 3, we formally define the sparse optimal transformation problem, propose the SICA penalty and the monotone transformation, and further develop the algorithm for estimating the sparse optimal transformations. The theoretical results on the estimation and selection consistency of sparse optimal transformation are given in Section 4, and the proofs are included in the supplementary materials. Experimental results based on simulation study and real data applications are reported in Section 5. Section 6 concludes the article.

## 2 NOTATIONS AND ASSUMPTIONS

Let $h(Y)$ and $f_j(X_j)$ denote the transformations of $Y$ and $X_j$, $j = 1, \ldots, p$. We assume the supports of $Y$ and $X_j$'s are compact, and they are further assumed to be [0,1] without loss of generality. Throughout the paper, $\{\mathbf{X}_i, Y_i\}_{i=1}^n$ is assumed to be an i.i.d. sample of $\mathbf{X}$ and $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})$.

For each $j = 1, \ldots, p$, let $\mathcal{H}_{X_j}$ denote the Hilbert space of measurable functions $f_j(X_j)$ with $\mathbb{E}[f_j(X_j)] = 0$ and the inner product $\langle f_j, f_j' \rangle = \mathbb{E}(f_j f_j')$, where $f_j'$ is an arbitrary function in $\mathcal{H}_{X_j}$. Note that the expectations are taken over the probability distribution of $X_j$ and $\mathbb{E}[f_j^2(X_j)] < \infty$. Let $\mathcal{H}_{\mathcal{X}}^+ = \mathcal{H}_{X_1} \oplus \mathcal{H}_{X_2} \oplus \cdots \oplus \mathcal{H}_{X_p}$ be the Hilbert space of functions of $\mathbf{X}$ that take an additive form: $\mathbf{f}(\mathbf{X}) = \sum_{j=1}^p f_j(X_j)$, with $f_j \in \mathcal{H}_{X_j}$. Let $L^2[0,1]$ be the Hilbert space of square integrable functions under the Lebesgue measure and $\{\psi_{jk} : k = 1, 2, \ldots\}$ denote a uniformly bounded, orthonormal basis of $L^2[0,1]$. To impose smoothness on each $f_j$, we only consider $f_j \in \mathcal{T}_j$, where $\mathcal{T}_j$ is the Sobolev ball of order two, that is, $\mathcal{T}_j = \{f_j \in \mathcal{H}_{X_j} : f_j = \sum_{k=1}^\infty \beta_{jk}\psi_{jk}, \sum_{k=1}^\infty \beta_{jk}^2 k^4 \leq C^2\}$ for some $0 < C < \infty$. To impose smoothness on $h$, we require that $h$ should be $r$ times continuously differentiable and its $r$-th derivative be Hölder continuous: $|h^{(r)}(y_1) - h^{(r)}(y_2)| \leq c|y_1 - y_2|^v$ for all $y_1$ and $y_2$, for some $0 < v \leq 1$ and $c > 0$. We use $\mathcal{M}$ to denote the set of functions satisfying this condition.

## 3 SPARSE OPTIMAL TRANSFORMATIONS

As discussed in the Introduction, in order to make the transformation of $Y$ interpretable and suitable for prediction, $h$ needs to be a monotone function. Without loss of generality, we require $h$ to be monotone increasing in this article. Then, the sparse optimal transformation (SPOT) problem can be defined as follows.

$$\begin{aligned} \min_{h\in\mathcal{M},\mathbf{f}:f_j\in\mathcal{T}_j} \quad & L(h,\mathbf{f})+\lambda\Omega(\mathbf{f}), \\ \text{s.t.} \quad & \mathbb{E}[h^2]=1, \quad h'\ge 0; \end{aligned} \tag{2}$$

where $L(h,\mathbf{f})=\frac{1}{2}\mathbb{E}\left[\left(h(Y)-\sum_{j=1}^{p} f_j(X_j)\right)^2\right]$, and $\Omega(\mathbf{f})=\sum_{j=1}^{p}\rho\left(\sqrt{\mathbb{E}[f_j^2(X_j)]}\right)$. Here, $\lambda$ is the regularization parameter and $\rho$ is a pre-specified penalty function.

## 3.1 SICA Penalty

As discussed in the Introduction, we choose to use the SICA penalty as $\rho$, which is denoted as $\rho := \rho_a(t)$ where

$$\rho_a(t)=\left(\frac{t}{a+t}\right)I(t\neq 0)+\left(\frac{a}{a+t}\right)t,\ t\in[0,\infty), \tag{3}$$

and

$$\rho_0(t)=\lim_{a\to 0+}\rho_a(t)=I(t\neq 0);$$
$$\rho_\infty(t)=\lim_{a\to\infty}\rho_a(t)=t.$$

It is clear that $\rho_0(\cdot)$ and $\rho_\infty(\cdot)$ correspond to the $L_0$ and $L_1$ penalty functions, respectively. As $a$ changes from zero to infinity, $\rho_a(\cdot)$ forms a smooth homotopy between the $L_0$ and $L_1$ penalty functions. Therefore, the SICA penalty with $0<a<\infty$ represents a compromise between the $L_0$ and $L_1$ penalty functions, while the $L_0$ and $L_1$ penalty functions can be considered the limiting cases.

Regularized regression methods using the $L_0$ penalty demonstrate different performances in parameter estimation, variable selection and computing than those using the $L_1$ penalty. The $L_0$ penalty is directly imposed on the number of variables, and thus is the original measure of model complexity. The $L_0$ penalty does not cause bias in estimation and can lead to consistency in variable selection under fairly general conditions (e.g. BIC of Schwarz [28]). It however suffers from the instability problem [2] and can become quickly infeasible in computing when the number of variables increases. On the other hand, as a convex relaxation of the $L_0$ penalty, the $L_1$ penalty enjoys the advantages of stability and simplicity in computing [31], but it can lead to noticeably large bias in estimation [7] and achieve variable selection consistency only under stringent conditions such as the irrepresentable condition for the lasso [34].

From (3), it can be seen that the SICA penalty in some sense can be considered a combination of the $L_0$ and $L_1$ penalty with the weights being dependent on $t$, and the tuning parameter $a$ determines where the SICA penalty stands between the $L_0$ and $L_1$ penalty. Lv and Fan [19] proposed a unified framework for regularizing least squares-based methods using the SICA penalty and investigated the properties of the resulting estimator under the linear model. It turns out that the SICA penalty possesses a number of advantages. Firstly, not like the $L_0$ penalty, the SICA penalty is continuous in $t$, therefore, stable and efficient algorithms can be developed to solve the SICA-regularized least squares problem. Secondly, the condition under which the SICA penalty can lead to variable selection consistency is much less restrictive than the irrepresentable condition under the $L_1$ or lasso penalty. The fundamental reason for the second advantage is given as follows. When the tuning parameter $a$ approaches to zero, the SICA penalty approaches the $L_0$ penalty, and helps the regularized method explore a broader solution or model space. (We note that $a$ cannot get too close to zero in practice; otherwise, the computation will start to become unstable.) In summary, the SICA penalty manages to combine the strengths of the $L_0$ and $L_1$ penalty while avoiding their limitations. We believe that the SICA penalty is not simply a variant of the popularly used $L_1$ penalty, and it is in fact a significant improvement and should be widely adopted in practice. Other good properties related the SICA penalty can be found in Nikolova [23], Lin and Lv [15] and Lv and Liu [20].

When the tuning parameter $a$ is sufficiently large, the behavior of the SICA penalty is very similar to the $L_1$ penalty, and in such a case, we propose to directly use the $L_1$ penalty. Therefore, we include both the SICA penalty and the $L_1$ penalty when we implement SPOT in a computing package. When the SICA penalty is used, we refer to our procedure as SPOT-SICA, and when the $L_1$ penalty is used, we refer to our procedure as SPOT-LASSO. We remark that SPOT-LASSO can be considered a special case of SPOT-SICA with $a=\infty$, and SPAM a special case of SPOT-LASSO with $h(Y)=Y$.

## 3.2 Monotone Transformation on Response

Let $\mathcal{S}_{q\ell_n}$ be the space of polynomial splines of degree $q\ge 1$ with equally-spaced knots. Let $\{B_m, m=1,\dots,\ell_n\}$ denote a normalized B-spline basis with $||B_m||_{\text{sup}}\le 1$, where $||\cdot||_{\text{sup}}$ is the sup-norm. Then, $\widetilde{h}(Y)=\sum_{m=1}^{\ell_n}\alpha_m B_m(Y)$ for any $\widetilde{h}(Y)\in\mathcal{S}_{q\ell_n}$. It is shown in De Boor [5] that for any $h\in\mathcal{M}$ defined in Section 2, there exists a function $\tilde{h}\in S_{q\ell_n}$ such that $||\tilde{h}-h||_{\text{sup}}=O(\ell_n^{-(r+v)})$, with $q\ge r+v$. The constraint that the transformation $h$ is monotone increasing in the SPOT problem (2) can be readily accommodated in B-spline approximation. According to Schumaker [27], a sufficient condition for a polynomial spline $\tilde{h}(Y)$ to be monotone increasing is that its coefficients satisfy the linear constraints $\alpha_m\ge\alpha_{m-1}$ for $m=2,\dots,\ell_n$. When using the centered B-spline basis, the linear constraints become $\alpha_1\ge 0, \alpha_m\ge\alpha_{m-1}$ for $m=2,\dots,\ell_n$. Let $\boldsymbol{\alpha}=(\alpha_1,\dots,\alpha_{\ell_n})^T$. The linear constraints can be further written as $D^T\boldsymbol{\alpha}\ge 0$, where $D$ is the $\ell_n\times\ell_n$ matrix defined as

$$D=I_{\ell_n}-\begin{bmatrix}\mathbf{0}_{\ell_n-1} & I_{\ell_n-1}\\ 0 & \mathbf{0}_{\ell_n-1}^T\end{bmatrix}.$$

Here, $I_k$ is the $k\times k$ identity matrix, and $\mathbf{0}_{\ell_n-1}$ is the $\ell_n-1$ dimensional vector of 0's. Denote $\mathbf{B}$ as the $n\times\ell_n$ matrix where $\mathbf{B}(i,k)=B_k(Y_i)$. Then, in terms of the sample, we have $\tilde{h}(\mathrm{Y})=\mathbf{B}\boldsymbol{\alpha}$ where $\tilde{h}(\mathrm{Y})=(\tilde{h}(Y_1),\dots,\tilde{h}(Y_n))^T$.

## 3.3 SPOT Algorithm

Recall that $\{\psi_{jk}: k=1,2,\dots\}$ is an orthonormal basis and $f_j=\sum_{k=1}^{\infty}\beta_{jk}\psi_{jk}$. We use $\widetilde{f}_j=\sum_{k=1}^{d_n}\beta_{jk}\psi_{jk}$ to approximate $f_j$, where $d_n$ is a truncation parameter. Thus, $\widetilde{f}_j$ is a smoothed approximation to $f_j$. It is well-known that for the second order Sobolev ball $\mathcal{T}_j$, we have $||f_j-\widetilde{f}_j||_2^2=O(1/d_n^4)$ for $f_j\in\mathcal{T}_j$ [24]. Let $\Psi_j$ denote the $n\times d_n$ matrix where $\Psi_j(i,k)=\psi_{jk}(X_{ij})$ and $\boldsymbol{\beta}_j:=(\beta_{j1},\dots,\beta_{jd_n})^T$. We have $\widetilde{f}_j(X_j)=\Psi_j\boldsymbol{\beta}_j$ where $\widetilde{f}_j(X_j)=(\widetilde{f}_j(X_{1j}),\dots,\widetilde{f}_j(X_{nj}))^T$. Recall that $\tilde{h}(\mathrm{Y})=\mathbf{B}\boldsymbol{\alpha}$ and $D$ defined in Section 3.2. The sample version of the SPOT problem (2) with the SICA penalty can be written as follows.

$$\begin{aligned} \min_{\substack{\boldsymbol{\alpha}\in R^{\ell_n}\\ \boldsymbol{\beta}_j\in R^{d_n}}} \quad & \frac{1}{2n}\left\|\mathbf{B}\boldsymbol{\alpha}-\sum_{j=1}^{p}\Psi_j\boldsymbol{\beta}_j\right\|_2^2+\lambda_n\sum_{j=1}^{p}\rho_a\left(\frac{\left\|\Psi_j\boldsymbol{\beta}_j\right\|_2}{\sqrt{n}}\right) \\ \text{s.t.} \quad & \frac{1}{n}\boldsymbol{\alpha}^T\mathbf{B}^T\mathbf{B}\boldsymbol{\alpha}=1; \quad D^T\boldsymbol{\alpha}\geq 0. \end{aligned} \tag{4}$$

We develop a coordinate descent procedure to solve (4). The estimation procedure is summarized in Algorithm 1. To facilitate the calculation of the SICA penalty, we apply the local linear approximation (LLA) method proposed in Zou and Li [35] to $\rho_a(t)$, which is $\rho_a(t)\approx\rho_a'(t_0)t+\rho_a(t_0)-\rho_a'(t_0)t_0$ and $\rho_a'(t_0)=a(a+1)/(a+t_0)^2$ in a neighborhood of $t_0$. We explain the two main components of Algorithm 1 below.

Suppose the current estimates of transformations are given as $\hat{h}^{(0)},\hat{f}_1^{(0)},\ldots,\hat{f}_p^{(0)}$, and we want to update $f_j$ next. Denote $\hat{R}_j^{(0)}=\hat{h}^{(0)}-\sum_{k\neq j}\hat{f}_k^{(0)}$. Applying the LLA method, the objective function in (4) can be simplified to

$$\frac{1}{2n}\left\|\hat{R}_j^{(0)}-\Psi_j\boldsymbol{\beta}_j\right\|_2^2+\lambda_n w_j\frac{1}{\sqrt{n}}\left\|\Psi_j\boldsymbol{\beta}_j\right\|_2, \tag{5}$$

where $w_j=a(a+1)/(a+t_j)^2$ and $t_j=\frac{1}{\sqrt{n}}||\hat{f}_j^{(0)}||_2$. Notice that $w_j$ only depends on the current estimate $\hat{f}_j^{(0)}$. Therefore, updating $\boldsymbol{\beta}_j$ in (5) is essentially equivalent to solving a weighted group lasso problem [12] with respect to one group, and the update of $\boldsymbol{\beta}_j$ admits an explicit expression as follows.

$$\hat{\boldsymbol{\beta}}_j=\left[1-\frac{\lambda_n w_j\sqrt{n}}{||\Psi_j(\Psi_j^T\Psi_j)^{-1}\Psi_j^T\hat{R}_j^{(0)}||_2}\right]_+(\Psi_j^T\Psi_j)^{-1}\Psi_j^T\hat{R}_j^{(0)}$$

where $[\cdot]_+$ denotes the positive part. Therefore, $f_j$ can be updated as

$$\hat{f}_j=\Psi_j\hat{\boldsymbol{\beta}}_j=\left[1-\frac{\lambda_n w_j\sqrt{n}}{||\hat{P}_j^{(0)}||_2}\right]_+\hat{P}_j^{(0)} \tag{6}$$

where $\hat{P}_j^{(0)}=\Psi_j(\Psi_j^T\Psi_j)^{-1}\Psi_j^T\hat{R}_j^{(0)}$.

Note that the objective function for SPOT-LASSO is equivalent to (5) with $w_j=1$. In such a case, we do not need to use the LLA method, and Algorithm 1 can be used directly to calculate SPOT-LASSO by specifying $w_j=1$.

After updating $f_j$ for $j=1,\ldots,p$, we fix $\hat{\mathbf{f}}=\hat{f}_1+\ldots+\hat{f}_p$ and further update $h$ (i.e., update $\boldsymbol{\alpha}$). Problem (4) becomes

$$\begin{aligned} \min_{\boldsymbol{\alpha}\in R^{\ell_n}} \quad & \frac{1}{2n}\left\|\mathbf{B}\boldsymbol{\alpha}-\hat{\mathbf{f}}\right\|_2^2 \\ \text{s.t.} \quad & \frac{1}{n}\boldsymbol{\alpha}^T\mathbf{B}^T\mathbf{B}\boldsymbol{\alpha}=1; \quad D^T\boldsymbol{\alpha}\geq 0; \end{aligned} \tag{7}$$

which is equivalent to a typical quadratic programing problem (with re-normalizing to make the equality constraint hold). Standard numeric packages, such as the R package "quadprog", can be used to solve problem (7), and we obtain $\widehat{\boldsymbol{\alpha}}$ as the estimate of $\boldsymbol{\alpha}$.

**Algorithm 1** SPOT-SICA Coordinate Descent Algorithm

```
Input: Data $\{\mathbf{X}_i, Y_i\}_{i=1}^n$, tuning parameters $\lambda$, $a$
Initialize $\hat{h} = Y/||Y||_2$, $\hat{f}_j = \mathbf{0}_n$ for $j = 1, \ldots, p$
Iterate (I) - (II) until convergence:
(I) Update $\hat{f}_j$, for each $j = 1, \ldots, p$;
    Compute the residual: $\widehat{R}_j = \hat{h} - \sum_{k\neq j}\hat{f}_k$
    Calculate $\widehat{P}_j = \Psi_j(\Psi_j^T\Psi_j)^{-1}\Psi_j^T\widehat{R}_j$
    Compute weight $w_j$:
        $w_j = a(a+1)/(a + ||\hat{f}_j||_2/\sqrt{n})^2$ for finite $a$
        $w_j = 1$ for $a = \infty$ ($L_1$ penalty)
    Soft thresholding, $\hat{f}_j = \left[1 - \lambda w_j\sqrt{n}/||\widehat{P}_j||_2\right]_+ \widehat{P}_j$
    Centering, $\hat{f}_j = \hat{f}_j - \text{mean}\ (\hat{f}_j)$
(II) Update $\hat{h}$;
    Solve $\widehat{\boldsymbol{\alpha}}$ in problem (7) by Quadratic Programming
    Obtain $\hat{h} = \mathbf{B}\widehat{\boldsymbol{\alpha}}$
Output: Fitted functions $\hat{h}$ and $\hat{f}_j, j = 1, \ldots, p$
```

## 4 THEORETICAL PROPERTIES

In this section, we discuss the theoretical properties of SPOT-SICA in variable selection and parameter estimation. In particular, we establish the consistency of SPOT-SICA under the transformation model and a given estimate of the response transformation. We assume that observations of $(Y,\mathbf{X})$ come from the following transformation model $h^*(Y)=\sum_{j=1}^p f_j^*(X_j)+\epsilon$, where $h^*$ and $f_j^*$ are the optimal transformations and $\epsilon$ is random error. Rewriting the transformation model in terms of an orthonormal basis $\{\psi_{jk}\}$, we have that

$$h^*(Y)=\sum_{j=1}^{p}\sum_{k=1}^{\infty}\beta_{jk}^*\psi_{jk}(X_j)+\epsilon. \tag{8}$$

The transformation model is a general model that encompasses a broad class of models in both statistics and econometrics [4, 14, 17]. We use the transformation model to facilitate our theoretical discussion, and show that under the transformation model, SPOT-SICA can recover the true model with probability approaching one asymptotically. In the case that the true distribution of $\mathbf{X}$ and $Y$ is more complex, the transformation model can also be used as an approximate model due to its flexibility, and our numerical results show that SPOT-SICA can still be used as an effective tool for variable selection (see Example 4 in the supplementary materials).

Let $S$ denote the set of true variables $S=\{j,f_j^*\neq 0\}$, and $s_n$ the cardinality of $S$, and $S^c$ its complement. We show that SPOT-SICA can correctly identify $S$ and consistently estimate $\boldsymbol{\beta}_j^*$ in (8) for $j\in S$.

Recall that $\Psi_j$ is the $n\times d_n$ matrix obtained from the sample, we use $\Psi_S$ to denote the $n\times s_n d_n$ matrix formed by stacking the matrices $\Psi_j, j\in S$ one after another. We state the assumptions under which the main results hold.

Assumption 1:
(A) Let $\tau_n=\min_{j\in S}||\Psi_j\beta_j^*||_2/\sqrt{n}$, where $||\cdot||_2$ is $\ell_2$-norm of a vector. It holds that $n^\alpha\tau_n\to\infty$ with $\alpha\in(0,1/2)$.
(B) It holds that $\rho'(\tau_n/2)=o(n^{-\alpha}d_n^{-1}s_n^{-1/2})$ and $\sup_{t\geq\tau_n/2}\rho''(t)=o(1)$.

(C) There exists a positive constant $c_0$ such that

$$c_0 \leq \min_{j \in S} \Lambda_{\min}\left(\frac{1}{n}\Psi_j^T \Psi_j\right) \leq \Lambda_{\max}\left(\frac{1}{n}\Psi_S^T \Psi_S\right) \leq c_0^{-1}$$

where $\Lambda_{\min}$ and $\Lambda_{\max}$ are the smallest and largest eigenvalues of a matrix, respectively.
(D) It holds that

$$\max_{j \in S^c} \left\| \Psi_j (\Psi_j^T \Psi_j)^{-1} \Psi_j^T \Psi_S \ (\Psi_S^T \Psi_S)^{-1} \right\|_{\infty, 2} \leq \frac{\sqrt{c_0}}{2\sqrt{n}} \frac{\rho'(0+)}{\rho'(\tau_n/2)} \qquad (9)$$

where for a matrix $A$, $||A||_{\infty,2} = \sup_{||x||_\infty = 1} ||Ax||_2$ with $x$ being a vector.
(E) The errors $\epsilon_i, i = 1, \ldots, n$, are independent and identically distributed as $N(0, \sigma^2)$.

Condition (C) and (D) in Assumption 1 are similar to the Irrepresentable Condition in Zhao and Yu [34] for $L_1$ penalty, which ensures selection consistency of Lasso. The conditions are adopted from Fan et al. [9], and similar conditions are required for SPAM (Theorem 2 in Ravikumar et al. [24]). The conditions here are fairly general compared to conditions in other procedures. In particular, if $\Psi_S$ is orthogonal, Condition (C) is satisfied with $c_0 = 1$. Condition (D) is automatically satisfied if $a \longrightarrow 0$ (e.g., $a = o(\tau_n)$).

Equation (9) reflects the restriction on the design matrix for SPOT-SICA to be consistent in variable selection. For fixed sample size $n$, the quantify $\rho'(0+)/\rho'(\tau_n/2)$ plays a critical role in Assumption (D). For the SICA penalty, we have $\rho'(0+)/\rho'(\tau_n/2) = (1 + \tau_n/(2a))^2$. The smaller $a$ is, the less restrictive Assumption (D) becomes. As $a \longrightarrow 0$, $\rho'(0+)/\rho'(\tau_n/2)$ approaches $\infty$. Therefore, for any given fixed design matrix, Assumption (D) will eventually be satisfied when $a$ is sufficiently small, and SPOT-SICA will have a high probability of selecting the true model. On the other hand, as $a \longrightarrow \infty$, $\rho_a$ approaches the $L_1$ penalty $\rho_\infty$, and $\rho'(0+)/\rho'(\tau_n/2)$ approaches 1; In other words, the right hand side of (9) becomes smaller and smaller, and Assumption (D) becomes more and more restrictive. When $a$ is too large, Assumption (D) may fail to hold, and SPOT-SICA or its limiting version SPOT-LASSO may fail to select the true variables. In theory, it appears that a small $a$ should always be preferred. Unfortunately, this is not true, because as we remarked previously, when $a$ is too small, the SICA penalty in general will incur computational instability and produce inferior results.

Assumption 2:
There exits an $L_\infty$-consistent estimate $\hat{h}^*$ of $h^*$ and $||\hat{h}^* - h^*||_{L_\infty} = \sup_{y \in [0,1]} |\hat{h}^*(y) - h^*(y)| = O_p(v_n)$ for some sequence $v_n = o(\lambda_n)$, where $\lambda_n$ is the regularization parameter in (4).

Assumption 2 assumes there exists a good estimate of the transformation $h^*$. This assumption is valid since there are several procedures proposed for obtaining such an estimate in the literature [4, 11, 17]. In particular, Chiappori et al. [4] showed that under certain conditions, $h^*$ can be estimated at the parametric rate. Due to space limitation as well as for ease of presentation, we do not present the details along that direction but instead state it as an assumption.

THEOREM 4.1. *Assume that $d_n + \log p = O(n\lambda_n^2)$, $\lambda_n n^\alpha d_n \sqrt{s_n} \rightarrow 0$, $\log(pd_n) = o(n^{1-2\alpha} s_n^{-1} d_n^{-2})$, and $s_n d_n^{-2} + v_n = o(\lambda_n)$. Then under Assumptions 1 and 2, with probability approaching one as n goes to infinity, there exits a local minimizer $\hat{\boldsymbol{\beta}}$ of (4) such that:*
*(1) $\hat{\boldsymbol{\beta}}_{S^c} = 0$;*
*(2) $||\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^*||_\infty \leq c_0^{1/2} n^{-\alpha} d_n^{-1/2}$;*
*where $|| \cdot ||_\infty$ stands for the infinity norm of a vector.*

Theorem 4.1 establishes the weak oracle property for SPOT-SICA in that SPOT-SICA not only identifies the true model, but also estimates the true coefficients consistently. The sketch of the proof is given in the supplementary materials and the main idea of the proof follows Fan et al. [9]. Note that Theorem 4.1 states a result of a local solution, it has been proved in [18] that any local minimizer will fall within statistical precision of the true parameter vector under appropriate conditions on the penalty function.

# 5 EXPERIMENTS

In this section, we compare the performances of SPOT-SICA, SPOT-LASSO and SPAM in variable selection and prediction through both synthetic and real life examples. For SPAM, we use its implementation in the R package "SAM". Similar to the implementation of SPAM, we use B-spline bases for function approximation in SPOT-SICA and SPOT-LASSO.

## 5.1 Effectiveness on Synthetic Data

We test the methods using data sampled from two types of models, the additive model and the transformation model. In the first example, we consider an additive model where SPAM is expected to work well. In the second example, a typical transformation model is considered. For each training data set, we also generate a validation data set and a test data set. Validation datasets are used to choose the tuning parameters $\lambda$ and $a$, and test datasets are used to measure the prediction accuracy of the estimated models in terms of mean squared error (MSE). The goal of using separate validation datasets and test datasets is to facilitate fair comparisons of different methods. We replicate each simulation 100 times, and report the averages and standard deviations (in parentheses) of precisions, recalls, sizes of the selected variables, $F_1$ scores, as well as MSEs of the estimated models on the test datasets. More simulation examples can be found in the supplementary materials.

**Example 1.** $Y = \sum_{j=1}^p f_j(X_j) + \epsilon$ where $\epsilon \sim N(0, 8/9)$; The functions are given by $f_1(x) = -2\sin(2x)$, $f_2(x) = x^2$, $f_3(x) = \frac{2\sin(x)}{2-\sin(x)}$, $f_4(x) = \exp(-x)$, $f_5(x) = x^3 + 1.5(x-1)^2$, $f_6(x) = x$, $f_7(x) = 3\sin(\exp(-0.5x))$, $f_8(x) = -5\Phi(x, 0.5, 0.8^2)$, and $f_j = 0$ for $j \geq 9$. Here, $\Phi(\cdot, \mu, \sigma^2)$ is the Gaussian cumulative distribution function with mean $\mu$ and standard deviation $\sigma$.

We generate covariates with a compound symmetry covariate structure as follows. Each covariate $X_j = (W_j + tU/3)/(1 + t/3), j = 1, \ldots, p$, where $W_1, \ldots, W_p$ and $U$ are from $Unif(-2.5, 2.5)$. As $t$ increases, the correlation between any two predictors will increase, which renders the variable selection problem more difficult in general. The sample size is $n = 200$ for train/validation/test dataset, and we consider the dimension of covariates $p = 50, 200$ and $1000$.

**Table 1: Comparison of different methods on simulated data from Examples 1 and 2.**

| Example | $p$ | $t$ | Method | Precision | Recall | Size | $F_1$ score | MSE |
|---|---|---|---|---|---|---|---|---|
| 1 | 1000 | 0 | SPAM | 0.14 (0.04) | 1.00 (0.00) | 64.08 (18.85) | 0.24 (0.06) | 2.11 (0.28) |
| 1 | 1000 | 0 | SPOT-LASSO | 0.24 (0.12) | 1.00 (0.00) | 56.02 (52.07) | 0.37 (0.16) | 1.94 (0.35) |
| 1 | 1000 | 0 | SPOT-SICA | 0.86 (0.21) | 1.00 (0.00) | 10.39 (5.24) | 0.91 (0.15) | 1.35 (0.17) |
| 1 | 1000 | 3 | SPAM | 0.11 (0.03) | 1.00 (0.00) | 74.41 (19.69) | 0.20 (0.05) | 2.18 (0.36) |
| 1 | 1000 | 3 | SPOT-LASSO | 0.05 (0.02) | 1.00 (0.00) | 168.49 (43.81) | 0.10 (0.04) | 2.69 (0.42) |
| 1 | 1000 | 3 | SPOT-SICA | 0.86 (0.17) | 1.00 (0.00) | 9.86 (2.98) | 0.91 (0.11) | 1.44 (0.24) |
| 1 | 1000 | 6 | SPAM | 0.13 (0.04) | 0.90 (0.12) | 64.79 (24.96) | 0.22 (0.06) | 2.28 (0.33) |
| 1 | 1000 | 6 | SPOT-LASSO | 0.15 (0.22) | 0.85 (0.20) | 127.24 (75.92) | 0.18 (0.16) | 2.68 (0.35) |
| 1 | 1000 | 6 | SPOT-SICA | 0.74 (0.21) | 0.79 (0.13) | 9.68 (4.66) | 0.74 (0.11) | 2.18 (2.51) |
| 2 | 1000 | 0 | SPAM | 0.14 (0.05) | 0.98 (0.06) | 41.80 (17.70) | 0.24 (0.08) | 3.91 (0.56) |
| 2 | 1000 | 0 | SPOT-LASSO | 0.23 (0.15) | 0.98 (0.06) | 45.16 (56.81) | 0.34 (0.17) | 2.39 (0.44) |
| 2 | 1000 | 0 | SPOT-SICA | 0.83 (0.24) | 1.00 (0.03) | 7.81 (9.13) | 0.88 (0.19) | 2.02 (0.33) |
| 2 | 1000 | 3 | SPAM | 0.11 (0.05) | 0.87 (0.15) | 44.06 (15.98) | 0.19 (0.07) | 2.91 (0.43) |
| 2 | 1000 | 3 | SPOT-LASSO | 0.18 (0.18) | 0.85 (0.16) | 61.00 (55.92) | 0.25 (0.18) | 2.30 (0.39) |
| 2 | 1000 | 3 | SPOT-SICA | 0.75 (0.31) | 0.89 (0.15) | 11.57 (18.51) | 0.76 (0.25) | 1.99 (0.34) |
| 2 | 1000 | 6 | SPAM | 0.10 (0.05) | 0.72 (0.18) | 40.58 (15.73) | 0.17 (0.07) | 2.94 (0.40) |
| 2 | 1000 | 6 | SPOT-LASSO | 0.12 (0.10) | 0.71 (0.26) | 59.24 (57.58) | 0.18 (0.11) | 2.36 (0.35) |
| 2 | 1000 | 6 | SPOT-SICA | 0.60 (0.36) | 0.73 (0.20) | 16.97 (23.42) | 0.55 (0.26) | 2.12 (0.32) |

Each component function $f_j (j = 1, \ldots, 8)$ are appropriately scaled as in Ravikumar et al. [24] and Yin et al. [32].

The simulation results for $p = 1000$ are summarized in the upper panel of Table 1. The complete results for all $p = 50, 200$ and 1000 can be found in the supplementary materials. We can see from Table 1 that SPOT-SICA always outperforms SPAM and SPOT-LASSO in terms of both $F_1$ score and prediction accuracy. The superior performance of SPOT-SICA is due to the use of SICA penalty for variable selection and estimation. Because of the advantages of SICA discussed in Section 3.1, SPOT-SICA can simultaneously screen out more spurious variables and produce less biased estimates of the function components, thus achieve better selection precision and lower prediction error. The performances of SPAM and SPOT-LASSO are mostly comparable in variable selection, because they both use the $L_1$ penalty. In terms of prediction, SPAM and SPOT-LASSO perform similarly in the cases when the predictors are sampled independently ($t = 0$). When data are sampled from more complex structures ($t = 3$ and $t = 6$), SPAM outperforms SPOT-LASSO, since SPOT-LASSO does not utilize the additive structure of the model.

**Example 2.** (Transformation Model)

$$Y = \log\left(4 + \sin(2\pi X_1) + |X_2| + X_3^2 + X_4^3 + X_5 + \epsilon\right)$$

where $\epsilon \sim N(0, 1/4)$. We sample covariates according to the same procedure in Example 1, except that we sample $W_1, \ldots, W_p$ and $U$ now from $Unif(-1, 1)$. The change is to ensure that the term in the *log*-function is positive.

The simulation results for $p = 1000$ are summarized in the lower panel of Table 1. The complete results for all $p = 50, 200$ and 1000 can be found in the supplementary materials. We see that it is consistent in all cases that SPOT-SICA outperforms SPOT-LASSO, and SPOT-LASSO outperforms SPAM, in both variable selection precision and prediction accuracy. In addition, although the results on the average size of estimated supports are similar for SPAM and SPOT-LASSSO, SPAM is much worse compared to SPOT-LASSO in terms of prediction accuracy. This observation supports the claim that the additional transformation on $Y$ in SPOT is providing more flexibility in capturing the complex dependence structure. One sample of the estimated optimal transformations from SPOT-SICA is visualized in Figure 1, which match the true functions well. To further assess the variability of the estimates, we run SPOT-SICA on bootstrapped samples and plot resulting transformations in Figure 1, as suggested in Breiman and Friedman [3].

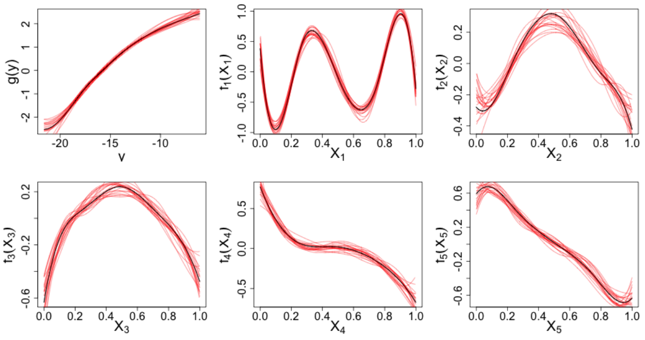

**Figure 1: Transformations of $Y$ and $X_1$ to $X_5$ from SPOT-SICA ($a = 1$) in Example 2 ($p = 50, t = 0$). The black line is the estimated transformation from original data, red lines are estimated transformations from 20 bootstrapped samples.**

**Example 3**. (More transformation models)
Let $V = 4 + \sin(2\pi X_1) + |X_2| + X_3^2 + X_4^3 + X_5 + \epsilon$, where $\epsilon \sim N(0, 1/4)$. We consider the following transformation models,

(3.1) $Y = 20/V$;
(3.2) $Y = 10\sqrt{V}$;

**Table 2: Comparison of different methods on simulated data from Example 3.**

| Model | $p$ | Method | Precision | Recall | Size | $F_1$ score | MSE |
|---|---|---|---|---|---|---|---|
| 3.1 | 1000 | SPAM | 0.15 (0.08) | 0.95 (0.11) | 41.22 (19.74) | 0.24 (0.10) | 1.32 (0.40) |
| 3.1 | 1000 | SPOT-LASSO | 0.24 (0.19) | 0.99 (0.06) | 40.80 (43.76) | 0.35 (0.20) | 0.84 (0.35) |
| 3.1 | 1000 | SPOT-SICA | 0.67 (0.28) | 0.97 (0.11) | 11.81 (18.42) | 0.74 (0.23) | 0.75 (0.28) |
| 3.2 | 1000 | SPAM | 0.13 (0.05) | 0.99 (0.05) | 43.89 (17.47) | 0.23 (0.07) | 4.23 (0.55) |
| 3.2 | 1000 | SPOT-LASSO | 0.22 (0.16) | 0.99 (0.03) | 59.10 (64.71) | 0.33 (0.21) | 2.55 (0.34) |
| 3.2 | 1000 | SPOT-SICA | 0.79 (0.29) | 1.00 (0.03) | 11.16 (19.43) | 0.84 (0.25) | 2.16 (0.27) |
| 3.3 | 1000 | SPAM | 0.13 (0.05) | 0.99 (0.05) | 44.02 (18.92) | 0.23 (0.07) | 2.99 (0.45) |
| 3.3 | 1000 | SPOT-LASSO | 0.24 (0.15) | 0.99 (0.05) | 40.28 (49.65) | 0.37 (0.19) | 1.79 (0.29) |
| 3.3 | 1000 | SPOT-SICA | 0.80 (0.29) | 0.99 (0.04) | 9.63 (12.58) | 0.85 (0.24) | 1.55 (0.24) |
| 3.4 | 1000 | SPAM | 0.14 (0.06) | 0.98 (0.06) | 41.51 (17.39) | 0.24 (0.09) | 2.82 (0.52) |
| 3.4 | 1000 | SPOT-LASSO | 0.25 (0.16) | 0.98 (0.06) | 36.39 (41.96) | 0.37 (0.19) | 1.71 (0.32) |
| 3.4 | 1000 | SPOT-SICA | 0.75 (0.30) | 0.99 (0.04) | 11.68 (17.94) | 0.81 (0.25) | 1.49 (0.25) |
| 3.5 | 1000 | SPAM | 0.15 (0.08) | 0.92 (0.18) | 40.80 (23.17) | 0.24 (0.10) | 0.70 (0.33) |
| 3.5 | 1000 | SPOT-LASSO | 0.24 (0.20) | 0.94 (0.20) | 38.42 (41.32) | 0.35 (0.22) | 0.47 (0.28) |
| 3.5 | 1000 | SPOT-SICA | 0.66 (0.29) | 0.95 (0.14) | 13.29 (22.98) | 0.73 (0.25) | 0.42 (0.25) |

(3.3) $Y = V^2/5$;
(3.4) $Y = \exp\{V/3\}$;
(3.5) $Y = 10\exp\{1/V\}$.

All predictors $X_j$ are generated independently from $Unif(-1, 1)$. Sample size is $n = 200$ for train/validation/test dataset. Results for $p = 1000$ from all models in Example 3 are summarized in Table 2. The complete results for all $p = 50, 200$ and 1000 can be found in the supplementary materials. The results are consistent to the statement in Example 2: SPOT-SICA consistently outperforms SPOT-LASSO, and SPOT-LASSO outperforms SPAM.

**Example 4**. (Some general models) We consider the following general models.

(4.1) $Y = \exp(X_1) + X_2^2\epsilon$;
(4.2) $Y = (1 + X_1)^{X_2} + 0.1\epsilon$;
(4.3) $Y = X_1^3 + X_2^2 X_3 + 0.1\epsilon$;
(4.4) $Y = X_1 + X_2 + (X_3 + X_4)^3 + 0.1\epsilon$;

where $\epsilon \sim N(0, 1)$. All three models considered here do not belong to transformation models. In particular, Model (4.1) represents one case that heterogeneity exists in the model; Model (4.3) incorporates the interaction terms of $X_2$ and $X_3$, or it can be considered that $X_2$ and $X_3$ form a group in the model; Model (4.4) represents another group structure in the model, where the additive term $X_3 + X_4$ can be considered to be in one function. We test our methods on these models to see how they performs under more general model settings.

All predictors $X_j$ are generated independently from $Unif(-1, 1)$. Sample size is $n = 200$ for train/validation/test dataset. Results for $p = 19$ from all models in Example 4 are summarized in Table 3. It is expected that variable selection is more difficult in these models compared to additive models in Example 1 and transformation models in Examples 2 and 3. However, we see from the results that even when the assumption on the transformation model does not hold, our proposed method can still be applied as a fairly effective tool for variable selection.

**Table 3: Comparison of different methods on simulated data from Example 4.**

| | Method | Precision | Recall | Size | $F_1$ score | MSE |
|---|---|---|---|---|---|---|
| 4.1 | SPAM | 0.33 (0.21) | 0.73 (0.25) | 5.92 (3.25) | 0.42 (0.20) | 0.21 (0.04) |
| 4.1 | SPOT-LASSO | 0.54 (0.28) | 0.74 (0.25) | 3.33 (1.78) | 0.60 (0.25) | 0.21 (0.04) |
| 4.1 | SPOT-SICA | 0.58 (0.30) | 0.76 (0.25) | 3.28 (1.97) | 0.63 (0.26) | 0.21 (0.04) |
| 4.2 | SPAM | 0.45 (0.36) | 0.88 (0.26) | 7.64 (6.42) | 0.50 (0.29) | 12.84 (37.45) |
| 4.2 | SPOT-LASSO | 0.44 (0.33) | 0.94 (0.23) | 7.62 (6.35) | 0.53 (0.32) | 11.55 (34.64) |
| 4.2 | SPOT-SICA | 0.58 (0.36) | 0.94 (0.20) | 5.81 (5.61) | 0.65 (0.32) | 11.29 (34.29) |
| 4.3 | SPAM | 0.33 (0.15) | 0.84 (0.17) | 8.81 (3.52) | 0.46 (0.15) | 0.05 (0.01) |
| 4.3 | SPOT-LASSO | 0.61 (0.32) | 0.76 (0.15) | 5.28 (3.44) | 0.62 (0.21) | 0.05 (0.01) |
| 4.3 | SPOT-SICA | 0.79 (0.30) | 0.74 (0.14) | 3.63 (2.48) | 0.72 (0.18) | 0.04 (0.01) |
| 4.4 | SPAM | 0.37 (0.13) | 1.00 (0.00) | 11.88 (3.57) | 0.53 (0.13) | 0.64 (0.11) |
| 4.4 | SPOT-LASSO | 0.81 (0.28) | 1.00 (0.00) | 6.24 (4.03) | 0.86 (0.21) | 0.56 (0.12) |
| 4.4 | SPOT-SICA | 0.93 (0.19) | 1.00 (0.02) | 4.81 (2.56) | 0.95 (0.14) | 0.53 (0.12) |

## 5.2 Role of Parameter $a$ in Variable Selection

In this experiment, using the same model as in Example 2, we investigate the role played by the tuning parameter $a$ in the SICA penalty in model selection accuracy. As discussed in Section 4, SPOT-SICA can achieve variable selection consistency under Assumption (D). The smaller $a$ is, the less restrictive the assumption is. To demonstrate this effect, we choose two values of $a$, $a = 1$ and $\infty$, and compare their performances under different design matrices. We vary $t$ from $\{1, 2, 3, 4, 5, 6\}$ to represent different levels of variable selection difficulty.

For any fixed $a$ and a given sample, the performance of SICA depends on the regularization parameter $\lambda$. SPOT-SICA is declared to have a success if there exists a $\lambda$ under which SPOT-SICA correctly select all true variables. For each value of $t$, we simulate 10 samples of $\mathbf{X}$ that leads to 10 design matrices. For each design matrix, we randomly sample the error term 100 times, and then apply SPOT-SICA and record their successes and failures. Consequently, we obtain 10 success rates, each over 100 random replicates. We plot these success rates at each value of $t$ in Figure 2. From Figure 2, we see that SPOT-SICA ($a = 1$) outperforms SPOT-LASSO ($a = \infty$) by consistently selecting the correct model. As expected, when $t$ increases, selecting the correct model becomes more difficult for both values of $a$. However, SPOT-SICA still have a higher chance to select the correct model even when SPOT-LASSO fails.

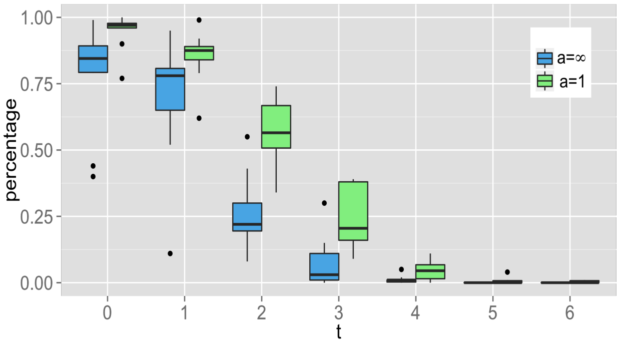

**Figure 2: Impact of $a$ on selection consistency of SPOT under different correlation structure controlled by $t$.**

Next, we choose two fixed values of $t$, which are $t = 0$ and $t = 2$, but vary $a$ from $\{0.05, 0.10, 0.50, 1.00, 2.00, 5.00\}$. For each fixed pair of $t$ and $a$, we repeat the previous procedure and record the average success rate. Results are presented in Table 4. Results from the $L_1$ penalty ($a = \infty$) are also recorded in the last column. We see that as $a$ becomes larger, the performance of the SICA penalty is approaching that of the $L_1$ penalty. When $a$ gets closer to zero, the chance of selecting a true model will first increase and then decrease, this suggests that the computational difficulty increases as the SICA penalty approaches the $L_0$ penalty. The pattern exists for both $t = 0$ and $t = 2$. This phenomenon has also been pointed out in Lv and Fan [19].

**Table 4: Average percentages of times that the true model can be selected with different choices of $a$.**

| $a$ | 0.05 | 0.10 | 0.50 | 1.00 | 2.00 | 5.00 | ∞ |
|---|---|---|---|---|---|---|---|
| $t$=0 | 0.945 | 0.967 | 0.982 | 0.947 | 0.903 | 0.834 | 0.781 |
| $t$=2 | 0.573 | 0.632 | 0.679 | 0.578 | 0.427 | 0.318 | 0.256 |

## 5.3 Real Data Application

We apply SPOT-SICA to two real datasets from the UCI Machine Learning Repository[1], which are the Boston Housing Data and the Communities and Crime Data.

*5.3.1 Boston Housing Data.* The *Boston Housing Data* was collected to study the house values in the suburbs of Boston; The dataset contains n=506 observations with 10 covariates, which are RM, AGE, DIS, TAX, PTRATIO, BLACK, LSTAT, CRIM, INDUS, NOX. To explore the variable selection property of SPOT-SICA, we follow the approach of Ravikumar et al. [24] and add 20 noise variables in the analysis. The first ten noise variables are randomly drawn from $Unif(0, 1)$, and the other ten noise variables are a random permutation of the original ten covariates.

We adopt the commonly used "one-standard-error" rule with 10-fold cross-validation to select the tuning parameters $\lambda$ and $a$, where we choose the most parsimonious model whose error is no more than one standard error above the error of the best model. We apply the SPOT-SICA to the 30 dimensional dataset with the selected tuning parameters. SPOT-SICA correctly zeros out both types of irrelevant variables, and it identifies five nonzero components out of the original ten covariates. The important variables are RM, DIS, TAX, PTRATIO, LSTAT. The estimated transformation functions are depicted in Figure 3. From Figure 3, we found that the monotone transformation of the response may be needed to yield a better fitted model. Furthermore, aside from the commonly recognized important variables, which are RM, TAX, PTRATIO and LSTAT, SPOT-SICA suggests that DIS is also important, which exhibits a clear nonlinear effect on the response MEDV.

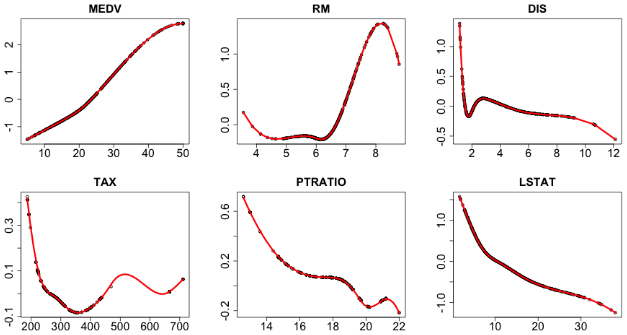

**Figure 3: Estimated transformations of the response and selected predictors by SPOT-SICA for the Boston Housing Data.**

*5.3.2 Communities and Crime Data.* The *Communities and Crime Data* was first collected in Redmond and Baveja [25] and it consists of 1994 observations from 128 variables including ethnicity proportions, income, poverty rate, divorce rate etc., and was previously analyzed by [21] and [29]. We consider modeling the violent crime rate from other covariates in the dataset. By removing the covariates with missing values, we narrow down to 98 covariates.

We apply SPOT-SICA to the dataset, with tuning parameters selected by 10-fold cross-validation and the "one-standard-error" rule. Out of 98 covariates, 10 are selected by SPOT-SICA as important variables in modeling the violent crime rate, which is fewer than 24 variables reported in Maldonado and Weber [21]. Moreover, the resulting estimates from SPOT-SCIA exhibit a higher prediction accuracy, with an average out-of-bag mean absolute error smaller than 0.093, which is better than the results from the proposed method in Maldonado and Weber [21]. The obtained transformations from SPOT-SICA are depicted in Figure 4. The labels above each graph corresponds to the orders of the covariates in the original data[2]. It is interesting to observe a clear nonlinear transformation of the response. Additionally, most selected variables exhibit nonlinear effects on the transformed response and a few others have nearly

[1] http://archive.ics.uci.edu/ml/

[2] http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime

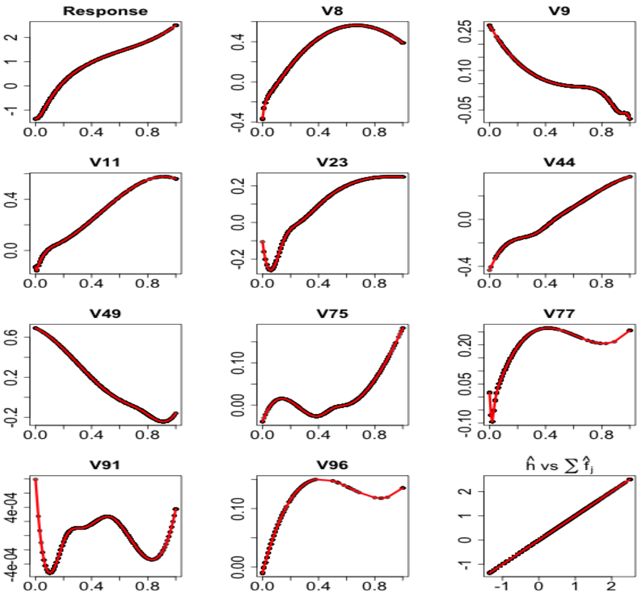

**Figure 4: Estimated transformations of the response and 10 selected predictors from SPOT-SICA for the Communities and Crime Data. The last graph is the plot of the estimated response transformation against the sum of all selected components.**

linear effects. Thus, our method effectively reduces the dimensionality of the data and is able to capture sensible linear/nonlinear relationships between the response and covariates.

## 6 CONCLUSIONS

In this article, we develop a novel method called SPOT for exploring the dependence structure between the response $Y$ and the predictor vector $\mathbf{X}$ in high dimensional data analysis. SPOT can consistently select important variables and automatically generate meaningful optimal transformations, under which the dependence structure can be best explored. SPOT demonstrates promising results on both simulated and real data in terms of selection consistency, estimation accuracy, prediction power, and interpretability. One interesting direction to improve SPOT is to consider further transformations in addition to the optimal transformations, in order to capture the dependence of $Y$ and $\mathbf{X}$ missed by optimal transformations. Another future direction is to investigate more relaxed conditions under which SPOT can possess selection and estimation consistency.

## REFERENCES

[1] Sivaraman Balakrishnan, Kriti Puniyani, and John D Lafferty. 2012. Sparse Additive Functional and Kernel CCA. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*. 911–918.

[2] Leo Breiman. 1996. Heuristics of Instability and Stabilization in Model Selection. *The Annals of Statistics* (1996), 2350–2383.

[3] Leo Breiman and Jerome H Friedman. 1985. Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association* 80, 391 (1985), 580–598.

[4] Pierre-André Chiappori, Ivana Komunjer, and Dennis Kristensen. 2015. Nonparametric identification and estimation of transformation models. *Journal of Econometrics* 188, 1 (2015), 22–39.

[5] Carl De Boor. 2001. A practical guide to splines, revised Edition, Vol. 27 of Applied Mathematical Sciences. (2001).

[6] Sam Efromovich. 2007. Conditional density estimation in a regression setting. *The Annals of Statistics* (2007), 2504–2535.

[7] Jianqing Fan and Runze Li. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.* 96, 456 (2001), 1348–1360.

[8] Jianqing Fan, Qiwei Yao, and Howell Tong. 1996. Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika* 83, 1 (1996), 189–206.

[9] Yingying Fan, Gareth M James, and Peter Radchenko. 2015. Functional additive regression. *The Annals of Statistics* 43, 5 (2015), 2296–2325.

[10] Trevor J Hastie and Robert J Tibshirani. 1990. *Generalized additive models*. Vol. 43. CRC Press.

[11] Joel L Horowitz and Enno Mammen. 2004. Nonparametric estimation of an additive model with a link function. *The Annals of Statistics* 32, 6 (2004), 2412–2443.

[12] Jian Huang, Patrick Breheny, and Shuangge Ma. 2012. A selective review of group selection in high-dimensional models. *Statistical science: a review journal of the Institute of Mathematical Statistics* 27, 4 (2012).

[13] Jian Huang, Joel L Horowitz, and Fengrong Wei. 2010. Variable selection in nonparametric additive models. *Annals of statistics* 38, 4 (2010), 2282.

[14] David Jacho-Chávez, Arthur Lewbel, and Oliver Linton. 2010. Identification and nonparametric estimation of a transformed additively separable model. *Journal of Econometrics* 156, 2 (2010), 392–407.

[15] Wei Lin and Jinchi Lv. 2013. High-dimensional sparse additive hazards regression. *J. Amer. Statist. Assoc.* 108, 501 (2013), 247–264.

[16] Yi Lin and Hao Helen Zhang. 2006. Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics* 34, 5 (2006), 2272–2297.

[17] Oliver Linton, Stefan Sperlich, and Ingrid Van Keilegom. 2008. Estimation of a semiparametric transformation model. *The Annals of Statistics* (2008), 686–718.

[18] Po-Ling Loh and Martin J Wainwright. 2015. Regularized M-estimators with Nonconvexity: Statistical and Algorithmic Theory for Local Optima. *Journal of Machine Learning Research* 16 (2015), 559–616.

[19] Jinchi Lv and Yingying Fan. 2009. A unified approach to model selection and sparse recovery using regularized least squares. *The Annals of Statistics* (2009), 3498–3528.

[20] Jinchi Lv and Jun S Liu. 2014. Model selection principles in misspecified models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76, 1 (2014), 141–167.

[21] Sebastián Maldonado and Richard Weber. 2010. Feature selection for support vector regression via kernel penalization. In *Neural Networks (IJCNN), The 2010 International Joint Conference on*. IEEE, 1–7.

[22] Lukas Meier, Sara Van de Geer, and Peter Bühlmann. 2009. High-dimensional additive modeling. *The Annals of Statistics* 37, 6B (2009), 3779–3821.

[23] Mila Nikolova. 2000. Local strong homogeneity of a regularized estimator. *SIAM J. Appl. Math.* 61, 2 (2000), 633–658.

[24] Pradeep Ravikumar, John Lafferty, Han Liu, and Larry Wasserman. 2007. SpAM: Sparse Additive Models. In *Advances in Neural Information Processing Systems*. 1201–1208.

[25] Michael Redmond and Alok Baveja. 2002. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research* 141, 3 (2002), 660–678.

[26] Murray Rosenblatt. 1969. Conditional probability density and regression estimators. *Multivariate Analysis II* 25 (1969), 31.

[27] Larry Schumaker. 1981. *Spline functions: basic theory*. Wiley, New York.

[28] Gideon Schwarz. 1978. Estimating the dimension of a model. *The Annals of Statistics* 6, 2 (1978), 461–464.

[29] Le Song, Eric P Xing, and Ankur P Parikh. 2011. Kernel embeddings of latent tree graphical models. In *Advances in Neural Information Processing Systems*. 2708–2716.

[30] Masashi Sugiyama, Ichiro Takeuchi, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Daisuke Okanohara. 2010. Conditional density estimation via least-squares density ratio estimation. In *International Conference on Artificial Intelligence and Statistics*. 781–788.

[31] Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), 267–288.

[32] Junming Yin, Xi Chen, and Eric P Xing. 2012. Group Sparse Additive Models. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*. 871–878.

[33] Ming Yuan and Yi Lin. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68, 1 (2006), 49–67.

[34] Peng Zhao and Bin Yu. 2006. On model selection consistency of Lasso. *The Journal of Machine Learning Research* 7 (2006), 2541–2563.

[35] Hui Zou and Runze Li. 2008. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of statistics* 36, 4 (2008), 1509.