Forward Selection and Estimation in High Dimensional Single Index Models

Shikai Luo and Subhashis Ghosal
North Carolina State University

August 29, 2016

Abstract

We propose a new variable selection and estimation technique for high dimensional single index models with unknown monotone smooth link function. Among many predictors, typically, only a small fraction of them have significant impact on prediction. In such a situation, more interpretable models with better prediction accuracy can be obtained by variable selection. In this article, we propose a new penalized forward selection technique which can reduce high dimensional optimization problems to several one dimensional optimization problems by choosing the best predictor and then iterating the selection steps until convergence. The advantage of optimizing in one dimension is that the location of optimum solution can be obtained with an intelligent search by exploiting smoothness of the criterion function. Moreover, these one dimensional optimization problems can be solved in parallel to reduce computing time nearly to the level of the one-predictor problem. Numerical comparison with the LASSO and the shrinkage sliced inverse regression shows very promising performance of our proposed method.

Keywords: Forward selection; High dimension; Sparsity; Single index model; Penalization; Variable selection

1 Introduction

Information extraction from high dimensional data is a fundamental statistical problem and has many important applications. Due to rapid advances of scientific techniques and computing resources, overwhelmingly large data are collected in various fields of sciences and industries in various form, such as images, gene-expressions and internet data. How to build accurate, interpretable and computationally efficient statistical models is a key question for high dimensional data analysis.

High dimensional data typically involves a lot of predictors, many of which are nearly irrelevant. Removing irrelevant variables from the predictive model is essential since the presence of too many variables may cause overfitting, which leads to poor prediction of future outcomes. Nowadays, regularized regression methods are widely used for variable selection in linear models or even generalized linear models. The LASSO (Tibshirani, 1996) is a popular method for regression that uses an ℓ_1-penalty to achieve a sparse solution. Other regularization approaches similar to the LASSO were also proposed in the literature (Fan and Li, 2001; Zou and Hastie, 2005; Yuan and Lin, 2006; Tibshirani et al., 2005).

Park and Hastie (2007) proposed an ℓ_1-regularization path algorithm for generalized linear models with a known link function. Based on coordinate descent, Friedman et al. (2010) proposed a fast algorithm to compute the regularization paths for generalized linear models with penalties include ℓ_1-norm (the LASSO), squared ℓ_2-norm (ridge regression) and a combination of the two (the elastic net).

In some situations, parametric specification of a link function may not be possible. A model where the distribution of the response variable depends on a linear combination of predictors is called a single index model. These models are a lot more flexible than linear models yet the coefficients there are still easily interpretable, making these models widely popular. However, classical approaches to estimation in a single index model apply only when the number of predictors is relatively few. To treat higher dimensional predictors, the estimation procedure must be accompanied by a variable selection step. Recently, several approaches have been proposed to estimate parameters of single index models in the high-dimensional setting. Wang and Yin (2008) extended the MAVE method of Xia et al. (2002) by introducing an ℓ_1-regularization. Peng and Huang (2011) minimized a penalized least squares criterion to perform automatic variable selection in single index models. However, these methods are applicable only in the case of $n > p$.

Methods based on sufficient dimension reduction (Cook, 2009) have also been used to estimate in single index models. Ni et al. (2005), Li and Yin (2008), Bondell and Li (2009), Wang et al. (2012) and Zhu and Zhu (2009) considered regularization methods in the inverse regression framework. These methods use structural properties of single index models to estimate regression coefficients without having to estimate the link function. Radchenko (2015) on the other hand considered a direct approach for estimation in a high dimensional

single index model based on penalized least squares by expanding the link function in a B-spline basis expansion. He showed that the minimizer exits and a solution path can be obtained, and that under mild conditions on ambient dimension, sparsity level, the number of basis elements and the tuning parameter, the resulting estimator converges to the truth at a certain rate.

Recently, Hwang et al. (2009) proposed a penalized forward selection technique for variable selection in linear regression. The procedure, called the forward iterative regression and shrinkage technique (FIRST), reduces p-dimensional optimization problems to several one dimensional optimization problems each of which admits an analytical solution. FIRST typically leads to smaller prediction error and substantially sparser models than common methods like the LASSO without compromising the predictive power. Moreover, the componentwise approach to the minimization of the objective function extends to similar optimization problems related to the LASSO.

In this paper, we combine both forward selection and penalization to propose variable selection and estimation technique for a high dimensional single index model with an unknown monotone increasing smooth link function. The monotonicity of the link function makes the single index regression coefficients more easily interpretable as the direction of steepest increase and is essential for our ordering based method.

The model we consider in this paper is as follows:

$$Y|\boldsymbol{X} \sim P(\cdot, g(\boldsymbol{X}^T\boldsymbol{\beta})) \tag{1.1}$$

where $P(\cdot, \theta)$ is a stochastically increasing family with scalar parameter θ, $\boldsymbol{X}$ is a p-dimensional covariate, $\boldsymbol{\beta}$ is the corresponding coefficient which is assumed to be of unit norm, and g is an unknown but strictly increasing and smooth function. In this semiparametric single index model setting, Y depends on $\boldsymbol{X}$ only through one single index $\boldsymbol{X}^T\boldsymbol{\beta}$ connected by the link function g. One special case of the single index model (1.1) is given as:

$$Y = g(\boldsymbol{X}^T\boldsymbol{\beta}) + \varepsilon \tag{1.2}$$

where ε is a random variable with mean 0 and finite variance.

We consider estimation in the high dimensional situation where p is relatively large,

possibly larger than the sample size n. Under a sparsity condition that most predictors are irrelevant, we can estimate $\boldsymbol{\beta}$ based on relatively low sample sizes, like what the LASSO does for linear models by putting an ℓ_1-penalty on regression coefficients. In the next section, we propose a penalization method of estimating $\boldsymbol{\beta}$ under sparsity for a single index model. Several simulation results to judge the performance of the proposed method is presented in Section 3. A real data analysis is presented in Section 4. Finally the paper concludes with a discussion in Section 5.

2 Selection and Estimation Procedure

Suppose the data we observed are $(\boldsymbol{X}_i, Y_i), i = 1, 2, \ldots, n$, which are independently and identically distributed. We propose a simultaneous variable selection and estimation procedure by maximizing the monotone association between Y and $\boldsymbol{X}^T\boldsymbol{\beta}$ under an ℓ_1-penalty. Note that because of non-linearity, Pearson's correlation coefficient is no longer a suitable measure of association. A more meaningful measure is given by Kendall's tau coefficient between Y and $\boldsymbol{X}^T\boldsymbol{\beta}$, since monotonicity of g implies that increments of Y are likely to be synchronized with those of $\boldsymbol{X}^T\boldsymbol{\beta}$, except for the random disturbance due to the error. Thus a natural estimator of $\boldsymbol{\beta}$ can be proposed by maximizing the following association between Y and $\boldsymbol{X}^T\boldsymbol{\beta}$ with respect to $\boldsymbol{\beta}$:

$$\tau_n(\boldsymbol{\beta}) = \frac{1}{n(n-1)} \sum_{1 \le i_1 \ne i_2 \le n} \text{sign}(Y_{i_2} - Y_{i_1}) \text{sign}(\boldsymbol{X}_{i_2}^T\boldsymbol{\beta} - \boldsymbol{X}_{i_1}^T\boldsymbol{\beta}). \tag{2.1}$$

In spite of its appeal, $\tau_n(\boldsymbol{\beta})$ is not continuous in $\boldsymbol{\beta}$. In fact, it is a step function and hence it necessarily has multiple modes. Therefore, we use a smoothed version of $\tau_n(\boldsymbol{\beta})$ given by:

$$\tau_n^*(\boldsymbol{\beta}) = \frac{1}{n(n-1)} \sum_{1 \le i_1 \ne i_2 \le n} \text{sign}(Y_{i_2} - Y_{i_1}) \tanh\{(\boldsymbol{X}_{i_2}^T\boldsymbol{\beta} - \boldsymbol{X}_{i_1}^T\boldsymbol{\beta})/h\}, \tag{2.2}$$

where $\tanh(x) = (e^x - e^{-x})/(e^x + e^{-x})$ is hyperbolic tangent function and h is a small constant used to smooth the sign function. It may be noted that the smoothing in (2.2) does not use a density function as the kernel and there is no bias variance trade-off, unlike in common smoothing problems like density estimation or nonparametric regression. The sole purpose

of the smoothing is to make the maximization problem free of multi-modality. Thus the value of the bandwidth parameter arbitrarily close to zero is preferable to emulate the sign function, although the value $h = 0$ cannot be used. We found that a value $h \in [0.1, 1]$ works as a good choice. A value of h is chosen and fixed. The best sparse linear combination will be searched by a forward selection method coordinate-wise for each predictor as done in FIRST. Our procedure can be described by the following algorithm.

For a given $\lambda > 0$ and $\epsilon > 0$, do the following steps:

Algorithm

Step 1. Find the index j_1 that maximizes the absolute value of T_j with respect to j where

$$T_j := \frac{1}{n(n-1)} \sum_{i_1 \neq i_2} \text{sign}(Y_{i_2} - Y_{i_1}) \text{sign}(\boldsymbol{X}_{i_2 j} - \boldsymbol{X}_{i_1 j})$$

and set $\hat{\boldsymbol{\beta}}^{(1)} = \text{sign}(T_{j_1}) \mathbf{e}_{j_1}$, where $\mathbf{e}_j = (0, \ldots, 0, 1, 0, \ldots, 0)^T$ with 1 at the jth entry and 0 at other entries, $j = 1, \ldots, p$.

Step 2. Suppose that $X_{j_1}, \ldots, X_{j_{k-1}}$ are already selected and current single index coefficient is $\hat{\boldsymbol{\beta}}^{(k-1)}$. For all remaining $j \notin \{j_1, \ldots, j_{k-1}\}$, we solve the following optimization problems in parallel if desired:

$$\hat{\beta}_j = \text{argmax}_{\beta_j} \{\tau_n^*(\hat{\boldsymbol{\beta}}^{(k-1)} + \beta_j \mathbf{e}_j) - \lambda |\beta_j|\}.$$

Let $j_k = \text{argmax}\{\tau_n^*(\hat{\boldsymbol{\beta}}^{(k-1)} + \hat{\beta}_j \mathbf{e}_j) : j \notin \{j_1, \ldots, j_{k-1}\}\}$. If $\tau_n^*(\hat{\boldsymbol{\beta}}^{(k-1)} + \hat{\beta}_{j_k} \mathbf{e}_{j_k}) - \tau_n^*(\hat{\boldsymbol{\beta}}^{(k-1)}) < \epsilon$, we stop this step and set $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{(k-1)}$. Otherwise, we set

$$\hat{\boldsymbol{\beta}}^{(k)} = (\hat{\boldsymbol{\beta}}^{(k-1)} + \hat{\beta}_{j_k} \mathbf{e}_{j_k}) / \|\hat{\boldsymbol{\beta}}^{(k-1)} + \hat{\beta}_{j_k} \mathbf{e}_{j_k}\|_2,$$

where $\|\cdot\|_2$ denotes the ℓ_2-norm, and repeat this step until we stop.

Step 3. Re-estimate β by maximizing (2.2) with selected variables from *Step 2*.

Since $\tau_n^*(\beta)$ is between -1 and 1, the procedure stops in at most $2/\epsilon$ steps. The resulting procedure will be called Smoothed Kendall's association Iterative Maximizer Model Selector (SKIMMS).

For the single index model (1.2), the proposed method SKIMMS along with a kernel smoothing step gives rise to an estimator of the unknown link function g. We apply a simple isotonic regression method to estimate $g(t)$ after obtaining the coefficient vector $\hat{\boldsymbol{\beta}}$. We first denote $Z_i = \boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}$, and continue the following steps:

- Sort $\{Z_1, \ldots, Z_n\}$ in increasing order, denote the sorted sequence by $\{Z_{(1)}, \ldots, Z_{(n)}\}$, and then concordantly sort $\{Y_1, \ldots, Y_n\}$ as $\{Y_{(1)}, \ldots, Y_{(n)}\}$ that may not be sorted;

- Run a pool-adjacent-violators (PAV) algorithm (Ayer et al., 1955; Hanson et al., 1973; Burdakov et al., 2004) on the concordantly sorted Y's and denote the resulting sequence by $\{Y^{(1)}, \ldots, Y^{(n)}\}$. Specifically, let $D_n = \{(Z_{(i)}, Y_{(i)}), i = 1, \ldots, n\}$. Then PAV algorithm computes a non-decreasing sequence of values $\{Y^{(i)}, i = 1, \ldots, n\}$ such that

$$S = \sum_{i=1}^{n} (Y^{(i)} - Y_{(i)})^2$$

  is minimized which consists of three steps:

  - D_{r+1} is formed by extending D_r with a data point $(Z_{(r+1)}, Y_{(r+1)})$;
  - If the values $Y^{(1)}, \ldots, Y^{(r)}$ denote the solution obtained for D_r, then a preliminary solution for D_{r+1} is formed by setting $Y^{(r+1)} = Y_{(r+1)}$. Thereafter, the final solution for D_{r+1} is derived by pooling adjacent $Y^{(\cdot)}$-values that violate the monotonicity constraints. To be more precise, $Y^{(r+1)}, \ldots, Y^{(r+1-k)}$ are assigned the common value $(Y^{(r+1)} + \cdots + Y^{(r+1-k)})/(k+1)$, where k is the smallest non-negative integer such that the new values of $Y^{(1)}, \ldots, Y^{(r+1)}$ form a non-decreasing sequence.
  - The optimal solution $\{Y^{(1)}, \ldots, Y^{(n)}\}$ for D_n is composed of clusters of identical values.

- Take a smooth, symmetric kernel function $K(t)$, and estimate g as follows:

$$\hat{g}(t) = \frac{\sum_{j=1}^{n} K(\frac{t - Z_{(j)}}{b}) \times Y^{(j)}}{\sum_{j=1}^{n} K(\frac{t - Z_{(j)}}{b})},$$

where b is chosen by a generalized cross validation technique (Craven and Wahba, 1978; Wahba, 1990). The function $\hat{g}$ can be easily shown to be monotone increasing, smooth and the monotonicity is strict as long as there are at least two different Y's.

Then for any given new observed $\boldsymbol{X}_{\text{new}}$, we can predict the corresponding response as $\hat{Y}_{\text{new}} = \hat{g}(\boldsymbol{X}_{\text{new}}^T \hat{\boldsymbol{\beta}})$

2.1 Tuning

There is one extremely important parameter in above procedure determining its performance, namely λ. It needs to be tuned carefully to assure a proper selection and estimation result. In our numerical experiments, we use five-fold cross validation, i.e., the data set is randomly split into five approximately equal parts. A set of possible values of λ are chosen over a grid and the Kendall's tau coefficient $\tau_n(\boldsymbol{\beta})$ is computed on one-fifth of the data set using the estimate of $\boldsymbol{\beta}$ from the rest of the data. The one-fifth part of the data is revolved over different choices to minimize the effect of uneven random splits of the data. This is in line with the FIRST procedure introduced by Hwang et al. (2009) for linear models. The value of λ maximizing the five-fold average of Kendall's tau coefficient is then chosen in the final selection and estimation stage using all data. The tuning process can be manipulated in parallel over the grid of λ to increase the computation speed.

3 Simulation Studies

In this section, we present several examples to demonstrate the performance of our proposed method. For each example, we vary the number of observations n and the correlation ρ between covariates. The number of observations is taken as either $n = 100$ or $n = 200$, while p is fixed at 150. Both the $n < p$ and $n > p$ situations are considered. We consider our method, LASSO and also the shrinkage sliced inverse regression (SSIR) (Ni et al., 2005). For SSIR, we fix the dimension of effective dimension reduction space as 1 for single index model. The R package `edrGraphicalTools` is used to select the optimal number of slices of SSIR. We only provide results for SSIR when $n = 200$, $p = 150$ since SSIR is very slow and is a poor performer in $p > n$ settings. For combinations with $p = 150$, $n = 100, 200$ and $\rho = 0.25, 0.5$,

each experiment is repeated 100 times. In our simulation, we fix $h = 0.5, \epsilon = 0.001$ and b is chosen via generalized cross validation (Craven and Wahba, 1978; Wahba, 1990).

Example 1. The data is generated according to the following simple model,

$$Y_i = \arctan(2\boldsymbol{X}_i^T\boldsymbol{\beta}) + 0.2\varepsilon_i, \quad i = 1, \ldots, n. \tag{3.1}$$

Example 2. The data comes from another model with more significant predictors:

$$Y_i = \exp(\boldsymbol{X}_i^T\boldsymbol{\beta}) + 0.5\varepsilon_i, \quad i = 1, \ldots, n. \tag{3.2}$$

Example 3. We consider another more difficult model with 9 significant predictors:

$$Y_i = \exp(\arctan(2\boldsymbol{X}_i^T\boldsymbol{\beta})) + 0.2\varepsilon_i, \quad i = 1, \ldots, n. \tag{3.3}$$

where $\boldsymbol{\beta} = (3, 5, 3, 5, 0, \ldots, 0)^T$ for Example 1, $\boldsymbol{\beta} = (3, 3, -4, -4, 5, 5, 0, \ldots, 0)^T$ for Example 2 and $\boldsymbol{\beta} = (3, 3, 3, -4, -4, -4, 5, 5, 5, 0, \ldots, 0)$ for Example 3. $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^T$ is a p-dimensional predictor generated from a multivariate normal distribution $N(0, \boldsymbol{\Sigma})$ with entries of $\boldsymbol{\Sigma} = (\!(\sigma_{ij})\!)_{p\times p}$ being $\sigma_{ii} = 1, i = 1, \ldots, p$, and $\sigma_{ij} = \rho, i \neq j$, the noise ε_i is independent of the predictors, and is generated from standard normal with 10% of the outliers following the Cauchy distribution.

The results for Examples 1–3 are shown in Tables 1–3, respectively. In the tables, FP denotes false positives, FN denotes false negatives, PE denotes prediction error based on test dataset of size 1000; mean of FP, FN, PE over 100 simulation replicates are reported with corresponding standard deviation in the parentheses. Note that SSIR does not estimate the link function and hence it cannot be used for prediction. Thus the PE entries for SSIR in the tables are left blank.

From Tables 1–3, we can see that SSIR fails to select most of the important variables. From Table 1, we can see that the proposed approach SKIMMS selects true variables more correctly as evidenced by significantly lower false positives (FP), false negatives (FN) and prediction errors (PE). Table 2 shows that SKIMMS gives higher false positives, but much lower false negatives and prediction error. As the sample size increases, FP, FN and PE

decreases for both SKIMMS and LASSO. However, the rate of decrease of SKIMMS is much faster than that of the LASSO. For Example 3, we can see from Table 3 that SKIMMS gives much lower false positives but slightly higher false negatives than the LASSO. Because of the inherent difficulty of defining the true model when predictors are highly correlated with each other, SKIMMS may miss some true predictors in the models it selects. In particular, the models selected by SKIMMS are typically considerably simpler than those selected by the LASSO and moreover, SKIMMS has smaller prediction error than the LASSO.

Table 1: Simulation results for Example 1

	ρ	Method	FP	FN	PE
$n = 100$	0.25	SKIMMS	4.10(2.23)	0(0)	0.06(0.01)
		LASSO	8.31(4.39)	0(0)	0.21(0.02)
	0.50	SKIMMS	3.10(2.02)	0.02(0.14)	0.06(0.02)
		LASSO	9.23(3.99)	0.09(0.32)	0.25(0.02)
$n = 200$	0.25	SKIMMS	1.44(1.07)	0(0)	0.04(0.00)
		LASSO	7.98(4.12)	0(0)	0.18(0.01)
		SSIR	8.56(3.57)	2.75(1.28)	-
	0.50	SKIMMS	1.22(1.11)	0(0)	0.04(0.00)
		LASSO	8.98(3.31)	0(0)	0.23(0.01)
		SSIR	8.72(3.50)	2.97(1.21)	-

Table 2: Simulation results for Example 2

	ρ	Method	FP	FN	PE
$n = 100$	0.25	SKIMMS	9.16(3.71)	0.14(0.35)	1.25(0.60)
		LASSO	5.4(6.01)	1.51(1.92)	2.52(0.87)
	0.50	SKIMMS	7.39(3.12)	0.7(0.83)	1.12(0.51)
		LASSO	5.1(5.64)	2.31(1.95)	1.93(0.63)
$n = 200$	0.25	SKIMMS	7.1(3.29)	0(0)	0.66(0.24)
		LASSO	3.92(4.05)	0.69(1.62)	2.16(0.87)
		SSIR	5.44(2.24)	4.74(1.31)	-
	0.50	SKIMMS	4.73(2.13)	0.04(0.20)	0.51(0.15)
		LASSO	4.59(5.93)	1.1(1.53)	1.65(0.52)
		SSIR	6.10(3.02)	4.21(1.39)	-

Table 3: Simulation results for Example 3

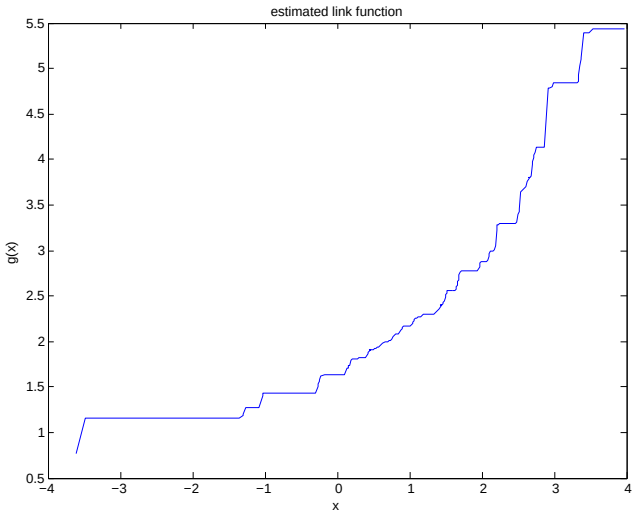
	ρ	Method	FP	FN	PE
$n = 100$	0.25	SKIMMS	6.64(3.06)	0.33(0.89)	0.15(0.18)
		LASSO	18.7(7.83)	0.03(0.17)	0.37(0.05)
	0.50	SKIMMS	5.97(2.72)	1.03(1.11)	0.20(0.16)
		LASSO	17.11(8.54)	0.35(0.86)	0.38(0.07)
$n = 200$	0.25	SKIMMS	4.52(1.57)	0(0)	0.05(0.01)
		LASSO	18.73(8.37)	0(0)	0.28(0.02)
		SSIR	5.53(2.75)	7.57(1.47)	-
	0.50	SKIMMS	4.33(1.21)	0.14(0.35)	0.06(0.03)
		LASSO	18.67(7.07)	0(0)	0.28(0.02)
		SSIR	6.91(2.79)	6.89(1.60)	-

4 Real Data Analysis

We use the Boston housing data to illustrate the proposed method SKIMMS. This dataset consist of values of houses in suburbs of Boston along with several potential predictors of house prices. This dataset was taken from the StatLib library maintained at Carnegie Mellon University and is available at `https://archive.ics.uci.edu/ml/datasets/Housing`. The data contains 506 instances with 13 continuous attributes and 1 binary valued attribute. All variables have been scaled to have standard deviation one before analysis. It turns out that the variables selected by SKIMMS and the LASSO are quite similar. The common variables selected by the two approaches include per capita crime rate by town, Charles River dummy variable, average number of rooms per dwelling, weighted distances to five Boston employment centers, pupil-teacher ratio by town, proportion of blacks by town and lower status of the population. The proposed SKIMMS procedure also selects another variable called index of accessibility to radial highways. The LASSO selects two other variables, proportion of residential land zoned for lots and nitric oxides concentration. We compare the five fold cross-validation prediction errors of the proposed method SKIMMS with that of the LASSO. We find that while the former is 0.230, the prediction error of the LASSO is significantly larger at 0.328. Furthermore, as we can see from the figure below, the estimated link function given by the proposed method shows obvious non-linearity. It appears that in the beginning the link function is somewhat flat, but increases steeply in the later portion. This indicates that when the good factors are combined the right way, house values can

increase very quickly, while on the lower end point a threshold needs to be crossed before the factors can start positively affecting the house prices. For instance, a lot of square footage in a less-desirable neighborhood may not add much value, while that can hugely affect house values in desirable neighborhoods. This seems to go well with common senses in the housing market. The non-linearity of the underlying link function also explains why SKIMMS has considerably smaller prediction error than the LASSO. Using the estimates of $\boldsymbol{\beta}$ and g given by SKIMMS, house prices can be automatically appraised reasonably well.

Figure 1: Estimated Link Function



5 Discussion

In this paper, we suggest a new forward selection and estimation method for high dimensional sparse single index model using a recursive algorithm that iterates over possible predictors. This method takes an advantage of easy and fast computation of one dimensional penalized optimization with LASSO penalty and forward selection of best predictors that maximize a smoothed version of Kendall's tau coefficient. We have derived an iterative algorithm which can proceed in parallel to save substantial computing time. Throughout the simulation study and a real data example, we show that our algorithm shows very competitive performance in model prediction and selection accuracy when compared with the LASSO and shrinkage sliced inverse regression (SSIR). Our algorithm has reasonable computational costs, which make the method useful for analyzing high dimensional sparse data.

Acknowledgment: We thank the referee for pointing to us an important reference paper. Research of the second author is partially supported by National Security Agency grant number H98230-12-1-0219.

References

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W., Silverman, E., et al. (1955). An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics*, 26(4):641–647.

Bondell, H. D. and Li, L. (2009). Shrinkage inverse regression estimation for model-free variable selection. *Journal of the Royal Statistical Society: Series B*, 71(1):287–299.

Burdakov, O., Grimvall, A., and Hussian, M. (2004). A generalised pav algorithm for monotonic regression in several variables. In *COMPSTAT, Proceedings of the 16th Symposium in Computational Statistics*, pages 761–767. Citeseer.

Cook, R. D. (2009). *Regression Graphics: Ideas for Studying Regressions Through Graphics*, volume 482. John Wiley and Sons.

Craven, P. and Wahba, G. (1978). Smoothing noisy data with spline functions. *Numerische Mathematik*, 31(4):377–403.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1.

Hanson, D. L., Pledger, G., and Wright, F. (1973). On consistency in monotonic regression. *The Annals of Statistics*, pages 401–421.

Hwang, W. Y., Zhang, H. H., and Ghosal, S. (2009). First: Combining forward iterative selection and shrinkage in high dimensional sparse linear regression. *Statistics and Its Interface*, 2:341–348.

Li, L. and Yin, X. (2008). Sliced inverse regression with regularizations. *Biometrics*, 64(1):124–131.

Ni, L., Cook, R. D., and Tsai, C.-L. (2005). A note on shrinkage sliced inverse regression. *Biometrika*, 92(1):242–247.

Park, M. Y. and Hastie, T. (2007). l_1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B*, 69(4):659–677.

Peng, H. and Huang, T. (2011). Penalized least squares for single index models. *Journal of Statistical Planning and Inference*, 141(4):1362–1379.

Radchenko, P. (2015). High dimensional single index models. *Journal of Multivariate Analysis*, 139:266–282.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, pages 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B*, 67(1):91–108.

Wahba, G. (1990). *Spline Models for Observational Data*, volume 59. SIAM.

Wang, Q. and Yin, X. (2008). A nonlinear multi-dimensional variable selection method for high dimensional data: Sparse MAVE. *Computational Statistics and Data Analysis*, 52(9):4512–4520.

Wang, T., Xu, P.-R., and Zhu, L.-X. (2012). Non-convex penalized estimation in high-dimensional models with single-index structure. *Journal of Multivariate Analysis*, 109:221–235.

Xia, Y., Tong, H., Li, W., and Zhu, L.-X. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B*, 64(3):363–410.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68(1):49–67.

Zhu, L.-P. and Zhu, L.-X. (2009). Nonconcave penalized inverse regression in single-index models with high dimensional predictors. *Journal of Multivariate Analysis*, 100(5):862–875.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2):301–320.