Statistica Sinica **26** (2016), 527-546
doi:http://dx.doi.org/10.5705/ss.202014.0151

KERNEL ADDITIVE SLICED INVERSE REGRESSION

Heng Lian and Qin Wang

University of New South Wales and Virginia Commonwealth University

Abstract: In recent years, nonlinear sufficient dimension reduction (SDR) methods have gained increasing popularity. While there is a large literature on semiparametric models in regression, parsimonious structured nonlinear SDR has attracted little attention so far. In this paper, extending kernel sliced inverse regression, we study additive models in the context of SDR and demonstrate its potential usefulness due to its flexibility and parsimony. We clarify the improved convergence rate using additive structure due to the faster rate of decay of the kernel's eigenvalues. Additive structure also opens the possibility of nonparametric variable selection. This sparsification of the kernel, however, does not introduce additional tuning parameters, in contrast with sparse regression. Simulations and data sets are presented to illustrate the benefits and limitations of the approach.

Key words and phrases: Kernel method, nonlinear dimension reduction, sliced inverse regression, variable selection.

1. Introduction

In the classical theory of linear sufficient dimension reduction, with a p-dimensional predictor X and a univariate response Y as in the regression setting, we say the subspace spanned by a $p \times d$ matrix B with $d \leq p$ is a sufficient dimension reduction space if $Y \perp X | B^T X$, where $\perp$ denotes independence. Thus, $B^T X$ summarizes the information in the predictors relevant to predicting Y. Under mild assumptions, the intersection of all SDR spaces is itself an SDR space, termed the central subspace (Cook (1994, 1996, 1998); Yin, Li, and Cook (2008)).

Under some mild assumptions including the linear design condition, sliced inverse regression (SIR) extracts directions in the central subspace by the eigenvectors of the matrix

$$Cov(X)^{-1} Cov(E[X|Y]), \tag{1.1}$$

which can be easily estimated by slicing the range of Y given a sample and hence, the name of SIR (Li (1991)). Wu (2008) and Yeh, Huang, and Lee (2009) extended standard SIR to nonlinear dimension reduction via the kernel method. The kernel method is popular in machine learning that maps predictors into a typically infinite-dimensional space and performs linear operations in this new feature space that correspond to nonlinear operations when mapped back to the original space. This can be formulated via the theory of reproducing kernel Hilbert spaces (RKHS), a familiar topic in the statistical literature. Fukumizu, Bach, and Jordan (2004, 2009) and Fukumizu and Leng (2014) proposed the use of a cross-covariance operator on RKHS to characterize the conditional independence $Y \perp X | B^T X$ so as to achieve linear dimension reduction.

As with all other fully nonparametric approaches, the kernel method suffers from low convergence rate when p is sufficiently large. In regression, semiparametric models such as those with additive structures have a long history in statistics (Stone (1985); Liang et al. (2008); Xue (2009); Wang et al. (2011); Ma (2012)). Such efforts are not present in nonlinear SDR. Commonly used product/tensor kernels in multi-dimensional setting try to incorporate all-way interactions, infeasible and uninteresting, and using such kernels is sub-optimal when additive structures are present.

In Section 2, we show that additive structures in kernel SIR can be easily realized by using the additive kernel instead of the product/tensor kernel. Establishing theoretical advantages for doing so is nontrivial. We establish the convergence rate for kernel SIR (KSIR) and clarify that the faster convergence rate in kernel additive SIR (KASIR) is related to the faster rate of decay in the eigenvalues of the additive kernel operator. In the special case that the RKHS is the periodic Sobolev space and that the true dimension reduction directions have an additive structure, the convergence rate is $n^{-2m/(2m+1)}$ where m is the smoothness parameter of the Sobolev space, the well-known optimal rate in regression (for fixed dimension asymptotics). Unlike nonparametric regression, kernel SIR requires no optimization procedure. In Section 3, we propose a method to sparsify the additive kernel that corresponds to sparse additive models in regression for nonparametric variable selection. However, the optimization problem in sparse kernel additive SIR (SKASIR) turns out to be much harder and thus we only investigate the case of p relatively small (up to about 20) compared to sample size. Another notable practical difference from sparse additive regression is that no extra tuning parameter is introduced in SKASIR. Section 4 contains our simulation studies as well as an application to a data set to demonstrate the performance of KASIR and SKASIR, in comparison with KSIR and standard linear SIR. We conclude the paper with a discussion on limitations and future plans. The proof of the main result is contained in the supplementary material.

2. Kernel Additive SIR using Additive Kernel

We first review some theory of RKHS and kernel sliced inverse regression proposed in Wu (2008); Yeh, Huang, and Lee (2009); Wu, Liang, and Mukherjee

(2013). Let $L^2(P_X)$ be the space of square integrable functions with probability measure P_X on R^p. Given a kernel $K(\cdot, .)$, positive definite on $R^p \times R^p$, with the spectral decomposition

$$K(x,y) = \sum_{j=1}^{\infty} \lambda_j \phi_j(x)\phi_j(y), \tag{2.1}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ are the eigenvalues and ϕ_j are the eigenfunctions orthonormal in $L^2(P_X)$, the induced RKHS is

$$\mathcal{H}_K = \Big\{ f : f(x) = \sum_j a_j \phi_j(x), \text{ with } \sum_j \frac{a_j^2}{\lambda_j} < \infty \Big\}.$$

The inner product on $\mathcal{H}_K$ is $\langle f, g\rangle_{\mathcal{H}} = \sum_j a_j b_j/\lambda_j$ if $f = \sum_j a_j \phi_j$ and $g = \sum_j b_j \phi_j$. As $\langle f, K(\cdot,x)\rangle_{\mathcal{H}} = f(x)$, K is called the reproducing kernel. A popular approach in the machine learning literature constructs the feature map $\Phi : x \to \Phi(x) = (\sqrt{\lambda_1}\phi_1(x), \sqrt{\lambda_2}\phi_2(x), \ldots) \in l_2$, where l_2 is the the space of square-summable sequences and thus, based on (2.1), $K(x,y)$ is just the inner product $\langle \Phi(x), \Phi(y)\rangle_2$ in l_2. Using this feature map, performing standard linear SIR in the feature space l_2 corresponds to nonlinear SIR in the original space R^p. Thus, following the SIR procedure (1.1), we can extract directions $\beta \in l_2$ from the eigenvalue problem

$$Cov(E[\Phi(X)|Y])\beta = \mu Cov(\Phi(X))\beta,$$

where $Cov(\Phi(X)) = E[(\Phi(X) - E[\Phi(X)]) \otimes (\Phi(X) - E[\Phi(X)])]$, for example. Mathematically, since $\langle K(\cdot,x), K(\cdot,y)\rangle_{\mathcal{H}} = K(x,y)$, we can just take the feature map to be $x \to K(\cdot,x) \in \mathcal{H}_K$, and the eigenvalue problem becomes

$$\Gamma f = \mu \Sigma f, \tag{2.2}$$

where $\Gamma = Cov(E[K(\cdot,X)|Y])$ and $\Sigma = Cov(K(\cdot,X))$. For simplicity of notation we assume without loss of generality that $E_X K(\cdot,X) = 0$, where the subscript in the expectation indicates the variable over which the expectation is taken, for clarity.

Given an i.i.d. sample $(X_i, Y_i), i = 1, \ldots, n$, the relevant covariance operators can be easily estimated by slicing moment estimators as in linear SIR; $\Sigma = Cov(K(\cdot,X)) = E_X[K(\cdot,X) \otimes K(\cdot,X)]$ can be estimated by $\Sigma_n = n^{-1}\sum_i (K(\cdot,X_i) - \overline{K(\cdot,X)}) \otimes (K(\cdot,X_i) - \overline{K(\cdot,X)})$ where $\overline{K(\cdot,X)} = n^{-1}\sum_i K(\cdot,X_i)$. If $E_X K(X,X) < \infty$, Σ has a spectral decomposition, say

$$\Sigma = \sum_j \lambda_j \psi_j \otimes \psi_j,$$

with $\lambda_1 \geq \lambda_2 \geq \cdots$ and $\langle \psi_j, \psi_k \rangle_{\mathcal{H}} = \delta_{jk}$. With the assumption $E_X K(\cdot, X) = 0$, by direct calculation $\Sigma f = E_X K(\cdot, X) f(X)$, and thus the eigenvalues of Σ and K (as in (2.1)) are the same. The eigenvectors here are chosen to be orthonormal in $\mathcal{H}$ instead of in $L^2(P_X)$. Even without the assumption $E_X K(\cdot, X) = 0$, it can be shown that eigenvalues of Σ and K decay at the same rate which does not affect our arguments, for example in Proposition 1, which only depend on the rate of decay of eigenvalues.

To obtain the slicing estimator of $\Gamma = Cov(E[K(\cdot, X)|Y])$, the range of Y is divided into H slices and we estimate Γ by

$$\Gamma_n = \sum_{h=1}^{H} \hat{p}_h (\overline{K_h(\cdot, X)} - \overline{K(\cdot, X)}) \otimes (\overline{K_h(\cdot, X)} - \overline{K(\cdot, X)}),$$

where $\overline{K_h(\cdot, X)}$ is the average of $K(\cdot, X_i)$ concomitant to the Y_i in the hth slice, and $\hat{p}_n = n_h/n$ where n_h is the number of observations in the hth slice. To stabilize the eigenvalue problem, a scalar multiple of the identity operator I is added to Σ_n resulting in

$$\Gamma_n \hat{f} = \mu(\Sigma_n + sI)\hat{f}. \tag{2.3}$$

To find the eigenfunction above, the representer theorem for kernel SIR allows us to write $\hat{f} = \sum_{i=1}^n c_i (K(\cdot, X_i) - \overline{K(\cdot, X)})$ and, plugging this expression into the above displayed equation, the eigenvalue problem can be written in terms of $\mathbf{c} = (c_1, \ldots, c_n)$ as

$$\mathbf{KJKc} = \mu \mathbf{K}(\mathbf{K} + s\mathbf{I})\mathbf{c}. \tag{2.4}$$

This is used for computation, where $\mathbf{K}$ is the centered $n \times n$ kernel matrix, $\mathbf{J}$ is the $n \times n$ matrix with $\mathbf{J}_{ij} = 1/n_h$ if Y_i and Y_j are in the hth slice and zero otherwise.

By the representer theorem, it is easy to see that if we use kernel K that has an additive form, $K(x, y) = K_1(x_1, y_1) + \cdots + K_p(x_p, y_p)$ for $x = (x_1, \ldots, x_p), y = (y_1, \ldots, y_p)$ and p kernels $K_1, \ldots, K_p$, then f (as well as all functions in $\mathcal{H}_K$) also has this additive form. Theoretically, we need to assume the true f, the eigenfunction in (2.2), is in the RKHS generated by $K(x, y) = K_1(x_1, y_1) + \cdots + K_p(x_p, y_p)$. This is equivalent to saying that one can write $f(x) = f_1(x_1) + \cdots + f_p(x_p)$ and f_j is in the RKHS generated by K_j.

Although it is trivial to incorporate additive structure into KSIR, it is nevertheless difficult to see how this additively structured kernel induces faster convergence rates. Wu, Liang, and Mukherjee (2013) have shown the consistency of KSIR but a meaningful convergence rate remains elusive. The main result in this paper clarifies the role of the kernel's eigenvalues $\lambda_1, \lambda_2, \ldots,$ in determining the convergence rate. We first present the theorem for general kernels and then discuss its implication for the faster convergence rate for additive kernels. The following assumptions are used.

(A) $\exists C > 0$ such that $K(x,x) < C$ for all x in the range of the predictor.

(B) $\lambda_j \asymp j^{-d}$ for some $d > 1$.

(C) The operator $\Sigma^{-1}\Gamma$ has an eigenfunction f associated with its largest eigenvalue μ that has multiplicity one. We let $\hat{f}$ be the eigenfunction of $(\Sigma_n + sI)^{-1}\Gamma_n$ associated with its largest eigenvalue.

(D) The response is discrete and can take values only in $\{y_1, \dots, y_H\}$.

(E) If the SDR space is generated by $\{h_1, \dots, h_r\} \subset \mathcal{H}_K$, the linear $E(g(x)|h_1(x), \dots, h_r(x))$ is linear in $h_1(x), \dots, h_r(x)$ for $g \in \mathcal{H}_K$.

The assumption of uniform boundedness of K in (A) is required in our proof to show $K(\cdot,x) \otimes K(\cdot,x)$ is a bounded operator for all x. When we consider the range of x in a compact set as when we use the Sobolev kernel defined on $[0,1]$ later, assumption (A) is a very mild regularity assumption, which does imply $\sum_j \lambda_j < \infty$. Assumption (B) was used in Blanchard, Bousquet, and Massart (2008) and Caponnetto and De Vito (2007) to establish oracle inequalities for support vector machines classification and regression, respectively. That eigenvalues play a critical role in convergence rates is expected in regression since the Rademacher complexity of the RKHS can be exactly characterized by these eigenvalues (Bartlett and Mendelson (2003); Koltchinskii and Yuan (2010); Raskutti, Wainwright, and Yu (2012)). The polynomial decay assumption holds in some special cases, see Koltchinskii and Yuan (2010). If it does not hold, we can possibly derive some rate in terms of the specific values of $\lambda_1, \lambda_2, \dots$. However, the expression would be messy and it would be hard to see the effect of eigenvalues on the convergence rate. Assumption (C) simply re-states the estimator and the population counterpart for clearness. It can be shown that, under our assumptions, the eigenspace of $(\Sigma_n + sI)^{-1}\Gamma_n$ associated with its largest eigenvalue also has multiplicity one with probability approaching one. For simplicity, we only consider the first dimension reduction direction, the eigenfunction associated with the largest eigenvalue. Rates for subsequent directions can be shown with some additional arguments. Assumption (D) is typically assumed in the SIR literature, for example in Cook and Ni (2005) to simplify analysis, which directly applies to classification problems and is also reasonable for regression since the slicing estimator will in effect quantize the responses. Assumption (E) is used in Wu (2008); Yeh, Huang, and Lee (2009); Wu, Liang, and Mukherjee (2013). (E) implies that $E[K(\cdot,X)|Y = y_h] \in \text{span}\{\Sigma h_1, \dots, \Sigma h_r\}$ and, in particular, that $\Sigma^{-1}\Gamma$ is a bounded operator.

Theorem 1. *If* (A)–(E) *hold, and* $s = c_n n^{-d/(d+1)} \to 0$ *with* $c_n \to \infty$, *then*

$$min_{c \in \{-1,1\}} E_{X^*}[(c\hat{f}(X^*) - f(X^*))^2] = O_p(c_n n^{-d/(d+1)}),$$

where X^* *is an independent copy of* X.

The constant c is necessary here since the eigenfunction can only be identified up to sign change. This rate is optimal in regression up to an extra arbitrarily slowly diverging sequence c_n. This may be due to our method of proof but it is not clear to us how to improve this.

We can obtain an improved convergence rate if the kernel's eigenvalue has a fast decay to zero. For kernel methods, a commonly used kernel is the Gaussian $K(x,y) = \exp\{-a\|x-y\|^2\}$, or the more flexible form $K(x,y) = \exp\{-\sum_{j=1}^{p} a_j(x_j-y_j)^2\}$. Another common example from smoothing splines is the Sobolov space of order m with kernel $K(s,t) = \sum_{\nu=1}^{m}(s-1)^{\nu-1}(t-1)^{\nu-1}/((\nu-1)!)^2 + \int_0^1 (s-u)_+^{m-1}(t-u)_+^{m-1}/((m-1)!)^2 du$; the multivariate version is constructed by taking the product of one-dimensional kernels. For mth order Sobolev space of periodic functions, it is known that the eigenvalues decay at the rate j^{-2m}.

Let $\mathcal{H}_j$ be the RKHS induced by the one-dimensional kernel K_j. Suppose $\phi_{j1}, \phi_{j2}, \ldots$, are the eigenfunctions of K_j with eigenvalues $\lambda_{jk} \asymp k^{-d}$. Let $K^{(p)}(x,y) = \prod_j K_j(x_j,y_j)$ and $K^{(s)}(x,y) = \sum_j K_j(x_j,y_j)$. The following simple proposition shows that the eigenvalues of $K^{(s)}$ decay at the same rate as each K_j when the coordinates are independent.

Proposition 1. *Under this setup, and assuming that the p predictors are independent, the eigenvalues of $K^{(s)}(x,y)$ are of order j^{-d}.*

Proof. The RKHS associated with $K^{(s)}$ is the space of functions of the form $\sum_{j=1}^{p} f_j(x_j)$ (Aronszajn (1950)) with $f_j \in \mathcal{H}_j$. Using $E_X f_j(X) = 0$, it is easy to see that $Kf = (\sum_j K_j)(\sum_j f_j) = \sum_j K_j f_j$. Thus if $f = \sum_j f_j$ satisfies the eigenvalue equation $Kf = \lambda f$, we have $K_j f_j = \lambda f_j$, which implies that for each j, either f_j is an eigenvector of K_j with (common) eigenvalue λ, or $f_j = 0$. This in turn means that the set of eigenvalues of K is a subset of $\{\lambda_{jk}\}$ and the multiplicity of each eigenspace is at most p. Thus the decay rate of the eigenvalues of K is the same as that of its additive component.

In a p-dimensional space, when the function $K(\cdot, y)$ is in the Hölder class of smoothness $2m-p$ for all y, the eigenvalues of a positive definite kernel are upper bounded by $j^{-2m/p}$ under mild assumptions, and a kernel can be constructed that achieves this rate of decay (Kühn (1987)). An illustration here is the mth order Sobolev space of period functions on $[0,1]$ whose kernel has smoothness $2m-1$, with eigenvalues known to decay exactly as j^{-2m}. In general, although it is unclear whether the rate of decay $j^{-2m/p}$ is achieved by a specific kernel, it is generally believed that $j^{-2m/p}$ is the typical rate, which leads to the convergence rate of KSIR of $n^{-2m/(2m+p)}$, the same as the minimax rate for nonparametric regression in a p-dimensional space.

It seems hard to infer the eigenvalues of the product kernel $K^{(p)}$ based on the eigenvalues of the K_j, although it is natural to conjecture that the eigenvalues

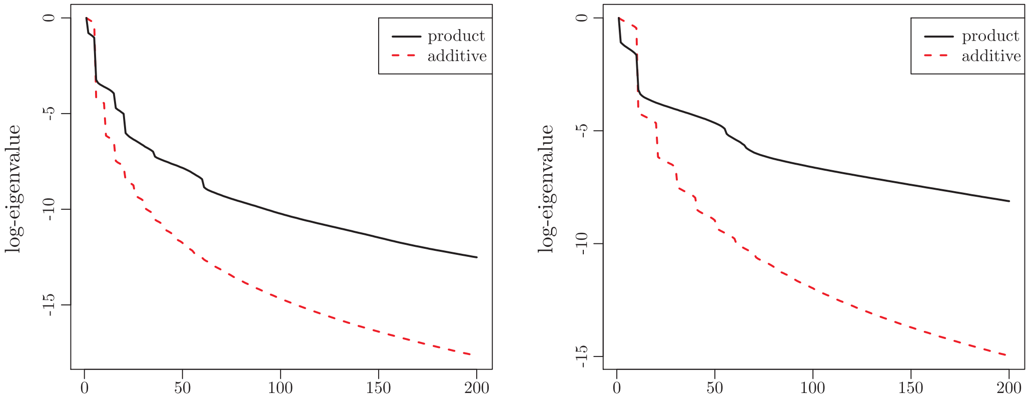

Figure 1. Logarithms of eigenvalues for the product (black and solid curve) and additive kernel (red and dashed curve). Left: $p = 5$; Right: $p = 10$.

of $K^{(p)}$ decay at a slower rate than those of K_j. To illustrate this numerically, we generated 100 sets of predictors with sample size $n = 300$ and $p = 5$ independently and uniformly distributed on $[0, 1]^p$. We then constructed the $n \times n$ kernel matrices of both the product type and the additive type using the kernel for the 2nd order Sobolev space. The averaged (over 100 data sets) logarithm of the largest 200 eigenvalues for the two kernel matrices are shown in the left panel of Figure 1. To facilitate comparison the eigenvalues are scaled such that the largest eigenvalue is always one. It is seen that the eigenvalues of the additive kernel (red and dashed curve) decay faster than those of the product kernel. The right panel of the same figure shows the results with $p = 10$. With larger p, the gap between the two curves is visually larger. Based on our results, the faster decay of the eigenvalues of the additive kernel leads to a faster convergence rate, if the additive assumption is valid.

3. Sparse Kernel Additive SIR

The additive kernel used in the previous section is $K(x, y) = K_1(x_1, y_1) + \cdots + K_p(x_p, y_p)$, equally weighted. To take into account the differing importance of the predictors, we can add a nonnegative weight to each component such that $K(x, y) = d_1K_1(x_1, y_1) + \cdots + d_pK_p(x_p, y_p)$, where a larger weight d_j roughly implies the more important role of the jth predictor, and $d_j = 0$ removes the jth predictor from the model. We use a data-driven procedure to determine the weights along with the sufficient dimension reduction space.

The variable selection problem for KASIR is fundamentally different from various sparse sliced inverse regression methods proposed previously for linear SDR (Li and Nachtsheim (2006); Li (2007); Bondell and Li (2009); Chen, Zou, and Cook (2010)). In linear SDR, sparse method naturally imposes sparsity on the eigenvector of an appropriately defined eigenvalue problem. In KASIR, although the SDR space is also estimated from the eigenvalue problem (2.4), sparsity of the eigenvector **c** does not perform variable selection.

The form of the additive kernel with weights $d_1K_1(x_1, y_1)+\cdots+d_pK_p(x_p, y_p)$ is similar to that used in multiple kernel learning, with the slight difference being that multiple kernel learning usually focuses on obtaining a good kernel for prediction rather than for variable selection and thus the kernels are not necessarily defined on different predictors. Our additive kernel with weights is also related to COSSO (Lin and Zhang (2006); Zhang (2006); Storlie et al. (2011)) which is an approach for component selection in additive splines, or more generally splines analysis of variance. Although many kernel algorithms have been able to incorporate kernel learning, for KASIR this poses much difficulty in computation.

It is well-known that the quotient trace problem

$$\max_{\mathbf{C}\in R^{p\times r},\{d_j\}} \operatorname{tr}((\mathbf{C}^T\mathbf{K}(\mathbf{K}+s\mathbf{I})\mathbf{C})^{-1}\mathbf{C}^T\mathbf{KJKC}),$$

is solved by the eigenvalue problem $\mathbf{KJKc}_i = \mu_i\mathbf{K}(\mathbf{K}+s\mathbf{I})\mathbf{c}_i$ associated with the largest r eigenvalues, where $\mathbf{c}_i, i = 1, \ldots, r$ are the columns of $\mathbf{C}$. Thus when weights are incorporated in the kernel, we can solve the following

$$\begin{aligned}&\max_{\mathbf{C}\in R^{p\times r},\{d_j\}} \operatorname{tr}((\mathbf{C}^T\mathbf{K}(\mathbf{K}+s\mathbf{I})\mathbf{C})^{-1}\mathbf{C}^T\mathbf{KJKC})\\ &s.t. \sum_j d_j = \tau, d_j \geq 0,\end{aligned}$$

where the kernel matrix $\mathbf{K} = \sum_j d_j\mathbf{K}_j$ implicitly depends on d_j. Unlike lasso problem where τ is treated as a tuning parameter (Tibshirani (1996)), here we can set the bound to be 1. It is straightforward to see that the constrained maximization problem with constraint $\sum_j d_j = \tau$ and smoothing parameter s is the same as the problem with constraint $\sum_j d_j = 1$ and smoothing parameter s/τ, in the sense that the maximizer $\mathbf{C}$ and the maximum value are the same. Since we choose the smoothing parameter s in the data-driven way, there is no loss of generality in setting $\tau = 1$. While it is more flexible to use two tuning parameters for controlling smoothness and variable selection separately, using one is not rare in additive regression, as in COSSO (Lin and Zhang (2006); Storlie et al. (2011)).

Given $\{d_j\}$, $\mathbf{C}$ can be obtained from the eigenvalue problem. However, given $\mathbf{C}$, the optimization problem is neither convex nor concave and finding $\{d_j\}$ is hard. For given $\{d_j\}$, we can solve the eigenvalue problem to get $\mathbf{C}$, written as $\mathbf{C}(\{d_j\})$ to emphasize the dependence on $\{d_j\}$. We can then use a general nonlinear solver for the problem

$$\max_{d_j, j=1\ldots,p} \ \mathrm{tr}((\mathbf{C}(\{d_j\})^T\mathbf{K}(\mathbf{K}+s\mathbf{I})\mathbf{C}(\{d_j\}))^{-1}\mathbf{C}(\{d_j\})^T\mathbf{KJKC}(\{d_j\}))$$
$$s.t. \sum_j d_j = 1, d_j \geq 0.$$

An alternative is to solve $\mathbf{C}$ for given $\{d_j\}$ and solve $\{d_j\}$ for given $\mathbf{C}$ (again using a general nonlinear solver). This alternating algorithm has trouble in achieving convergence, in our experience.

4. Numerical Examples

4.1. Simulations

Three simulation examples were used to compare five methods: KSIR, KASIR, SKASIR and standard linear SIR, and SSIR, a sparse version of linear SIR. The data were generated from three models:

1. $Y_i = 20\sin(X_{i1}X_{i2})/(1+\exp\{-3X_{i3}\}) + \epsilon_i, i = 1, \ldots, n,$
2. $Y_i = 10(\sin(3X_{i1}) + X_{i2})\log(|\sin(3X_{i1}) + X_{i2}|) + \epsilon_i, i = 1, \ldots, n,$
3. $Y_i = (1.5X_{i1} + 2X_{i2} - X_{i3})\exp\{1.5X_{i1} + 2X_{i2} - X_{i3}\} + \epsilon_i, i = 1, \ldots, n,$

where $\epsilon_i \overset{i.i.d.}{\sim} N(0,1)$. The predictors were generated from a multivariate Gaussian distribution with mean zero and covariance $Cov(X_{ij_1}, X_{ij_2}) = 0.2^{|j_1-j_2|}$, and then transformed to $[0,1]$ by applied the standard normal cdf. The first example is a general nonlinear model; the second has an additive structure while being nonlinear and KASIR is expected to perform well; the third is linear.

We took $n = 50$ or 100, and dimension $p = 10$. We used the kernel for the 2nd order Sobolev space. KSIR used the product of these one-dimensional kernels and KASIR/SKASIR used the sum of these one-dimensional kernels. Nonlinear optimization in SKASIR was implemented using the nloptr package in R. Although linear SIR and SSIR could be implemented in more traditional ways, we treated linear SIR/SSIR as a special case of KSIR using the linear kernel $K(s,t) = 1 + st$. The smoothing provided by KSIR could be advantageous even in linear SIR. For the smoothing parameter s, 15 equally spaced values in $[-7, 5]$ were used on the logarithmic scale. The number of slices was 10 in all numerical examples, and generally the results are not sensitive to any reasonable choice of the number of slices.

To quantify the performance of the methods, we generated independent test data of size n and computed the (absolute value of) Spearman correlation between the estimated index and the response on the test data. This is possible since in all simulated examples we only have one index. The whole procedure was repeated 100 times in each scenario. The results are reported in Figures 2−4. In these figures, dotted curves show the 0.1 and 0.9 quantiles over the 100 repetitions. The x-axis of this plot is the logarithm of the smoothing parameter $\log(s)$. The second and third rows show the values of d_j for SKASIR and SSIR, respectively, using the smoothing parameter that achieves the largest correlation value.

For Example 1, the model does not have additive structure, thus it is somewhat surprising to see that, when $n = 50$, both SSIR and SKASIR outperform KSIR, with SIR and KASIR (without variable selection) performing similarly to KSIR in terms of Spearman correlation. This suggests that the nonlinearity in Example 1 is not sufficiently strong and with small sample size, more parsimonious models can still be very competitive even though the model assumption is wrong. When $n = 100$, all methods become similar. We further performed simulations for this example with $n = 50$ and $p = 5$ with results shown in Figure 5. With a smaller dimension, we expect KSIR suffers less from the curse of dimensionality and, indeed, the results demonstrate that with $p = 5$, KSIR has the best performance. Returning to Figure 2, both SKASIR and SSIR put large weights on the first two predictors, followed by the third predictor. The contribution of the third predictor to the response is relatively small, as expected from the form of the regression function since the exponential function does not vary a lot on $[0, 1]$. Although the weights d_j for $j > 3$ can be shrunk to exactly zero in some cases, the kernel produced is often not sparse enough, especially when $n = 100$. On the other hand, even when the weights are not exactly zero, the weights for irrelevant predictors are generally much smaller and can still provide information on variable importance.

Example 2 is an ideal scenario for KASIR/SKASIR, which outperformed other methods. Sparse methods can also improve on non-sparse counterparts, for both additive and linear SIR. Additive models correctly identify the first two predictors as important, while linear methods only identify the second predictor. This is a natural since the first predictor has a strong nonlinear effect in the index (and designed such that it has no obviously increasing or decreasing trend for $X_1 \in [0, 1]$). SKASIR can select a much sparser model, while in SSIR none of the weights are sufficiently close to zero. The linear model is not sufficiently flexible, so all predictors strive to compensate for this by playing some role in prediction.

Example 3 has a linear index and thus linear standard SIR performs well and sparsity can improve performance to a small extent. However, KASIR/SKASIR performs almost the same as linear methods. In particular, the curve for SKASIR follows closely that for SSIR and the curve for KASIR follows closely that for SIR.

KSIR is the worst performer in this example. SKASIR and SSIR separate the first three predictors as important, although weights for SKASIR are somewhat less sparse.

For Example 2, we also carried out simulations with $n = 100$ and $p = 20$, with results shown in Figure 6. The results are qualitatively similar to the case $p = 10$. The additive model KASIR/SKASIR has the best performance and the important predictors are correctly identified.

4.2. NMMAPS data analysis

We used the NMMAPS (National Morbidity Mortality Air Pollution Study) database which contains daily mortality, weather and pollution data for 1987-2000, and considered data for the year 1997. We explored the relationship between daily mean ozone level and some predictors. The explanatory variables selected were mean temperature, relative humidity, mean CO_2 level, mean PM_{10} level, mean SO_2 level, daily humidity range, and daily temperature range. After excluding one day with missing observations, we had a sample size of $n = 364$. Scatterplot of the daily mean ozone level against the mean temperature in Figure 7 clearly shows some nonlinearity in the data, although this observation by itself does not mean nonlinear dimension reduction is more appropriate than linear dimension reduction.

With data, it is harder to assess the performance of different methods. We randomly partitioned the whole data set into a training part and a testing part of equal sizes. We performed dimension reduction using the five methods on the training data, with a sequence of smoothing parameters as used in simulations, and considered the number of indices r (projection directions) from 1 to 4. Gaussian process regression was used to learn a nonparametric function that maps the values of the index to the response. We used the R package tpg for Gaussian process regression with the default parameters choices. We also used the R package rpart for regression tree to fit a nonparametric regression function with default parameters choices. However, the predictions with trees were generally worse than Gaussian process regression, so we choose Gaussian process regression even though it is much slower. Using the estimated index and regression function, prediction mean squared errors (PMSE) on the test data is reported on Table 1 based on 100 random partitions of the data. The errors reported are based on the pair of (r, s) values that produce the smallest error for each method. It is seen that KSIR and KASIR perform similarly, and better than linear methods. Kernel weights do not help in prediction in this data set although it may help in interpreting the importance of different predictors. The average kernel weights for SKASIR were $(0.375, 0.111, 0.171, 0.038, 0.051, 0.132, 0.122)$, showing that the most important predictor appears to be the mean temperature. We show the

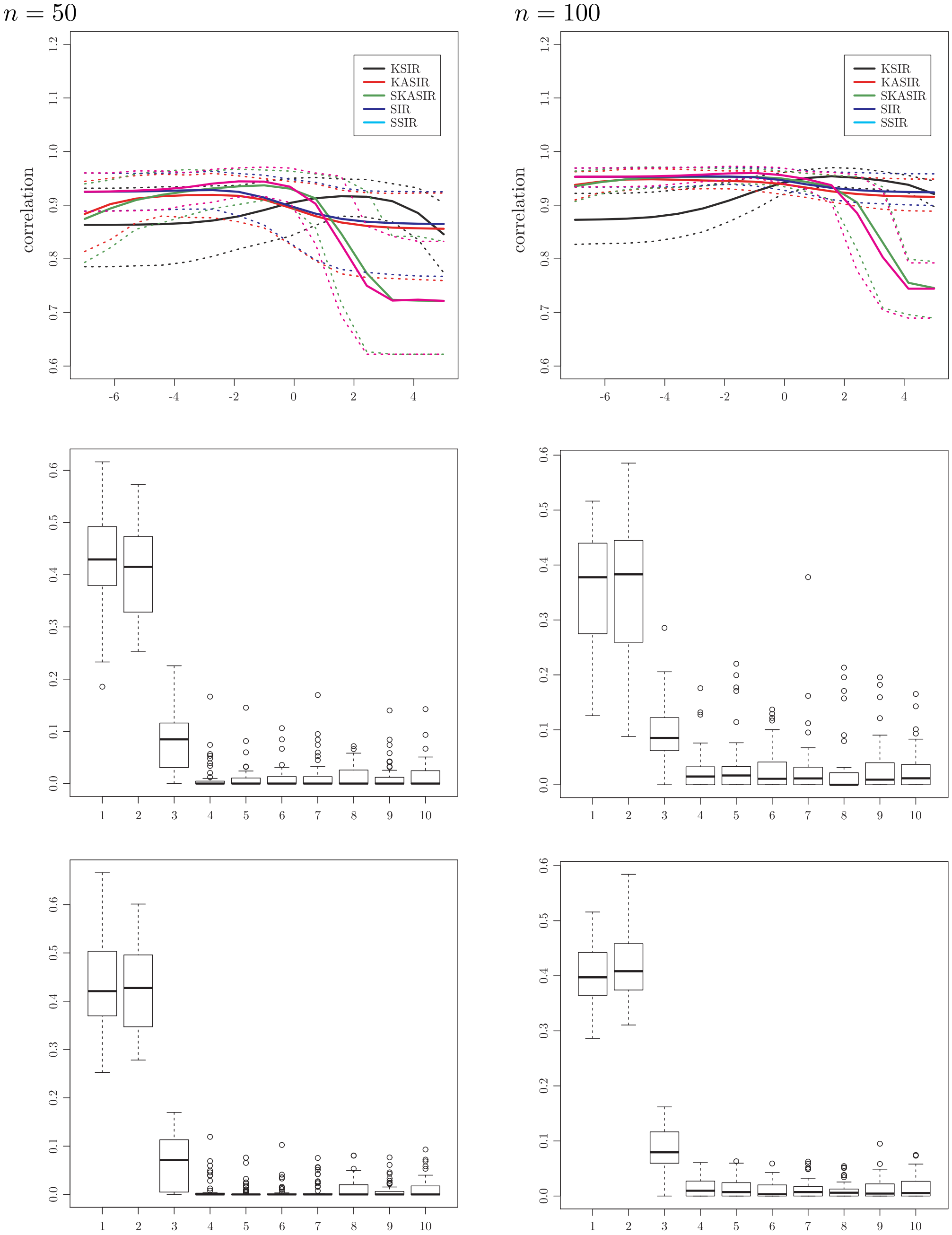

Figure 2. For Example 1, the Spearman correlation on test data for the five methods (KSIR, KASIR, SKASIR, SIR, SSIR) is shown in the first row. The second (third) row shows the values of d_j for SKASIR (SSIR).

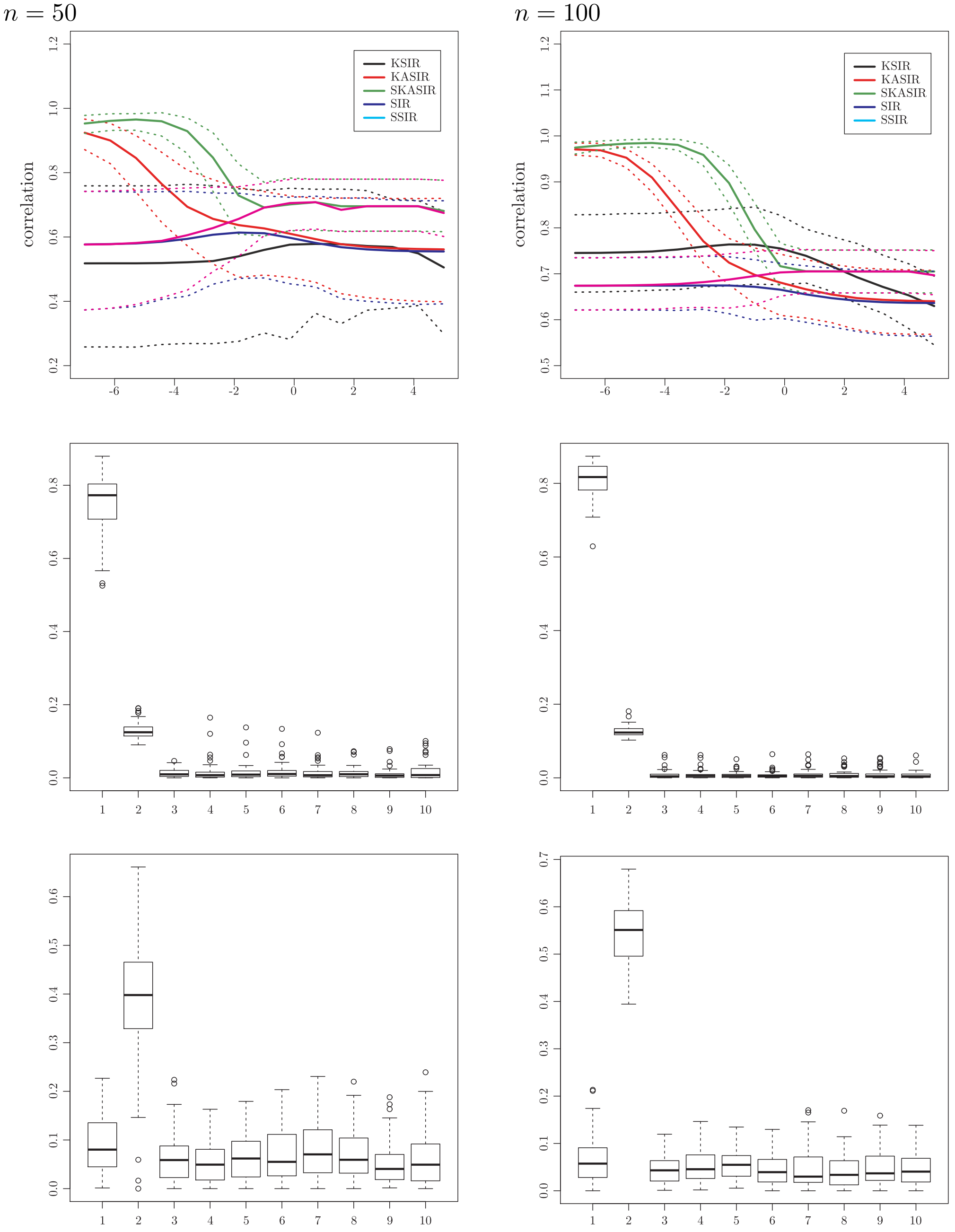

Figure 3. For Example 2, the Spearman correlation on test data for the five methods (KSIR, KASIR, SKASIR, SIR, SSIR) is shown in the first row. The second (third) row shows the values of d_j for SKASIR (SSIR).

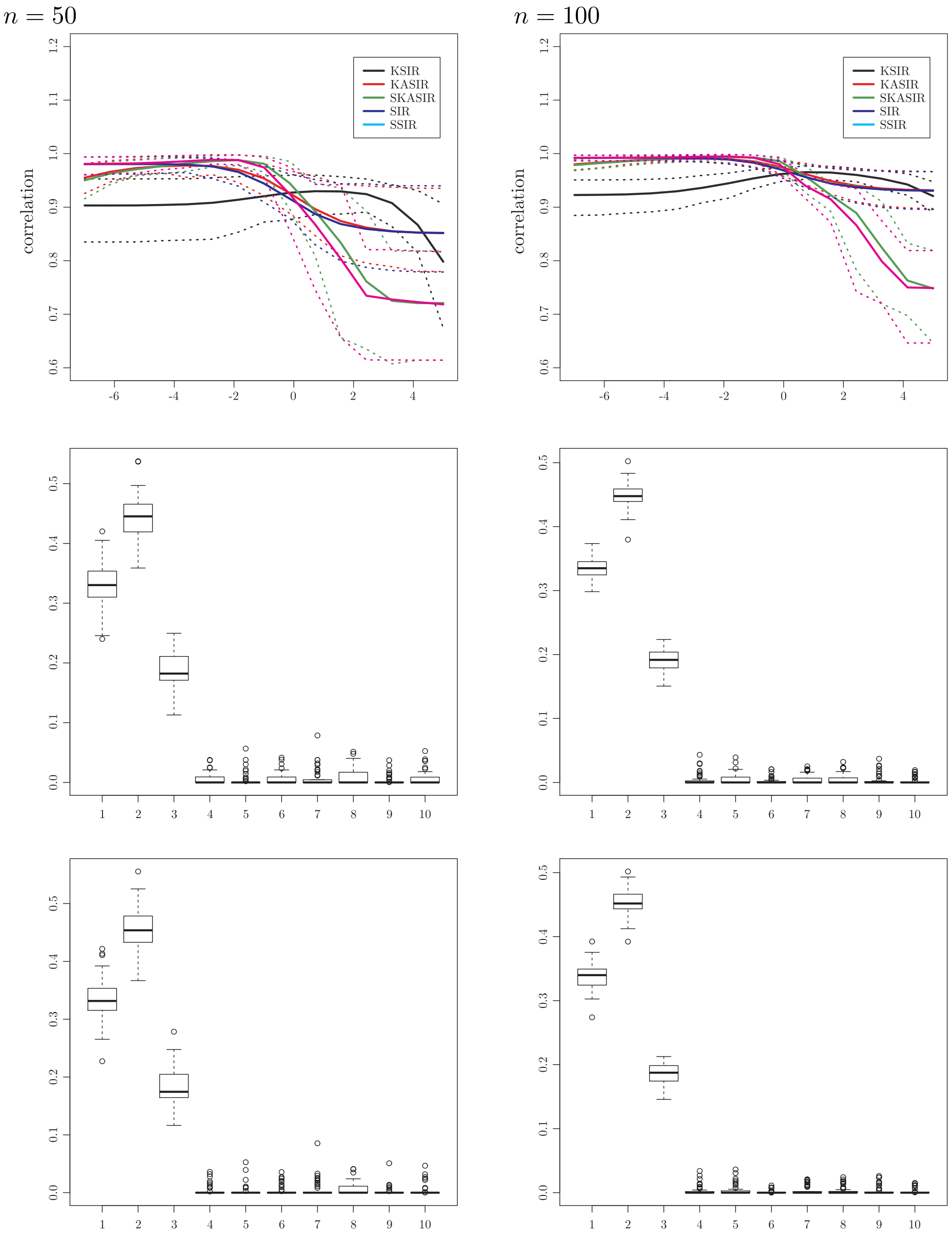

Figure 4. For Example 3, the Spearman correlation on test data for the five methods (KSIR, KASIR, SKASIR, SIR, SSIR) is shown in the first row. The second (third) row shows the values of d_j for SKASIR (SSIR).

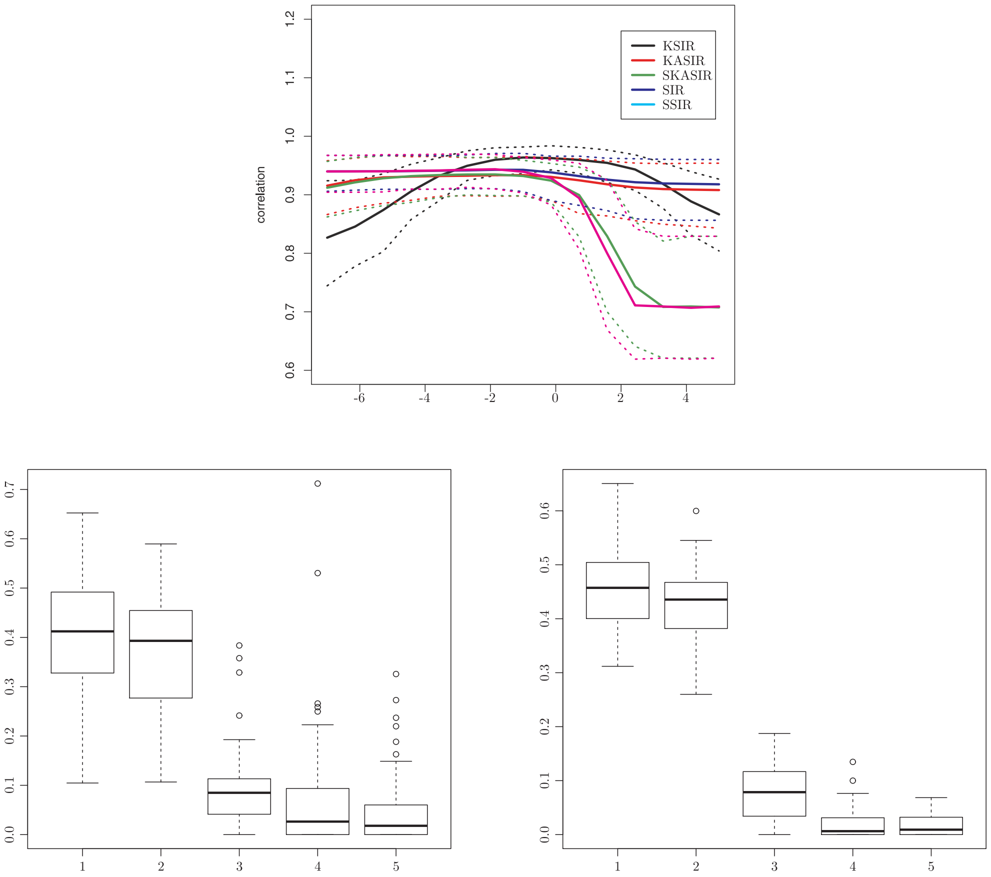

Figure 5. For Example 1 with $n = 50$ and $p = 5$, the Spearman correlation on test data for the five methods (KSIR, KASIR, SKASIR, SIR, SSIR) is shown in the first row. The left (right) panel on the second row shows the values of d_j for SKASIR (SSIR).

Table 1. Prediction MSE for the five SDR methods for the NMMAPS data.

KSIR	KASIR	SKASIR	SIR	SSIR
0.591	0.592	0.595	0.608	0.611

estimated component functions from the first dimension reduction direction for one of the 100 runs in Figure 8.

5. Conclusion and Discussion

In this paper we considered kernel additive sliced inverse regression and its sparse version that can perform variable selection. The advantages of the additive structure come from the fast eigenvalue decay rate for the additive kernel

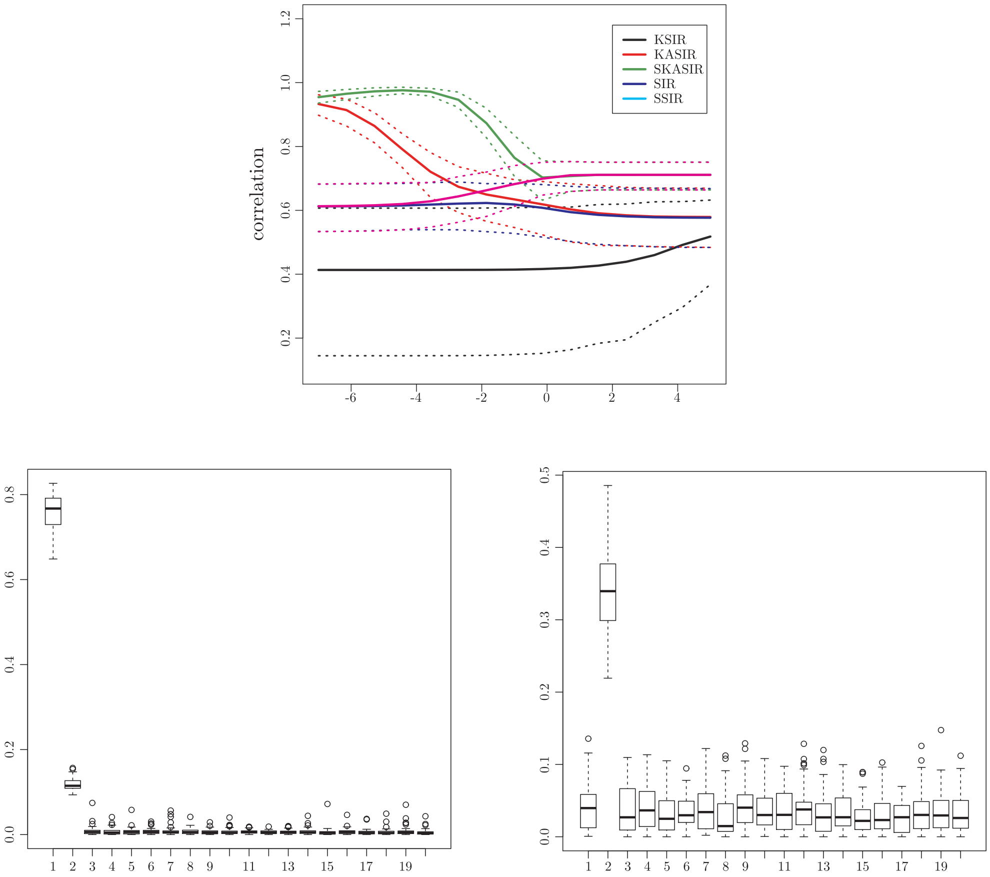

Figure 6. For Example 2 with $n = 100$ and $p = 20$, the Spearman correlation on test data for the five methods (KSIR, KASIR, SKASIR, SIR, SSIR) is shown in the first row. The left (right) panel on the second row shows the values of d_j for SKASIR (SSIR).

compared to general kernels in multi-dimensional case. We showed via numerical studies that KASIR is flexible, parsimonious and reliable and SKASIR can further identify important predictors, a goal that fully nonparametric KSIR method cannot achieve. Although we do not consider a partially linear structure, this is straightforward by just using a linear kernel in the additive combination of kernels whenever the predictor is in the linear part. In particular, in this way we can deal with both continuous and discrete predictors simultaneously.

Due to the necessity of using general-purpose nonlinear optimization software, the computation of SKASIR may be too slow and unstable for high dimensional problems, and in particular we cannot carry out simulations with p much larger than 20 in R, due to computational speed constraints. New formu-

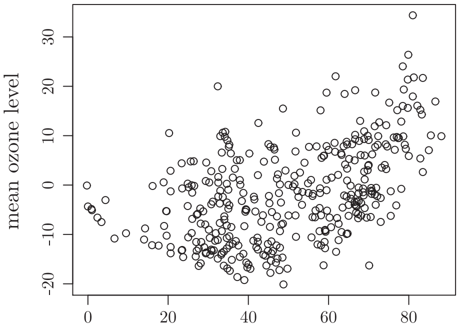

Figure 7. Scatterplot for the NMMAPS data with the predictor being mean temperature.

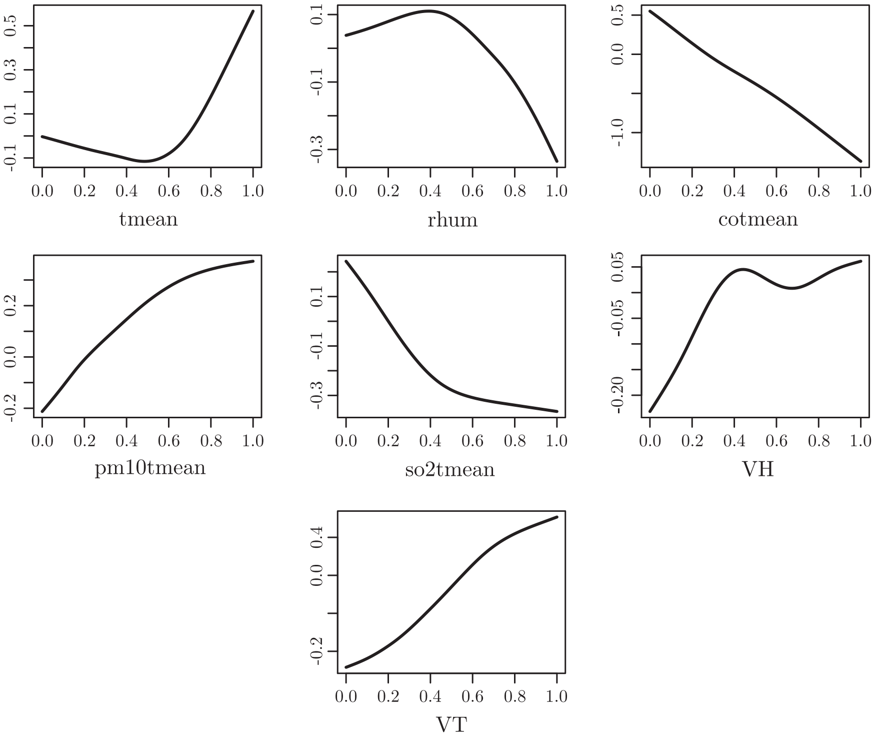

Figure 8. Estimated component functions from the first dimension reduction direction for one of the 100 runs.

lations or algorithms need to be proposed for it to work in problems in higher dimensions. In our formulation, given $\mathbf{C}$, the optimization problem for d_j can actually be posed as a quadratically constrained quadratic programming (QCQP), which is well-known to be NP-hard in general. Although previously some QCQP problems can be solved by semidefinite programming after relaxation, the nonconvexity of our formulation prevents any easy route for doing so and we have not been successful in this direction so far. Asymptotic properties of the sparse estimator are not established here due to the technical challenges, and are worth further investigation.

In the simulations, it is seen that the estimated weights are typically not sparse enough and although some weights are small, they are not always sufficiently small to be treated as zero. One could replace the constraint $\sum_j d_j = 1$ by $\sum_j d_j^q = 1$ with $0 < q < 1$, potentially leading to a sparser solution, as has been demonstrated in regression (Huang, Horowitz, and Ma (2008)).

We here only used the one-dimensional Sobolev kernel as the building block of product and additive kernels. Kernel choice is a challenging topic in itself, and it is hard to tune the kernel for different data sets. For kernel methods, the Gaussian kernel is popular and gives satisfactory performance in various problems. However, we have not used the Gaussian kernel: the choice of the bandwidth parameter in the Gaussian kernel is critical for its performance and it is not clear how to choose these parameters in an efficient way. Even though some bandwidth selection methods could be used, this extra complication disturbs the comparison between different methods. With the Sobolev kernel, there are no hyperparameters to choose. Careful treatment for kernels with hyperparameters needs further investigation. A related problem is automatic kernel choice. Selection/combination of different kernels is an interesting direction for future research.

A general and elegant theory of nonlinear SDR is reported in Lee, Li, and Chiaromonte (2013). By using the definition of a SDR σ-field to replace the concept of SDR space, the linear design condition is not necessary for nonlinear SDR, and they proposed generalized SIR. As they show, even without the linear design condition, KSIR can still be used to estimate the SDR σ-field. However, it seems challenging to establish asymptotic theory with this formulation.

The problem of determining the number of indices in the SDR space is also important. In data, the smallest PMSE is usually obtained when $r = 2$ suggesting that $r = 2$ may be appropriate. However, this is certainly problematic unless the goal is mean prediction. It is worthwhile to investigate the extension of other SDR method, such as SAVE (Cook and Weisberg (1991); Cook (2000); Zhu and Zhu (2007); Zhu, Zhu, and Feng (2010); Dong and Li (2010)), using product or additive kernels.

Supplementary Materials

The supplementary material online for this paper contains the proof of Theorem 1.

Acknowledgements

The authors sincerely thank Professor Hsin-Cheng Huang, an AE and two reviewers for their insightful comments that have led to significant improvements of the manuscript.

References

Aronszajn, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc* **68**, 337-404.

Bartlett, P. L. and Mendelson, S. (2003). Rademacher and Gaussian complexities: Risk bounds and structural results. *J. Machine Learn. Res.* **3**, 463-482.

Blanchard, G., Bousquet, O. and Massart, P. (2008). Statistical performance of support vector machines. *Ann. Statist.* **36**, 489-531.

Bondell, H. D. and Li, L. (2009). Shrinkage inverse regression estimation for model-free variable selection. *J. Roy. Statist. Soc. Ser. B* **71**, 287-299.

Caponnetto, A. and De Vito, E. (2007). Optimal rates for the regularized least-squares algorithm. *Found. Comput. Math.* **7**, 331-368.

Chen, X., Zou, C. and Cook, R. D. (2010). Coordinate-independent sparse sufficient dimension reduction and variable selection. *Ann. Statist.* **38**, 3696-3723.

Cook, R. D. (1994). On the interpretation of regression plots. *J. Amer. Statist. Assoc.* **89**, 177-189.

Cook, R. D. (1996). Graphics for regressions with a binary response. *J. Amer. Statist. Assoc.* **91**, 983-992.

Cook, R. D. (2000). SAVE: A method for dimension reduction and graphics in regression. *Comm. Statist. Theory Methods* **29**, 2109-2121.

Cook, R. D. and Ni, L. (2005). Sufficient dimension reduction via inverse regression. *J. Amer. Statist. Assoc.* **100**, 410-428.

Cook, R. D. and Weisberg, S. (1991). Sliced inverse regression for dimension reduction: Comment. *J. Amer. Statist. Assoc.* **86**, 328-332.

Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions Through Graphics*. Wiley, New York.

Dong, Y. and Li, B. (2010). Dimension reduction for non-elliptically distributed predictors: second-order methods. *Biometrika* **97**, 279-294.

Fukumizu, K., Bach, F. R. and Jordan, M. (2004). Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *J. Machine Learn. Res.* **5**, 73-99.

Fukumizu, K., Bach, F. R. and Jordan, M. (2009). Kernel dimension reduction in regression. *Ann. Statist.* **37**, 1871-1905.

Fukumizu, K. and Leng, C. (2014). Gradient based kernel dimension reduction for regression. *J. Amer. Statist. Assoc.* **109**, 359-370.

Huang, J., and Horowitz, J. L. and Ma, S. G. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist.* **36**, 587-613.

Kato, T. (1995). *Perturbation Theory for Linear Operators.* Springer Verlag.

Koltchinskii, V. and Yuan, M. (2010). Sparsity in multiple kernel learning. *Ann. Statist.* **38**, 3660-3695.

Kühn, T. (1987). Eigenvalues of integral operators with smooth positive definite kernels. *Archiv der Mathematik* **49**, 525-534.

Lee, K.-Y., Li, B. and Chiaromonte, F. (2013). A general theory for nonlinear sufficient dimension reduction: Formulation and estimation. *Ann. Statist.* **41**, 221-249.

Li, K. C. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.* **86**, 316-327.

Li, L. (2007). Sparse sufficient dimension reduction. *Biometrika* **94**, 603-613.

Li, L. and Nachtsheim, C. J. (2006). Sparse sliced inverse regression. *Technometrics* **48**, 503-510.

Liang, H., Thurston, S. W., Ruppert, D., Apanasovich, T. and Hauser, R. (2008). Additive partial linear models with measurement errors. *Biometrika* **95**, 667-678.

Lin, Y. and Zhang, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. *Ann. Statist.* **34**, 2272-2297.

Ma, S. (2012). Two-step spline estimating equations for generalized additive partially linear models with large cluster sizes. *Ann. Statist.* **40**, 2943-2972.

Raskutti, G., Wainwright, M. J. and Yu, B. (2012). Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *J. Machine Learn. Res.* **13**, 389-427.

Stone, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13**, 689-705.

Storlie, C. B., Bondell, H. D., Reich, B. J. and Zhang, H. H. (2011). Surface estimation, variable selection, and the nonparametric oracle property. *Statist. Sinica* **21**, 679-705.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

Wang, L., Liu, X., Liang, H. and Carroll, R. J. (2011). Estimation and variable selection for generalized additive partial linear models. *Ann. Statist.* **39**, 1827-1851.

Wu, H.-M. (2008). Kernel sliced inverse regression with applications to classification. *J. Comput. Graph. Statist.* **17**, 590-610.

Wu, Q., Liang, F. and Mukherjee, S. (2013). Kernel sliced inverse regression: regularization and consistency. *Abstract and Applied Analysis.*

Xue, L. (2009). Consistent variable selection in additive models. *Statist. Sinica* **19**, 1281-1296.

Yeh, Y.-R., Huang, S.-Y. and Lee, Y.-J. (2009). Nonlinear dimension reduction with kernel sliced inverse regression. *IEEE Trans. Knowledge and Data Engineering* **21**, 1590-1603.

Yin, X., Li, B. and Cook, R. D. (2008). Successive direction extraction for estimating the central subspace in a multiple-index regression. *J. Multivariate Anal.* **99**, 1733-1757.

Zhang, H. H. (2006). Variable selection for support vector machines via smoothing spline ANOVA. *Statist. Sinica* **16**, 659-674.

Zhu, L.P., Zhu, L.X., and Feng, Z.H (2010). Dimension reduction in regressions through cumulative slicing estimation. *J. Amer. Statist. Assoc.* **105**, 1455-1466.

Zhu, L.-P. and Zhu, L.-X. (2007). On kernel method for sliced average variance estimation. *J. Multivariate Anal.* **98**, 970-991.

School of Mathematics and Statistics, University of New South Wales, Sydney Australia 2052.
E-mail: heng.lian@unsw.edu.au

Department of Statistical Sciences & Operations Research, Virginia Commonwealth University, Richmond, VA 23284, USA.

E-mail: qwang3@vcu.edu

(Received December 2014; accepted May 2015)