Hindawi Publishing Corporation
Abstract and Applied Analysis
Volume 2013, Article ID 540725, 11 pages
http://dx.doi.org/10.1155/2013/540725

Research Article

Kernel Sliced Inverse Regression: Regularization and Consistency

Qiang Wu,[1] Feng Liang,[2] and Sayan Mukherjee[3]

[1] *Department of Mathematical Sciences, Middle Tennessee State University, Murfreesboro, TN 37130, USA*
[2] *Department of Statistics, University of Illinois at Urbana-Champaign, Urbana, IL 61820, USA*
[3] *Departments of Statistical Science, Mathematics, and Computer Science, Institute for Genome Sciences & Policy, Duke University, Durham, NC 27708, USA*

Correspondence should be addressed to Qiang Wu; qwu@mtsu.edu

Received 3 May 2013; Accepted 14 June 2013

Academic Editor: Yiming Ying

Kernel sliced inverse regression (KSIR) is a natural framework for nonlinear dimension reduction using the mapping induced by kernels. However, there are numeric, algorithmic, and conceptual subtleties in making the method robust and consistent. We apply two types of regularization in this framework to address computational stability and generalization performance. We also provide an interpretation of the algorithm and prove consistency. The utility of this approach is illustrated on simulated and real data.

1. Introduction

The goal of dimension reduction in the standard regression/classification setting is to summarize the information in the p-dimensional predictor variable X relevant to predicting the univariate response variable Y. The summary $S(X)$ should have $d \ll p$ variates and ideally should satisfy the following conditional independence property:

$$Y \perp\!\!\!\perp X \mid S(X). \tag{1}$$

Thus, any inference of Y involves only the summary statistic $S(X)$ which is of much lower dimension than the original data X.

Linear methods for dimension reduction focus on linear summaries of the data, $S(X) = (\beta_1^T X, \ldots, \beta_d^T X)$. The d-dimensional subspace, $\mathcal{S} = \operatorname{span}(\beta_1, \ldots, \beta_d)$, is defined as the effective dimension reduction (e.d.r.) space in [1], since $\mathcal{S}$ summarizes all the predictive information on Y. A key result in [1] is that under some mild conditions, the e.d.r. directions $\{\beta_j\}_{j=1}^d$ correspond to the eigenvectors of the matrix:

$$T = [\operatorname{cov}(X)]^{-1} \operatorname{cov}[\mathbb{E}(X \mid Y)]. \tag{2}$$

Thus, the e.d.r. directions or subspace can be estimated via an eigenanalysis of the matrix T, which is the foundation of the sliced inverse regression (SIR) algorithm proposed in [1, 2]. Further developments include sliced average variance estimation (SAVE) [3] and Principal Hessian directions (PHDs) [4]. The aforementioned algorithms cannot be applied in high-dimensional settings, where the number of covariates p is greater than the number of observations n, since the sample covariance matrix is singular. Recently, an extension of SIR has been proposed in [5], which can handle the case for $p > n$ based on the idea of partial least squares.

A common premise held in high-dimensional data analysis is that the intrinsic structure of data is in fact low dimensional, for example, the data is concentrated on a manifold. Linear methods such as SIR often fail to capture this nonlinear low-dimensional structure. However, there may exist a nonlinear embedding of the data into a Hilbert space, where a linear method can capture the low-dimensional structure. The basic idea in applying kernel methods is the application of a linear algorithm to the data mapped into a feature space induced by a kernel function. If projections onto this low-dimensional structure can be computed by inner products in this Hilbert space, the so-called kernel trick [6, 7] can be used to obtain simple and efficient algorithms. Since the embedding is nonlinear, linear directions in the feature space correspond to nonlinear directions in the original data

space. Nonlinear extensions of some classical linear dimensional reduction methods using this approach include kernel principle component analysis [6], kernel Fisher discriminant analysis [8], and kernel independent correlation analysis [9]. This idea was applied to SIR in [10, 11] resulting in the kernel sliced inverse regression (KSIR) method which allows for the estimation of nonlinear e.d.r. directions.

There are numeric, algorithmic, and conceptual subtleties to a direct application of this kernel idea to SIR, although it looks quite natural at first glance. In KSIR, the p-dimensional data are projected into a Hilbert space $\mathcal{H}$ through a feature map: $\phi : X \to \mathcal{H}$ and the nonlinear features are supposed to be recovered by the eigenfunctions of the following operator

$$T = \left[\operatorname{cov}(\phi(X))\right]^{-1} \operatorname{cov}\left[\mathbb{E}\left(\phi\left(X\right) \mid Y\right)\right]. \tag{3}$$

However, this operator T is actually not well defined in general, especially when $\mathcal{H}$ is infinite dimensional and the covariance operator $\operatorname{cov}(\phi(X))$ is not invertible. In addition, the key utility of representation theorems in kernel methods is that optimization in a possibly infinite dimensional Hilbert space $\mathcal{H}$ reduces to solving a finite dimensional optimization problem. In the KSIR algorithm developed in [10, 11], the representer theorem has been implicitly used and seems to work well in empirical studies. However, it is not theoretically justifiable since T is estimated empirically via observed data. Moreover, the computation of eigenfunctions of an empirical estimate of T from observations is often ill-conditioned and results in computational instability. In [11], a low rank approximation of the kernel matrix is used to overcome this instability and to reduce computational complexity. In this paper, our aim is to clarify the theoretical subtleties in KSIR and to motivate two types of regularization schemes to overcome the computational difficulties arising in KSIR. The consistency is proven and practical advantages of regularization are demonstrated via empirical experiments.

2. Mercer Kernels and Nonlinear e.d.r. Directions

The extension of SIR to use kernels is based on properties of reproducing kernel Hilbert spaces (RKHSs) and in particular Mercer kernels [12].

Given predictor variables $X \in \mathcal{X} \subseteq \mathbb{R}^p$, a Mercer kernel is a continuous, positive, and semidefinite function $k(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with the following spectral decomposition:

$$k\left(x,z\right) = \sum_j \lambda_j \phi_j\left(x\right)\phi_j\left(z\right), \tag{4}$$

where $\{\phi_j\}$ are the eigenfunctions and $\{\lambda_j\}$ are the corresponding nonnegative, nonincreasing eigenvalues. An important property of Mercer kernels is that each kernel k uniquely corresponds to an RKHS as follows:

$$\mathcal{H} = \left\{ f \mid f\left(x\right) = \sum_{j\in\Lambda} a_j \phi_j\left(x\right) \text{ with } \sum_{j\in\Lambda} \frac{a_j^2}{\lambda_j} < \infty \right\}, \tag{5}$$

where the cardinality of $\Lambda := \{j : \lambda_j > 0\}$ is the dimension of the RKHS which may be infinite [12, 13].

Given a Mercer kernel, there exists a unique map or embedding ϕ from $\mathcal{X}$ to a Hilbert space defined by the eigenvalues and eigenfunctions of the kernel. The map takes the following form:

$$\phi\left(x\right) = \left(\sqrt{\lambda_1}\phi_1\left(x\right), \sqrt{\lambda_2}\phi_2\left(x\right), \ldots, \sqrt{\lambda_{|\Lambda|}}\phi_{|\Lambda|}\left(x\right)\right). \tag{6}$$

The Hilbert space induced by this map with the standard inner product $k(x,z) = \langle\phi(x),\phi(z)\rangle$ is isomorphic to the RKHS (5), and we will denote both Hilbert spaces as $\mathcal{H}$ [13]. In the case where k is infinite dimensional, $\phi : \mathcal{X} \to \ell_2$.

The random variable $X \in \mathcal{X}$ induces a random element $\phi(X)$ in the RKHS. Throughout this paper, we will use Hilbert space valued random variables; so we now recall some basic facts. Let Z be a random element in $\mathcal{H}$ with $\mathbb{E}\|Z\| < \infty$, where $\|\cdot\|$ denotes the norm in $\mathcal{H}$ induced by its inner product $\langle\cdot,\cdot\rangle$. The expectation $\mathbb{E}(Z)$ is defined to be an element in $\mathcal{H}$, satisfying $\langle a, \mathbb{E}(Z)\rangle = \mathbb{E}\langle a, Z\rangle$, for all $a \in \mathcal{H}$. If $\mathbb{E}\|Z\|^2 \le \infty$, then the covariance operator of Z is defined as $\mathbb{E}[(Z - \mathbb{E}Z) \otimes (Z - \mathbb{E}Z)]$, where

$$(a \otimes b)\, f = \left\langle b, f\right\rangle a, \quad \text{for any } f \in \mathcal{H}. \tag{7}$$

Let $\mathcal{P}$ denote the measure for random variable X. Throughout, we assume the following conditions.

Assumption 1.

(1) For all $x \in \mathcal{X}$, $k(x,\cdot)$ is $\mathcal{P}$-measurable.

(2) There exists $M > 0$ such that $x \in \mathcal{X}$, $k(X,X) \le M$ (a.s.) with respect to $\mathcal{P}$.

Under Assumption 1, the random element $\phi(X)$ has a well-defined mean and a covariance operator because $\|\phi(x)\|^2 = k(x,x)$ is bounded (a.s.). Without loss of generality, we assume $\mathbb{E}\phi(X) = 0$, where 0 is the zero element in $\mathcal{H}$. The boundedness also implies that the covariance operator $\Sigma = \mathbb{E}[\phi(X) \otimes \phi(X)]$ is compact and has the following spectral decomposition:

$$\Sigma = \sum_{i=1}^{\infty} w_i e_i \otimes e_i, \tag{8}$$

where w_i and $e_i \in \mathcal{H}$ are the eigenvalues and eigenfunctions, respectively.

We assume the following model for the relationship between Y and X:

$$Y = F\left(\left\langle\beta_1, \phi\left(X\right)\right\rangle, \ldots, \left\langle\beta_d, \phi\left(X\right)\right\rangle, \varepsilon\right), \tag{9}$$

with $\beta_j \in \mathcal{H}$ and the distribution of ε is independent of X. This model implies that the response variable Y depends on X only through a d-dimensional summary statistic

$$S\left(X\right) = \left(\left\langle\beta_1, \phi\left(X\right)\right\rangle, \ldots, \left\langle\beta_d, \phi\left(X\right)\right\rangle\right). \tag{10}$$

Although $S(X)$ is a linear summary statistic in $\mathcal{H}$, it extracts nonlinear features in the space of the original predictor

variables X. We call $\{\beta_j\}_{j=1}^d$ the nonlinear e.d.r. directions, and $\mathcal{S} = \text{span}(\beta_1, \ldots, \beta_d)$ the nonlinear e.d.r. space. The following proposition [10] extends the theoretical foundation of SIR to this nonlinear setting.

Proposition 2. *Assume the following linear design condition for $\mathcal{H}$ that for any $f \in \mathcal{H}$, there exists a vector $b \in \mathbb{R}^d$ such that*

$$\mathbb{E}\left[\langle f, \phi(X)\rangle \mid S(X)\right] = b^T S(X), \quad \text{with } S(X) = \left(\langle \beta_1, \phi(X)\rangle, \ldots, \langle \beta_d, \phi(X)\rangle\right)^T. \tag{11}$$

Then, for the model specified in (9)*, the inverse regression curve $\mathbb{E}[\phi(X) \mid Y]$ is contained in the span of $(\Sigma\beta_1, \ldots, \Sigma\beta_d)$, where Σ is the covariance operator of $\phi(X)$.*

Proposition 2 is a straightforward extension of the multivariate case in [1] to a Hilbert space or a direct application of the functional SIR setting in [14]. Although the linear design condition (11) may be difficult to check in practice, it has been shown that such a condition usually holds approximately in a high-dimensional space [15]. This conforms to the argument in [11] that the linearity in a reproducing kernel Hilbert space is less strict than the linearity in the Euclidean space. Moreover, it is pointed out in [1, 10] that even if the linear design condition is violated, the bias of SIR variate is usually not large.

An immediate consequence of Proposition 2 is that nonlinear e.d.r. directions are the eigenvectors corresponding to the largest eigenvalues of the following generalized eigendecomposition problem:

$$\Gamma\beta = \lambda\Sigma\beta, \quad \text{where } \Sigma = \text{cov}\left[\phi(X)\right], \quad \Gamma = \text{cov}\left[\mathbb{E}\left(\phi(X) \mid Y\right)\right], \tag{12}$$

or equivalently from an eigenanalysis of the operator $T = \Sigma^{-1}\Gamma$. In the infinite dimensional case, a technical difficulty arises since the operator

$$\Sigma^{-1} = \sum_{i=1}^{\infty} w_i^{-1} e_i \otimes e_i \tag{13}$$

is not defined on the entire Hilbert space $\mathcal{H}$. So for the operator T to be well defined, we need to show that the range of Γ is indeed in the range of Σ. A similar issue also arose in the analysis of dimension reduction and canonical analysis for functional data [16, 17]. In these analyses, extra conditions are needed for operators like T to be well defined. In KSIR, this issue is resolved automatically by the linear design condition and extra conditions are not required as stated by the following Theorem; see Appendix A for the proof.

Theorem 3. *Under Assumption 1 and the linear design condition* (11)*, the following hold:*

(i) *the operator Γ is of finite rank $d_\Gamma \le d$. Consequently, it is compact and has the following spectral decomposition:*

$$\Gamma = \sum_{i=1}^{d_\Gamma} \tau_i u_i \otimes u_i, \tag{14}$$

where τ_i and u_i are the eigenvalues and eigenvectors, respectively. Moreover, $u_i \in \text{range}(\Sigma)$ for all $i = 1, \ldots, d_\Gamma$;

(ii) *the eigendecomposition problem* (12) *is equivalent to the eigenanalysis of the operator T, which takes the following form*

$$T = \sum_{i=1}^{d_\Gamma} \tau_i u_i \otimes \Sigma^{-1}(u_i). \tag{15}$$

3. Regularized Kernel Sliced Inverse Regression

The discussion in Section 2 implies that nonlinear e.d.r. directions can be retrieved by applying the original SIR algorithm in the feature space induced by the Mercer kernel. There are some computational challenges to this idea such as estimating an infinite dimensional covariance operator and the fact that the feature map is often difficult or impossible to compute for many kernels. We address these issues by working with inner products of the feature map and adding a regularization term to kernel SIR.

3.1. Estimating the Nonlinear e.d.r. Directions. Given n observations $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, our objective is to obtain an estimate of the e.d.r. directions $(\widehat{\beta}_1, \ldots, \widehat{\beta}_d)$. We first formulate a procedure almost identical to the standard SIR procedure except that it operates in the feature space $\mathcal{H}$. This highlights the immediate relation between the SIR and KSIR procedures.

(1) Without loss of generality, we assume that the mapped predictor variables are mean zero, that is, $\sum_{i=1}^n \phi(x_i) = 0$, for otherwise we can subtract $\overline{\phi} = (1/n)\sum_{i=1}^n \phi(x_i)$ from $\phi(x_i)$. The sample covariance is estimated by

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \phi(x_i) \otimes \phi(x_i). \tag{16}$$

(2) Bin the Y variables into H slices $G_1, \ldots, G_H$ and compute mean vectors of the corresponding mapped predictor variables for each slice

$$\psi_h = \frac{1}{n_h}\sum_{i \in G_h} \phi(x_i), \quad h = 1, \ldots, H. \tag{17}$$

Compute the sample between-group covariance matrix

$$\widehat{\Gamma} = \sum_{h=1}^{H} \frac{n_h}{n}\, \psi_h \otimes \psi_h. \tag{18}$$

(3) Estimate the SIR directions $\widehat{\beta}_j$ by solving the generalized eigendecomposition problem:

$$\widehat{\Gamma}\beta = \lambda\widehat{\Sigma}\beta. \tag{19}$$

This procedure is computationally impossible if the RKHS is infinite dimensional or the feature map cannot be computed (which is the usual case). However, the model given in (9) requires not the e.d.r. directions but the projection onto these directions, that is, the d summary statistics

$$v_1 = \langle \beta_1, \phi(x) \rangle, \ldots, v_d = \langle \beta_d, \phi(x) \rangle, \tag{20}$$

which we call the KSIR variates. Like other kernel algorithms, the kernel trick enables KSIR variates to be efficiently computed from only the kernel k, not the map ϕ.

The key quantity in this alternative formulation is the centred Gram matrix K defined by the kernel function $k(\cdot,\cdot)$, where

$$K_{ij} = \left\langle \phi(x_i), \phi\left(x_j\right) \right\rangle = k\left(x_i, x_j\right). \tag{21}$$

Note that the rank of K is less than n; so K is always singular.

Given the centered Gram matrix K, the following generalized eigendecomposition problem can be used to compute the KSIR variates:

$$KJKc = \lambda K^2 c, \tag{22}$$

where c denotes the n-dimensional generalized eigenvector and J denotes an $n \times n$ matrix with $J_{ij} = 1/n_m$ if i, j are in the mth group consisting of n_m observations and zero otherwise. The following proposition establishes the equivalence between two eigendecomposition problems, (22) and (19), in the recovery of KSIR variates $(v_1, \ldots, v_d)$.

Proposition 4. *Given the observations* $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, *let* $(\widehat{\beta}_1, \ldots, \widehat{\beta}_d)$ *and* $(\widehat{c}_1, \ldots, \widehat{c}_d)$ *denote the generalized eigenvectors of* (22) *and* (19), *respectively. Then, for any* $x \in \mathcal{X}$ *and* $j = 1, \ldots, d$, *the following holds:*

$$\begin{aligned} v_j &= \left\langle \widehat{\beta}_j, \phi(x) \right\rangle = \widehat{c}_j^T K_x, \\ K_x &= \left(k(x, x_1), \ldots, k(x, x_n)\right)^T, \end{aligned} \tag{23}$$

provided that $\widehat{\Sigma}$ *is invertible. When* $\widehat{\Sigma}$ *is not invertible, the conclusion holds modulo the null space of* $\widehat{\Sigma}$.

This result was proven in [10], in which the algorithm was further reduced to solving

$$JKc = \lambda Kc \tag{24}$$

by canceling K from both sides of (22).

Remark 5. It is important to remark that when $\widehat{\Sigma}$ is not invertible, Proposition 4 states that the equivalence between (19) and (22) holds modulo the null space of $\widehat{\Sigma}$ that is a requirement of the representer theorem, that is, $\widehat{\beta}_j$ are linear combinations of $\phi(x_i)$. Without this mandated condition, we will see that each eigenvector of (22) produces an eigenvector of (19), while the eigenvector of (19) is not necessarily recovered by (22).

Remark 6. It is necessary to clarify the difference between (12) and (19). In (12), it is natural to assume that β is orthogonal to the null space of Σ. To see this, let $\beta = \beta^0 + \beta^\star$ with β^0 belonging to the null space and $\beta^\star$ orthogonal to the null space. Then,

$$\mathbb{E}\left[\langle \beta, \phi(x) \rangle^2\right] = \langle \beta, \Sigma\beta \rangle = \langle \beta^\star, \Sigma\beta^\star \rangle = \mathbb{E}\left[\langle \beta^\star, \phi(x) \rangle^2\right], \tag{25}$$

that is, $\langle \beta, \phi(x) \rangle = \langle \beta^\star, \phi(x) \rangle$ (a.s.). It means that β_0 does not contribute to the KSIR variates and thus could be set as 0. However, in (19), if $\widehat{\beta}$ is an eigenvector and $\widehat{\beta}^\star$ is its orthogonal component relative to the null space of $\widehat{\Sigma}$, the identity $\langle \widehat{\beta}, \phi(x) \rangle = \langle \widehat{\beta}^\star, \phi(x) \rangle$ is in general not true for a new point x which is different from the observations. Thus, from a theoretical perspective, it is not as natural to assume the representer theorem, although it works well in practice. In this sense, the KSIR algorithm based on (22) or (24) does not have a thorough mathematical foundation.

3.2. Regularization and Stability. Except for the theoretical subtleties, in applications with relatively small samples, the eigendecomposition in (22) is often ill-conditioned resulting in overfitting as well as numerically unstable estimates of the e.d.r. space. This can be addressed by either thresholding eigenvalues of the estimated covariance matrix $\widehat{\Sigma}$ or by adding a regularization term to (22) or (24).

We motivate two types of regularization schemes. The first one is the traditional ridge regularization. It is used in both linear SIR and functional SIR [18–20], which solves the eigendecomposition problem

$$\widehat{\Gamma}\beta = \lambda\left(\widehat{\Sigma} + sI\right)\beta. \tag{26}$$

Here, and in the sequel, I denotes the identity operator and s is a tuning parameter. Assuming the representer theorem, its kernel form is given as

$$JKc = \lambda(K + nsI)c. \tag{27}$$

Another type of regularization is to regularize (22) directly:

$$KJKc = \lambda\left(K^2 + n^2 sI\right)c. \tag{28}$$

The following proposition, whose proof is in Appendix B, states that solving the generalized eigendecomposition problem (28) is equivalent to finding the eigenvectors of

$$\left(\widehat{\Sigma}^2 + sI\right)^{-1}\widehat{\Sigma}\widehat{\Gamma}. \tag{29}$$

Proposition 7. *Let* $\widehat{c}_j$ *be the eigenvectors of* (28), *and let* $\widehat{\beta}_j$ *be the eigenvectors of* (29). *Then, the following holds for the regularized KSIR variates:*

$$\widehat{v}_j = \widehat{c}_j^T\left[k(x, x_1), \ldots, k(x, x_n)\right] = \left\langle \widehat{\beta}_j, \phi(x) \right\rangle. \tag{30}$$

This algorithm is termed as the Tikhonov regularization. For linear SIR, it is shown in [21] that the Tikhonov regularization is more efficient than the ridge regularization.

Except for the computational stability, regularization also makes the matrix forms of KSIR, (27) and (28), interpretable by justifiable representer theorem.

Proposition 8. *For both ridge and the Tikhonov regularization scheme of KSIR, the eigenfunctions $\widehat{\beta}_j$ are linear combinations of $\phi(x_i)$, $i = 1, \ldots, n$.*

The conclusion follows from the observation that $\widehat{\beta}_j = (1/\lambda_j)(\widehat{\Gamma}\widehat{\beta}_j - \widehat{\Sigma}\widehat{\beta}_j)$ for the ridge regularization and $\widehat{\beta}_j = (1/\lambda_j)(\widehat{\Sigma}\widehat{\Gamma}\widehat{\beta}_j - \widehat{\Sigma}^2\widehat{\beta}_j)$ for the Tikhonov regularization.

To close, we remark that KSIR is computationally advantageous even for the case of linear models when $p \gg n$ due to the fact that the eigendecomposition problem is for $n \times n$ matrices rather than the $p \times p$ matrices in the standard SIR formulation.

3.3. Consistency of Regularized KSIR. In this subsection, we prove the asymptotic consistency of the e.d.r. directions estimated by regularized KSIR and provide conditions under which the rate of convergence is $O_p(n^{-1/4})$. An important observation from the proof is that the rate of convergence of the e.d.r. directions depends on the contribution of the small principal components. The rate can be arbitrarily slow if the e.d.r. space depends heavily on eigenvectors corresponding to small eigenvalues of the covariance operator.

Note that various consistency results are available for linear SIR [22–24]. These results hold only for the finite dimensional setting and cannot be adapted to KSIR where the RKHS is often infinite dimensional. Consistency of functional SIR has also been studied before. In [14], a thresholding method is considered, which selects a finite number of eigenvectors and uses results from finite rank operators. Their proof of consistency requires stronger and more complicated conditions than ours. The consistency for functional SIR with ridge regularization is proven in [19], but it is of a weaker form than our result. We remark that the consistency results for functional SIR can be improved using a similar argument in this paper.

In the following, we state the consistency results for the Tikhonov regularization. A similar result can be proved for the ridge regularization while the details are omitted.

Theorem 9. *Assume $\mathbb{E}k(X, X)^2 < \infty$, $\lim_{n\to\infty} s(n) = 0$ and $\lim_{n\to\infty} s\sqrt{n} = \infty$; then*

$$\left|\left\langle \widehat{\beta}_j, \phi(\cdot)\right\rangle - \left\langle \beta_j, \phi(\cdot)\right\rangle\right| = o_p(1), \quad j = 1, \ldots, d_\Gamma, \tag{31}$$

where d_Γ is the rank of Γ, $\langle \beta_j, \phi(\cdot)\rangle$ is the projection onto the jth e.d.r., and $\langle \widehat{\beta}_j, \phi(\cdot)\rangle$ is the projection onto the jth e.d.r. as estimated by regularized KSIR.

If the e.d.r. directions $\{\beta_j\}_{j=1}^{d_\Gamma}$ depend only on a finite number of eigenvectors of the covariance operator Σ, the rate of convergence is $O(n^{-1/4})$.

This theorem is a direct corollary of the following theorem which is proven in Appendix C.

Theorem 10. *Define the projection operator and its complement for each $N \geq 1$*

$$\Pi_N = \sum_{i=1}^{N} e_i \otimes e_i, \qquad \Pi_N^{\perp} = I - \Pi_N = \sum_{i=N+1}^{\infty} e_i \otimes e_i, \tag{32}$$

where $\{e_i\}_{i=1}^{\infty}$ are the eigenvectors of the covariance operator Σ as defined in (8), with the corresponding eigenvalues denoted by w_i.

Assume $\mathbb{E}k(X, X)^2 < \infty$. For each $N \geq 1$, the following holds:

$$\left\|(\widehat{\Sigma}^2 + sI)^{-1}\widehat{\Sigma}\widehat{\Gamma} - T\right\|_{HS} = O_p\left(\frac{1}{s\sqrt{n}}\right) + \sum_{j=1}^{d_\Gamma}\left(\frac{s}{w_N^2}\left\|\Pi_N(\widetilde{u}_j)\right\| + \left\|\Pi_N^{\perp}(\widetilde{u}_j)\right\|\right), \tag{33}$$

where $\widetilde{u}_j = \Sigma^{-1}u_j$ and $\{u_j\}_{i=1}^{d_\Gamma}$ are the eigenvectors of Γ as defined in (14) and $\|\cdot\|_{HS}$ denotes the Hilbert-Schmidt norm of a linear operator.

If $s = s(n)$ satisfy $s \to 0$ and $s\sqrt{n} \to \infty$ as $n \to \infty$, then

$$\left\|(\widehat{\Sigma}^2 + sI)^{-1}\widehat{\Sigma}\widehat{\Gamma} - T\right\|_{HS} = o_p(1). \tag{34}$$

4. Application to Simulated and Real Data

In this section, we compare regularized kernel sliced inverse regression (RKSIR) with several other SIR-related dimension reduction methods. The comparisons are used to address two questions: (1) does regularization improve the performance of kernel sliced inverse regression, and (2) does the nonlinearity of kernel sliced inverse regression improve the prediction accuracy?

We would like to remark that the assessment of nonlinear dimension reduction methods could be more difficult than that of linear ones. When the feature mapping ϕ for an RKHS is not available, we do not know the true e.d.r. directions or subspace. So in that case, we will use the prediction accuracy to evaluate the goodness of RKSIR.

4.1. Importance of Nonlinearity and Regularization. Our first example illustrates that both the nonlinearity and regularization of RKSIR can significantly improve prediction accuracy.

The regression model has ten predictor variables $X = (X_1, \ldots, X_{10})$ with each one following a normal distribution $X_i \sim N(0, 1)$. A univariate response is given as

$$Y = \left(\sin(X_1) + \sin(X_2)\right)\left(1 + \sin(X_3)\right) + \varepsilon \tag{35}$$

with noise $\varepsilon \sim N(0, 0.1^2)$. We compare the effectiveness of the linear dimension reduction methods SIR, SAVE, and PHD with RKSIR, by examining the predictive accuracy of a nonlinear kernel regression model on the reduced space. We generate 100 training samples and apply the above methods to

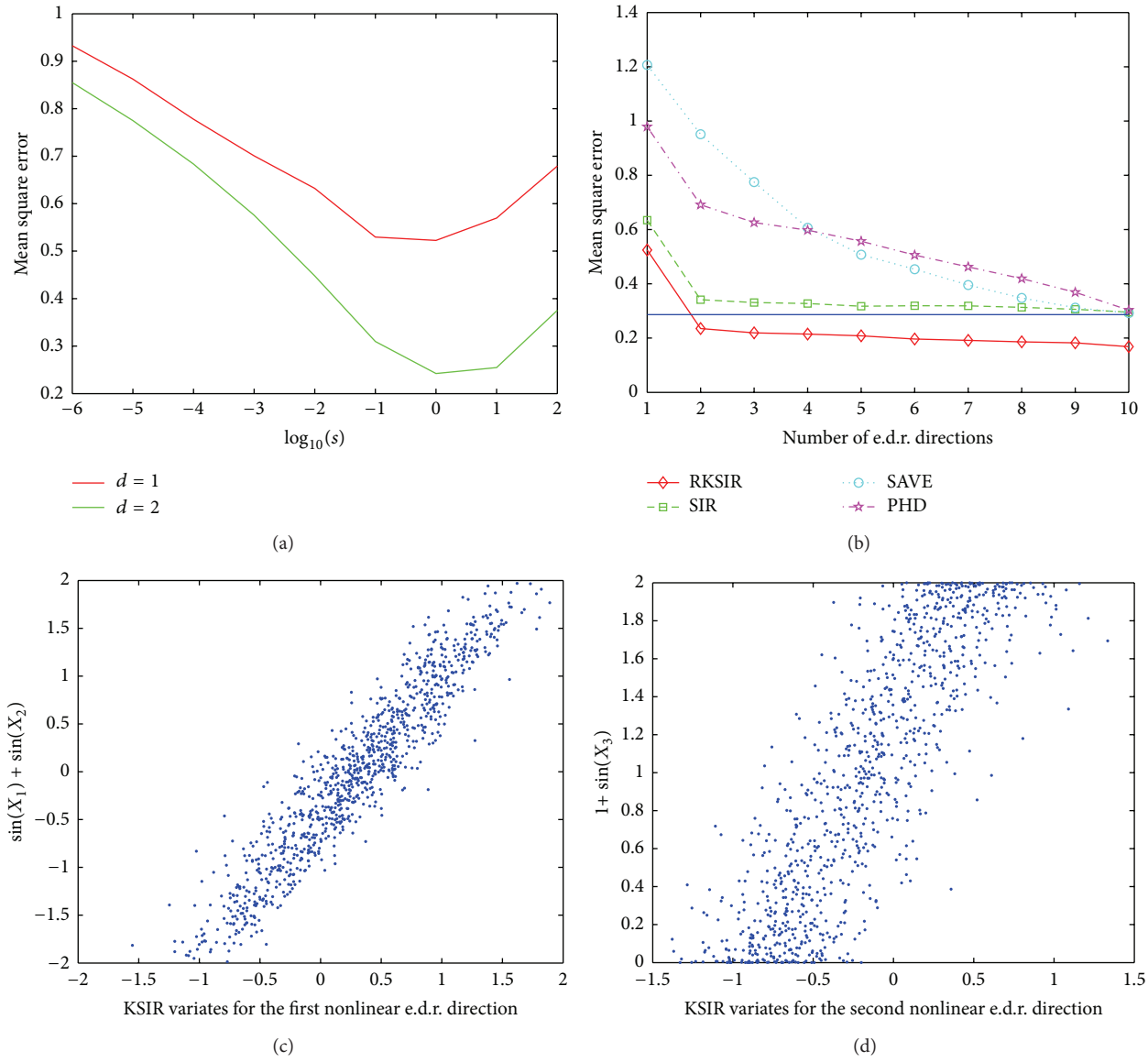

FIGURE 1: Dimension reduction for model (35): (a) mean square error for RKSIR with different regularization parameters; (b) mean square error for various dimension reduction methods. The blue line represents the mean square error without using dimension reduction. (c)-(d) RKSIR variates versus nonlinear factors.

compute the e.d.r. directions. In RKSIR, we used an additive Gaussian kernel as follows:

$$k(x, z) = \sum_{j=1}^{d} \exp\left(-\frac{(x_j - z_j)^2}{2\sigma^2}\right), \tag{36}$$

where $\sigma = 2$. After projecting the training samples on the estimated e.d.r. directions, we train a Gaussian kernel regression model based on the new variates. Then, the mean square error is computed on 1000 independent test samples. This experiment is repeated 100 times with all parameters set by cross-validation. The results are summarized in Figure 1.

Figure 1(a) displays the accuracy for RKSIR as a function of the regularization parameter, illustrating the importance of selecting regularization parameters. KSIR without regularization performs much worse than the RKSIR. Figure 1(b) displays the prediction accuracy of various dimension reduction methods.

RKSIR outperforms all the linear dimension reduction methods, which illustrates the power of nonlinearity introduced in RKSIR. It also suggests that there are essentially two nonlinear e.d.r. directions. This observation seems to agree with the model in (35). Indeed, Figures 1(c) and 1(d) show that the first two e.d.r. directions from RKSIR estimate the two nonlinear factors well.

4.2. Effect of Regularization. This example illustrates the effect of regularization on the performance of KSIR as a function of the anisotropy of the predictors.

The regression model has ten predictor variables $X = (X_1, \ldots, X_{10})$ and a univariate response specified by

$$Y = X_1 + X_2^2 + \varepsilon, \quad \varepsilon \sim N\left(0, 0.1^2\right), \tag{37}$$

where $X \sim N(0, \Sigma_X)$ and $\Sigma_X = Q\Delta Q$ with Q a randomly chosen orthogonal matrix and $\Delta = \text{diag}(1^\theta, 2^\theta, \ldots, 10^\theta)$. We will see that increasing the parameter $\theta \in [0, \infty)$ increases the anisotropy of the data that increases the difficulty of identifying the correct e.d.r. directions.

For this model, it is known that SIR will miss the direction along the second variable X_2. So we focus on the comparison of KSIR and RKSIR in this example.

If we use a second-order polynomial kernel $k(x, z) = (1 + x^T z)^2$ that corresponds to the feature space

$$\Phi(X) = \left\{1, X_i, \left(X_i X_j\right)_{i \le j}\right\}, \quad i, j = 1, \ldots, 10, \tag{38}$$

then $X_1 + X_2^2$ can be captured in one e.d.r. direction. Ideally the first KSIR variate $v = \langle \beta_1, \phi(X) \rangle$ should be equivalent to $X_1 + X_2^2$ modulo shift and scale

$$v - \mathbb{E}v \propto X_1 + X_2^2 - \mathbb{E}\left(X_1 + X_2^2\right). \tag{39}$$

So for this example given estimates of KSIR variates at the n data points $\{\widehat{v}_i\}_{i=1}^n = \{\langle \widehat{\beta}_1, \phi(x_i) \rangle\}_{i=1}^n$, the error of the first e.d.r. direction can be measured by the least squares fitting of v with respect to $(X_1 + X_2^2)$

$$\text{error} = \min_{a,b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \left(\widehat{v}_i - \left(a\left(x_{i,1} + x_{i,2}^2\right) + b\right)\right)^2. \tag{40}$$

We drew 200 observations from the model specified in (37), and then we applied the two dimension reduction methods KSIR and RKSIR. The mean and standard errors of 100 repetitions of this procedure are reported in Figure 2. The result shows that KSIR becomes more and more unstable as θ increases and the regularization helps to reduce this instability.

4.3. Importance of Nonlinearity and Regularization in Real Data. When SIR is applied to classification problems, it is equivalent to a Fisher discriminant analysis. For the case of multiclass classification, it is natural to use SIR and consider each class as a slice. Kernel forms of Fisher discriminant analysis (KFDA) [8] have been used to construct nonlinear discriminant surfaces and the regularization has improved performance of KFDA [25]. In this example, we show that this idea of adding a nonlinearity and a regularization term improves predictive accuracy in a real multiclass classification data set, the classification of handwritten digits.

The MNIST data set (Y. LeCun, http://yann.lecun.com/exdb/mnist/) contains 60,000 images of handwritten digits $\{0, 1, 2, \ldots, 9\}$ as training data and 10,000 images as test data. Each image consists of $p = 28 \times 28 = 784$ gray-scale pixel intensities. It is commonly believed that there is clear nonlinear structure in this 784-dimensional space.

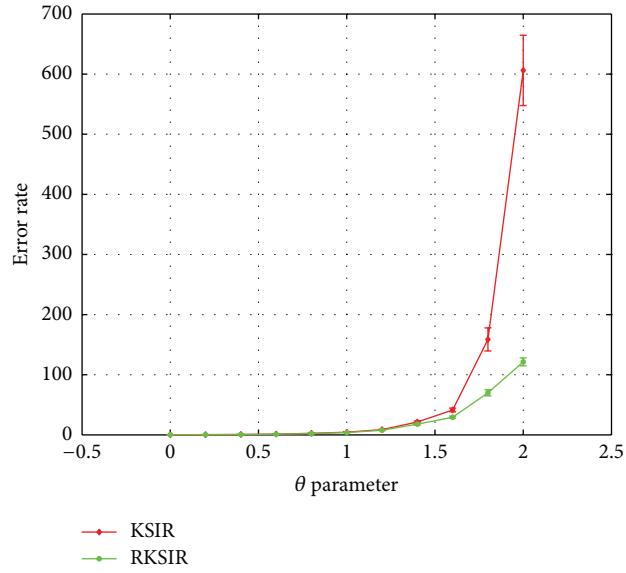

Figure 2: Error in e.d.r. as a function of θ.

We compared regularized SIR (RSIR) as in (26), KSIR, and RKSIR, on this data to examine the effect of regularization and nonlinearity. Each draw of the training set consisted of 100 observations of each digit. We then computed the top 10 e.d.r. directions using these 1000 observations and 10 slices, one for each digit. We projected the 10,000 test observations onto the e.d.r. directions and used a k-nearest neighbor (kNN) classifier with $k = 5$ to classify the test data. The accuracy of the kNN classifier without dimension reduction was used as a baseline. For KSIR and RKSIR we used a Gaussian kernel with the bandwith parameter set as the median pairwise distance between observations. The regularization parameter was set by cross-validation.

The mean and standard deviation of the classification accuracy over 100 iterations of this procedure are reported in Table 1. The first interesting observation is that the linear dimension reduction does not capture discriminative information, as the classification accuracy without dimension reduction is better. Nonlinearity does increase classification accuracy and coupling regularization with nonlinearity increases accuracy more. This improvement is dramatic for 2, 3, 5, and 8.

5. Discussion

The interest in manifold learning and nonlinear dimension reduction in both statistics and machine learning has led to a variety of statistical models and algorithms. However, most of these methods are developed in the unsupervised learning framework. Therefore, the estimated dimensions may not be optimal for the regression models. Our work incorporates nonlinearity and regularization to inverse regression approaches and results in a robust response driven nonlinear dimension reduction method.

RKHS has also been introduced into supervised dimension reduction in [26], where the conditional covariance

Table 1: Mean and standard deviations for error rates in classification of digits.

Digit	RKSIR	KSIR	RSIR	kNN
0	0.0273 (0.0089)	0.0472 (0.0191)	0.0487 (0.0128)	0.0291 (0.0071)
1	0.0150 (0.0049)	0.0177 (0.0051)	0.0292 (0.0113)	0.0052 (0.0012)
2	0.1039 (0.0207)	0.1475 (0.0497)	0.1921 (0.0238)	0.2008 (0.0186)
3	0.0845 (0.0208)	0.1279 (0.0494)	0.1723 (0.0283)	0.1092 (0.0130)
4	0.0784 (0.0240)	0.1044 (0.0461)	0.1327 (0.0327)	0.1617 (0.0213)
5	0.0877 (0.0209)	0.1327 (0.0540)	0.2146 (0.0294)	0.1419 (0.0193)
6	0.0472 (0.0108)	0.0804 (0.0383)	0.0816 (0.0172)	0.0446 (0.0081)
7	0.0887 (0.0169)	0.1119 (0.0357)	0.1354 (0.0172)	0.1140 (0.0125)
8	0.0981 (0.0259)	0.1490 (0.0699)	0.1981 (0.0286)	0.1140 (0.0156)
9	0.0774 (0.0251)	0.1095 (0.0398)	0.1533 (0.0212)	0.2006 (0.0153)
Average	0.0708 (0.0105)	0.1016 (0.0190)	0.1358 (0.0093)	0.1177 (0.0039)

operators on such kernel spaces are used to characterize the conditional independence between linear projections and the response variable. Therefore, their method estimates linear e.d.r. subspaces, while in [10, 11] and in this paper, RKHS is used to model the e.d.r. subspace, which leads to nonlinear dimension reduction.

There are several open issues in regularized kernel SIR method, such as the selection of kernels, regularization parameters, and number of dimensions. A direct assessment of the nonlinear e.d.r directions is expected to reduce the computational burden in procedures based on cross validation. While these are well established in linear dimension reduction, however, little is known for nonlinear dimension reduction. We would like to leave them for future research.

There are some interesting connections between KSIR and functional SIR, which are developed by Ferré and his coauthors in a series of papers [14, 17, 19]. In functional SIR, the observable data are functions and the goal is to find linear e.d.r. directions for functional data analysis. In KSIR, the observable data are typically not functions but mapped into a function space in order to characterize nonlinear structures. This suggests that computations involved in functional SIR can be simplified by a parametrization with respect to an RKHS or using a linear kernel in the parametrized function space. On the other hand, from a theoretical point of view, KSIR can be viewed as a special case of functional SIR, although our current theoretical results on KSIR are different from the ones for functional SIR.

Appendices

A. Proof of Theorem 3

Under the assumption of Proposition 2, for each $Y = y$,

$$\mathbb{E}\left[\phi(X) \mid Y = y\right] \in \operatorname{span}\left\{\Sigma\beta_i, i = 1, \ldots, d\right\}. \tag{A.1}$$

Since $\Gamma = \operatorname{cov}[\mathbb{E}(\phi(X) \mid Y)]$, the rank of Γ (i.e., the dimension of the image of Γ) is less than d, which implies that Γ is compact. With the fact that it is symmetric and semipositive, there exist d_Γ positive eigenvalues $\{\tau_i\}_{i=1}^{d_\Gamma}$ and eigenvectors $\{u_i\}_{i=1}^{d_\Gamma}$, such that $\Gamma = \sum_{i=1}^{d_\Gamma} \tau_i u_i \otimes u_i$.

Recall that for any $f \in \mathcal{H}$,

$$\Gamma f = \mathbb{E}\left\langle \mathbb{E}\left[\phi(X) \mid Y\right], f\right\rangle \mathbb{E}\left[\phi(X) \mid Y\right] \tag{A.2}$$

also belongs to

$$\operatorname{span}\left\{\Sigma\beta_i, i = 1, \ldots, d\right\} \subset \operatorname{range}(\Sigma) \tag{A.3}$$

because of (A.1); so

$$u_i = \frac{1}{\tau_i}\Gamma u_i \in \operatorname{range}(\Sigma). \tag{A.4}$$

This proves (i).

Since for each $f \in \mathcal{H}$, $\Gamma f \in \operatorname{range}(\Sigma^{-1})$, the operator $T = \Sigma^{-1}\Gamma$ is well defined over the whole space. Moreover,

$$\begin{aligned} Tf &= \Sigma^{-1}\left(\sum_{i=1}^{d_\Gamma}\left\langle u_i, f\right\rangle u_i\right) \\ &= \sum_{i=1}^{d_\Gamma}\left\langle u_i, f\right\rangle \Sigma^{-1}(u_i) = \left(\sum_{i=1}^{d_\Gamma}\Sigma^{-1}(u_i)\otimes u_i\right) f. \end{aligned} \tag{A.5}$$

This proves (ii).

B. Proof of Proposition 7

We first prove the proposition for matrices to simplify then notation; we then extend the result to the operators, where d_K is infinite and a matrix form does not make sense.

Let $\Phi = [\phi(x_1), \ldots, \phi(x_n)]$. It has the following SVD decomposition:

$$\begin{aligned} \Phi &= UDV^T \\ &= \begin{bmatrix} u_1 \cdots u_{d_K} \end{bmatrix} \begin{bmatrix} \overline{D}_{d\times d} & 0_{d\times(n-d)} \\ 0_{(d_K-d)\times d} & 0_{(d_K-d)\times(n-d)} \end{bmatrix} \begin{bmatrix} v_1^T \\ \vdots \\ v_n^T \end{bmatrix} = \overline{U}\,\overline{D}\,\overline{V}^T, \end{aligned} \tag{B.1}$$

where $\overline{U} = [u_1, \ldots, u_d]$, $\overline{V} = [v_1, \ldots, v_d]$ and $\overline{D} = \overline{D}_{d\times d}$ is a diagonal matrix of dimension $d \le n$.

We need to show the KSIR variates

$$\begin{aligned}\widehat{v}_j &= c_j^T \left[k\left(x, x_1\right), \ldots, k\left(x, x_n\right)\right] \\ &= c_j^T \Phi^T \Phi\left(x\right) = \left\langle \Phi c_j, \phi\left(x\right)\right\rangle = \left\langle \beta_j, \phi\left(x\right)\right\rangle.\end{aligned} \tag{B.2}$$

It suffices to prove that if (λ, c) is a solution to (28), then (λ, β) is also a pair of eigenvalue and eigenvector of $\left(\widehat{\Sigma}^2 + \gamma I\right)^{-1}\widehat{\Sigma}\widehat{\Gamma}$ and vice versa, where c and β are related by

$$\beta = \Phi c, \quad c = \overline{V}\,\overline{D}\,\overline{U}^T\beta. \tag{B.3}$$

Noting that facts $\widehat{\Sigma} = (1/n)\Phi\Phi^T$, $\widehat{\Gamma} = (1/n)\Phi J\Phi^T$, and $K = \Phi^T\Phi = \overline{V}\,\overline{D}^2\,\overline{V}^T$, the argument may be made as follows:

$$\begin{aligned}&\widehat{\Sigma}\widehat{\Gamma}\beta = \lambda\left(\widehat{\Sigma}^2 + \gamma I\right)\beta \\ &\quad\Longleftrightarrow \Phi\Phi^T\Phi J\Phi^T\Phi c = \lambda\left(\Phi\Phi^T\Phi\Phi^T\Phi c + n^2\gamma\Phi c\right) \\ &\quad\Longleftrightarrow \Phi KJKc = \lambda\Phi\left(K^2 + n^2\gamma\right)c \\ &\quad\left(\Longleftrightarrow \overline{V}\,\overline{V}^T KJKc = \lambda\overline{V}\,\overline{V}^T\left(K^2 + n^2\gamma I\right)c\right) \\ &\quad\Longleftrightarrow KJKc = \lambda\left(K^2 + n^2\gamma I\right)c.\end{aligned} \tag{B.4}$$

Note that the implication in the third step is necessary only in the $\Rightarrow$ direction which is obtained by multiplying both sides $\overline{V}\,\overline{D}^{-1}\overline{U}^T$ and using the facts $\overline{U}^T\overline{U} = I_d$. For the last step, since $\overline{V}^T\overline{V} = I_d$, we use the following facts:

$$\begin{aligned}\overline{V}\,\overline{V}^T K &= \overline{V}\,\overline{V}^T\overline{V}\,\overline{D}^2\overline{V}^T = \overline{V}\,\overline{D}^2\overline{V}^T = K, \\ \overline{V}\,\overline{V}^T c &= \overline{V}\,\overline{V}^T\overline{V}\,\overline{D}^{-1}\overline{U}^T\beta = \overline{V}\,\overline{D}^{-1}\overline{U}^T\beta = c.\end{aligned} \tag{B.5}$$

In order for this result to hold rigorously when the RKHS is infinite dimensional, we need to formally define Φ, Φ^T, and the SVD of Φ when d_K is infinite. For the infinite dimensional case, Φ is an operator from $\mathbb{R}^n$ to $\mathscr{H}_K$ defined by $\Phi v = \sum_{i=1}^n v_i\phi(x_i)$ for $v = (v_1, \ldots, v_n)^T \in \mathbb{R}^n$ and Φ^T is its adjoint, an operator from $\mathscr{H}_K$ to $\mathbb{R}^n$ such that $\Phi^T f = (\langle\phi(x_1), f\rangle_K, \ldots, \langle\phi(x_1), f\rangle_K)^T$ for $f \in \mathscr{H}_K$. The notions $\overline{U}$ and $\overline{U}^T$ are similarly defined.

The above formulation of Φ and Φ^T coincides the definition of $\widehat{\Sigma}$ as a covariance operator. Since the rank of $\widehat{\Sigma}$ is less than n, it is compact and has the following representation:

$$\widehat{\Sigma} = \sum_{i=1}^{d_K}\widehat{\sigma}_i u_i \otimes u_i = \sum_{i=1}^{d}\sigma_i u_i \otimes u_i, \tag{B.6}$$

where $d \le n$ is the rank and $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_d > \sigma_{d+1} = \cdots = 0$. This implies that each $\phi(x_i)$ lies in $\operatorname{span}(u_1, \ldots, u_d)$ and hence we can write $\phi(x_i) = \overline{U}\tau_i$, where $\overline{U} = (u_1, \ldots, u_d)$ should be considered as an operator from $\mathbb{R}^d$ to $\mathscr{H}_K$ and $\tau_i \in \mathbb{R}^d$. Denote $\Upsilon = (\tau_1, \ldots, \tau_n)^T \in \mathbb{R}^{n\times d}$. It is easy to check that $\Upsilon^T\Upsilon = \operatorname{diag}(n\sigma_1, \ldots, n\sigma_d)$. Let $\overline{D}_{d\times d} = \operatorname{diag}(\sqrt{n\sigma_1}, \ldots, \sqrt{n\sigma_d})$ and $\overline{V} = \Upsilon\overline{D}^{-1}$. Then, we obtain the SVD for Φ as $\Phi = \overline{U}\overline{D}\overline{V}^T$ which is well defined.

C. Proof of Consistency

C.1. Preliminaries. In order to prove Theorems 9 and 10, we use the properties of the Hilbert-Schmidt operators, covariance operators for the Hilbert space valued random variables, and the perturbation theory for linear operators. In this subsection we provide a brief introduction to them. For details, see [27–29] and references therein.

Given a separable Hilbert space $\mathscr{H}$ of dimension $p_{\mathscr{H}}$, a linear operator L on $\mathscr{H}$ is said to belong to the Hilbert-Schmidt class if

$$\|L\|_{\mathrm{HS}}^2 = \sum_{i=1}^{p_{\mathscr{H}}}\left\|Le_i\right\|_{\mathscr{H}}^2 < \infty, \tag{C.1}$$

where $\{e_i\}$ is an orthonormal basis. The Hilbert-Schmidt class forms a new Hilbert space with norm $\|\cdot\|_{\mathrm{HS}}$.

Given a bounded operator S on $\mathscr{H}$, the operators SL and LS both belong to the Hilbert-Schmidt class and the following holds:

$$\|SL\|_{\mathrm{HS}} \le \|S\|\,\|L\|_{\mathrm{HS}}, \qquad \|LS\|_{\mathrm{HS}} \le \|L\|_{\mathrm{HS}}\,\|S\|, \tag{C.2}$$

where $\|\cdot\|$ denotes the default operator norm

$$\|L\|^2 = \sup_{f\in\mathscr{H}}\frac{\|Lf\|^2}{\|f\|^2}. \tag{C.3}$$

Let Z be a random vector taking values in $\mathscr{H}$ satisfying $\mathbb{E}\|Z\|^2 < \infty$. The covariance operator

$$\Sigma = \mathbb{E}\left[(Z - \mathbb{E}Z)\otimes(Z - \mathbb{E}Z)\right], \tag{C.4}$$

is self-adjoint, positive, compact, and belongs to the Hilbert-Schmidt class.

A well-known result from perturbation theory for linear operators states that if a set of linear operators T_n converges to T in the Hilbert-Schmidt norm and the eigenvalues of T are nondegenerate, then the eigenvalues and eigenvectors of T_n converge to those of T with same rate or convergence as the convergence of the operators.

C.2. Proof of Theorem 10. We will use the following result from [14]:

$$\left\|\widehat{\Sigma} - \Sigma\right\|_{\mathrm{HS}} = O_p\left(\frac{1}{\sqrt{n}}\right), \qquad \left\|\widehat{\Gamma} - \Gamma\right\|_{\mathrm{HS}} = O_p\left(\frac{1}{\sqrt{n}}\right). \tag{C.5}$$

To simplify the notion, we denote $\widehat{T}_s = \left(\widehat{\Sigma}^2 + sI\right)^{-1}\widehat{\Sigma}\widehat{\Gamma}$. Also, define

$$T_1 = \left(\widehat{\Sigma}^2 + sI\right)^{-1}\Sigma\Gamma, \qquad T_2 = \left(\Sigma^2 + sI\right)^{-1}\Sigma\Gamma. \tag{C.6}$$

Then,

$$\begin{aligned}&\left\|\widehat{T}_s - T\right\|_{\mathrm{HS}} \\ &\quad\le \left\|\widehat{T}_s - T_1\right\|_{\mathrm{HS}} + \left\|T_1 - T_2\right\|_{\mathrm{HS}} + \left\|T_2 - T\right\|_{\mathrm{HS}}.\end{aligned} \tag{C.7}$$

For the first term, observe that

$$\left\|\widehat{T}_s - T_1\right\|_{\mathrm{HS}} \le \left\|\left(\widehat{\Sigma}^2 + sI\right)^{-1}\right\| \left\|\widehat{\Sigma}\widehat{\Gamma} - \Sigma\Gamma\right\|_{\mathrm{HS}} = O_p\left(\frac{1}{s\sqrt{n}}\right). \tag{C.8}$$

For the second term, note that

$$T_1 = \sum_{j=1}^{d_\Gamma} \tau_j \left(\left(\widehat{\Sigma}^2 + sI\right)^{-1} \Sigma u_j\right) \otimes u_j,$$
$$T_2 = \sum_{j=1}^{d_\Gamma} \tau_j \left(\left(\Sigma^2 + sI\right)^{-1} u_j\right) \otimes u_j. \tag{C.9}$$

Therefore,

$$\left\|T_1 - T_2\right\|_{\mathrm{HS}} = \sum_{j=1}^{d_\Gamma} \tau_j \left\|\left(\left(\widehat{\Sigma}^2 + sI\right)^{-1} - \left(\Sigma^2 + sI\right)^{-1}\right) \Sigma u_j\right\|. \tag{C.10}$$

Since $u_j \in \mathrm{range}(\Sigma)$, there exists $\widetilde{u}_j$ such that $u_j = \Sigma\widetilde{u}_j$. Then,

$$\left(\left(\Sigma^2 + sI\right)^{-1} - \left(\widehat{\Sigma}^2 + sI\right)^{-1}\right) \Sigma u_j = \left(\widehat{\Sigma}^2 + sI\right)^{-1} \left(\widehat{\Sigma}^2 - \Sigma^2\right) \left(\Sigma^2 + sI\right)^{-1} \Sigma^2 \widetilde{u}_j, \tag{C.11}$$

which implies

$$\begin{aligned} \left\|T_1 - T_2\right\| &\le \sum_{j=1}^{d_\Gamma} \tau_j \left\|\left(\widehat{\Sigma}^2 + sI\right)^{-1}\right\| \left\|\widehat{\Sigma}^2 - \Sigma^2\right\|_{\mathrm{HS}} \\ &\quad \times \left\|\left(\widehat{\Sigma}^2 + sI\right)^{-1} \Sigma^2\right\| \left\|\widetilde{u}_j\right\| \\ &= O_p\left(\frac{1}{s\sqrt{n}}\right). \end{aligned} \tag{C.12}$$

For the third term, the following holds:

$$\left\|T_2 - T\right\|_{\mathrm{HS}}^2 = \sum_{j=1}^{d_\Gamma} \tau_j \left\|\left(\left(\Sigma^2 + sI\right)^{-1} \Sigma - \Sigma^{-1}\right) u_j\right\|^2, \tag{C.13}$$

and for each $j = 1, \ldots, d_\Gamma$,

$$\begin{aligned} &\left\|\left(\Sigma^2 + sI\right)^{-1} \Sigma u_j - \Sigma^{-1} u_j\right\| \\ &\le \left\|\left(\Sigma^2 + sI\right)^{-1} \Sigma^2 \widetilde{u}_j - \widetilde{u}_j\right\| \\ &= \left\|\sum_{i=1}^{\infty} \left(\frac{w_j^2}{s + w_j^2} - 1\right) \left\langle \widetilde{u}_j, e_i \right\rangle e_i\right\| \\ &= \left(\sum_{i=1}^{\infty} \frac{s^2}{(s + w_i^2)^2} \left\langle \widetilde{u}_j, e_i \right\rangle^2\right)^{1/2} \\ &\le \frac{s}{w_N} \left(\sum_{i=1}^{N} \left\langle \widetilde{u}_j, e_i \right\rangle^2\right)^{1/2} + \left(\sum_{i=N+1}^{\infty} \left\langle \widetilde{u}_j, e_i \right\rangle^2\right)^{1/2} \\ &= \frac{s}{w_N^2} \left\|\Pi_N\left(\widetilde{u}_j\right)\right\| + \left\|\Pi_N^{\perp}\left(\widetilde{u}_j\right)\right\|. \end{aligned} \tag{C.14}$$

Combining these terms results in (33).

Since $\|\Pi_N^{\perp}(\widetilde{u}_j)\| \to 0$ as $N \to \infty$, consequently, we have

$$\left\|\widehat{T}_s - T\right\|_{\mathrm{HS}} = o_p(1) \tag{C.15}$$

if $s \to 0$ and $s\sqrt{n} \to \infty$.

If all the e.d.r. directions β_i depend only on a finite number of eigenvectors of the covariance operator, then there exist some $N > 1$ such that $\mathcal{S}^* = \mathrm{span}\{\Sigma e_i,\ i = 1, \ldots, N\}$. This implies that

$$\widetilde{u}_j = \Sigma^{-1} u_j \in \Sigma^{-1}\left(\mathcal{S}^*\right) \subset \mathrm{span}\left\{e_i,\ i = 1, \ldots, N\right\}. \tag{C.16}$$

Therefore, $\|\Pi_N^{\perp}(\widetilde{u}_j)\| = 0$. Let $s = O(n^{-1/4})$; the rate is $O(n^{1/4})$.

Acknowledgments

The authors acknowledge the support of the National Science Foundation (DMS-0732276 and DMS-0732260) and the National Institutes of Health (P50 GM 081883). Any opinions, findings and conclusions, or recommendations expressed in this work are those of the authors and do not necessarily reflect the views of the NSF or NIH.

References

[1] K. Li, "Sliced inverse regression for dimension reduction (with discussion)," *Journal of the American Statistical Association*, vol. 86, no. 414, pp. 316–342, 1991.

[2] N. Duan and K. Li, "Slicing regression: a link-free regression method," *The Annals of Statistics*, vol. 19, no. 2, pp. 505–530, 1991.

[3] R. Cook and S. Weisberg, "Comment," *Journal of the American Statistical Association*, vol. 86, no. 414, pp. 328–332, 1991.

[4] K. C. Li, "On principal hessian directions for data visualization and dimension reduction: another application of stein's lemma," *Journal of the American Statistical Association*, vol. 87, no. 420, pp. 1025–1039, 1992.

[5] L. Li, R. D. Cook, and C.-L. Tsai, "Partial inverse regression," *Biometrika*, vol. 94, no. 3, pp. 615–625, 2007.

[6] B. Schölkopf, A. J. Smola, and K. Müller, "Kernel principal component analysis," in *Proceedings of the Artificial Neural Networks (ICANN '97)*, W. Gerstner, A. Germond, M. Hasler, and J.-D. Nicoud, Eds., vol. 1327 of *Lecture Notes in Computer Science*, pp. 583–588, Springer, Lausanne, Switzerland, October 1997.

[7] B. Schölkopf and A. J. Smola, *Learning with Kernels*, The MIT Press, Massachusetts, Mass, USA, 2002.

[8] G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach," *Neural Computation*, vol. 12, no. 10, pp. 2385–2404, 2000.

[9] F. R. Bach and M. I. Jordan, "Kernel independent component analysis," *Journal of Machine Learning Research*, vol. 3, no. 1, pp. 1–48, 2002.

[10] H.-M. Wu, "Kernel sliced inverse regression with applications to classification," *Journal of Computational and Graphical Statistics*, vol. 17, no. 3, pp. 590–610, 2008.

[11] Y.-R. Yeh, S.-Y. Huang, and Y.-J. Lee, "Nonlinear dimension reduction with kernel sliced inverse regression," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 11, pp. 1590–1603, 2009.

[12] J. Mercer, "Functions of positive and negative type and their connection with the theory of integral equations," *Philosophical Transactions of the Royal Society of London A*, vol. 209, no. 441–458, pp. 415–446, 1909.

[13] H. König, *Eigenvalue Distribution of Compact Operators*, vol. 16 of *Operator Theory: Advances and Applications*, CH, Basel, Switzerland, 1986.

[14] L. Ferré and A. Yao, "Functional sliced inverse regression analysis," *Statistics*, vol. 37, no. 6, pp. 475–488, 2003.

[15] P. Hall and K.-C. Li, "On almost linearity of low-dimensional projections from high-dimensional data," *The Annals of Statistics*, vol. 21, no. 2, pp. 867–889, 1993.

[16] G. He, H.-G. Müller, and J.-L. Wang, "Functional canonical analysis for square integrable stochastic processes," *Journal of Multivariate Analysis*, vol. 85, no. 1, pp. 54–77, 2003.

[17] L. Ferré and A. Yao, "Smoothed functional inverse regression," *Statistica Sinica*, vol. 15, no. 3, pp. 665–683, 2005.

[18] W. Zhong, P. Zeng, P. Ma, J. S. Liu, and Y. Zhu, "RSIR: regularized sliced inverse regression for motif discovery," *Bioinformatics*, vol. 21, no. 22, pp. 4169–4175, 2005.

[19] L. Ferré and N. Villa, "Multilayer perceptron with functional inputs: an inverse regression approach," *Scandinavian Journal of Statistics*, vol. 33, no. 4, pp. 807–823, 2006.

[20] L. Li and X. Yin, "Sliced inverse regression with regularizations," *Biometrics*, vol. 64, no. 1, pp. 124–131, 2008.

[21] C. Bernard-Michel, L. Gardes, and S. Girard, "Gaussian regularized sliced inverse regression," *Statistics and Computing*, vol. 19, no. 1, pp. 85–98, 2009.

[22] T. Hsing and R. J. Carroll, "An asymptotic theory for sliced inverse regression," *The Annals of Statistics*, vol. 20, no. 2, pp. 1040–1061, 1992.

[23] J. Saracco, "An asymptotic theory for sliced inverse regression," *Communications in Statistics*, vol. 26, no. 9, pp. 2141–2171, 1997.

[24] L. X. Zhu and K. W. Ng, "Asymptotics of sliced inverse regression," *Statistica Sinica*, vol. 5, no. 2, pp. 727–736, 1995.

[25] T. Kurita and T. Taguchi, "A kernel-based fisher discriminant analysis for face detection," *IEICE Transactions on Information and Systems*, vol. 88, no. 3, pp. 628–635, 2005.

[26] K. Fukumizu, F. R. Bach, and M. I. Jordan, "Kernel dimension reduction in regression," *The Annals of Statistics*, vol. 37, no. 4, pp. 1871–1905, 2009.

[27] T. Kato, *Perturbation Theory for Linear Operators*, Springer, Berlin, Germany, 1966.

[28] F. Chatelin, *Spectral Approximation of Linear Operators*, Academic Press, 1983.

[29] G. Blanchard, O. Bousquet, and L. Zwald, "Statistical properties of kernel principal component analysis," *Machine Learning*, vol. 66, no. 2-3, pp. 259–294, 2007.