

# Consistency of regularized sliced inverse regression for kernel models

Qiang Wu, Feng Liang, and Sayan Mukherjee\*

September 4, 2008

## Abstract

We develop an extension of the sliced inverse regression (SIR) framework for dimension reduction using kernel models and Tikhonov regularization. The result is a numerically stable nonlinear dimension reduction method. We prove consistency of the method under weak conditions even when the reproducing kernel Hilbert space induced by the kernel is infinite dimensional. We illustrate the utility of this approach on simulated and real data.

*Keywords:* Dimension reduction, sliced inverse regression, kernel methods, regularization

---

\*Qiang Wu and Sayan Mukherjee are members of the Departments of Statistical Science and Computer Science and Institute for Genome Sciences & Policy, Duke University, Durham, NC 27708; Feng Liang is a member of the Department of Statistics, University of Illinois at Urbana-Champaign, IL 61820

# 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, that is,  $S(X) = (\beta_1^T X, \dots, \beta_d^T X)$ . The  $d$ -dimensional subspace,  $\mathcal{S} = \text{span}(\beta_1, \dots, \beta_d)$ , is defined as the effective dimension reduction (e.d.r.) space in Li [1991] since  $\mathcal{S}$  summaries all the information we need to know about  $Y$ . A key result in Li [1991] is that under some mild conditions the e.d.r. directions  $\{\beta_j\}_{j=1}^d$  correspond to the eigenvectors of the matrix

$$T = [\text{cov}(X)]^{-1} \text{cov}[\mathbb{E}(X|Y)].$$

Thus the e.d.r. directions or subspace can be estimated via an eigenanalysis of matrix  $T$ , which is the foundation of the sliced inverse regression algorithm proposed by Li [1991] and Duan and Li [1991]. Further developments include sliced average variance estimation [Cook and Weisberg 1991]. Recently this framework was extended to the high-dimensional setting where there are more covariates  $p$  than observations  $n$  in Li *et al.* [2007].

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, often there exists a nonlinear embedding of the data into a Hilbert space where a linear method can capture the low-dimensional structure. If projections onto this low-dimensional structure can be computed by inner products in this Hilbert space, kernel methods [Schölkopf *et al.* 1997, Schölkopf and Smola 2002] can be used to obtain simple and efficient algorithms. The basic idea in applying kernel methods is the application of a linear algorithm to the data mapped into a feature space induced by the kernel function. 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 are kernel principle component analysis [Schölkopf *et al.* 1997] and kernel independent correlation analysis [Bach and Jordan 2002]. This idea was applied to SIR in Wu [2008] 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 in a direct application of this kernel idea to SIR. If the kernel function is infinite dimensional, the embedding induced by the kernel function maps a point in the  $p$ -dimensional covariate space into an infinite dimensional Hilbert space, then  $T$  is an operator and is not well defined. In addition, for infinite or high-dimensional feature spaces an empirical estimate of  $T$  based from observations is often ill-conditioned and results in computational instability. We seek to overcome these difficulties by proposing a modified version of KSIR that incorporates a Tikhonov regularization term. The method is stated as an eigen-decomposition problem as well as a least squares problem in Section 3.

Addition of this regularization term has both theoretical and practical implications. The theoretical contribution is a proving asymptotic consistency of the e.d.r. estimates and with a rate of  $O(n^{1/4})$  under stated conditions, see Section 4. The practical contribution is empirical evidence that addition of the regularization term improves predictive accuracy, see Section 5.

## 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 (RKHS) and in particular Mercer kernels [Mercer 1909].

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

$$k(x, z) = \sum_j \lambda_j \phi_j(x) \phi_j(z),$$

where  $\{\phi_j\}$  are the eigenfunctions and  $\{\lambda_j\}$  are the corresponding non-negative, non-increasing eigenvalues. An important property of Mercer kernels is that each kernel  $k$  uniquely corresponds to a reproducing kernel Hilbert space (RKHS)

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

where the cardinality of  $\Lambda := \{j : \lambda_j > 0\}$  is the dimension of the RKHS which may be infinite [Mercer 1909, König 1986].

The key idea is given a Mercer kernel there exists a unique map or embedding  $\phi$  from  $\mathcal{X}$  to a Hilbert space defined by the eigenvalues and eigenfunctions of the kernel. The map takes the form

$$\phi(x) = \left( \sqrt{\lambda_1} \phi_1(x), \sqrt{\lambda_2} \phi_2(x), \dots, \sqrt{\lambda_{|\Lambda|}} \phi_{|\Lambda|}(x) \right). \quad (3)$$

The Hilbert space induced by this map with the standard inner product  $k(x, z) = \langle \phi(z), \phi(x) \rangle$  is isomorphic to the RKHS (2) and we will denote both these Hilbert spaces as  $\mathcal{H}$  [König 1986]. In the case where  $k$  is infinite dimensional,  $\phi : \mathcal{X} \rightarrow \ell_2$ .

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  $\mathbb{E}\|Z\|^2 < \infty$ , then the covariance operator of  $Z$  is defined as  $\mathbb{E}[(Z - \mathbb{E}Z) \otimes (Z - \mathbb{E}Z)]$ , where

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

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  $x \in \mathcal{X}$ ,  $k(X, X) \leq M$  almost surely (a.s.) with respect to  $\mathcal{P}$ .
3.  $\mathcal{H}$  is separable.

Under Assumption 1, the random element  $\phi(X)$  has a well-defined mean and covariance operator because  $\|\phi(x)\|^2 = k(x, x)$  is bounded (a.s.). Without loss of generality, we assume  $\mathbb{E}\phi(X) = \mathbf{0}$  where  $\mathbf{0}$  is the vector of zeros 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

$$\Sigma = \sum_{i=1}^{\infty} w_i e_i \otimes e_i, \tag{4}$$

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  $X$ ,

$$Y = F(\langle \beta_1, \phi(X) \rangle, \dots, \langle \beta_d, \phi(X) \rangle, \varepsilon), \tag{5}$$

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

$$S(X) = (\langle \beta_1, \phi(X) \rangle, \dots, \langle \beta_d, \phi(X) \rangle).$$

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  $\mathcal{S} = \text{span}(\beta_1, \dots, \beta_d)$  the nonlinear e.d.r. space. The following proposition [Wu 2008] extends the theoretical foundation of SIR to this nonlinear setting.

**Proposition 1.** 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

$$\mathbb{E}[\langle f, \phi(X) \rangle | S(X)] = b^T S(X), \text{ with } S(X) = (\langle \beta_1, \phi(X) \rangle, \dots, \langle \beta_d, \phi(X) \rangle)^T. \quad (6)$$

Then for the model specified in equation (5) the inverse regression curve  $\mathbb{E}[\phi(X)|Y]$  is contained in the span of  $(\Sigma\beta_1, \dots, \Sigma\beta_d)$ , where  $\Sigma$  is the covariance operator of  $\phi(X)$ .

Proposition 1 is a straightforward extension of the multivariate case in Li [1991] to a Hilbert space or a direct application of the functional SIR setting in Ferré and Yao [2003]. Although the linear design condition (6) may be difficult to check in practice, it has been shown that such a condition usually holds approximately in a high-dimensional space [Hall and Li 1993],

An immediate consequence of this proposition is that nonlinear e.d.r. directions are the eigenvectors corresponding to the largest eigenvalues of the following generalized eigen-decomposition problem

$$\Gamma\beta = \lambda\Sigma\beta, \quad \text{where } \Sigma = \text{cov}[\phi(X)], \quad \Gamma = \text{cov}[\mathbb{E}(\phi(X)|Y)], \quad (7)$$

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

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

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 image of  $\Gamma$  is indeed in the range of  $\Sigma^{-1}$ . A similar issue also arose in the analysis of dimension reduction and canonical analysis for functional data [He et al. 2003, Ferré and Yao 2005]. 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 1.** Under Assumption 1 and the linear design condition (6) in Proposition 1 the following hold:

- (i) The operator  $\Gamma$  is of finite rank  $d_\Gamma \leq d$ . Consequently, it is compact and has the following spectral decomposition

$$\Gamma = \sum_{i=1}^{d_\Gamma} \tau_i u_i \otimes u_i, \quad (8)$$

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

- (ii) *The eigen-decomposition problem (7) is equivalent to the eigenanalysis of the operator  $T$ , which takes the following form*

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

### 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 fundamental computational challenges to this idea such as estimating a 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 Tikhonov regularization term to kernel SIR. We also provide a least squares formulation of this problem which enables us to select the regularization parameter via cross-validation.

#### 3.1 Estimating the nonlinear E.D.R. directions

Given  $n$  observations  $\{(x_1, y_1), \dots, (x_n, y_n)\}$  our objective is to obtain an estimate of the e.d.r. directions  $(\hat{\beta}_1, \dots, \hat{\beta}_d)$ . 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  $\sum_{i=1}^n \phi(x_i) = \mathbf{0}$ . The sample covariance is estimated by

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n \phi(x_i) \otimes \phi(x_i)$$

2. Bin the  $Y$  variables into  $H$  slices  $G_1, \dots, G_H$  and compute mean vectors of the corresponding mapped predictor variables for each group

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

Compute the sample between-group covariance matrix

$$\hat{\Gamma} = \sum_{h=1}^H \frac{n_h}{n} \psi_h \otimes \psi_h.$$

3. Estimate the SIR directions  $\hat{\beta}_j$  by solving the generalized eigen-decomposition problem

$$\hat{\Gamma} \beta = \lambda \hat{\Sigma} \beta. \tag{9}$$

This procedure is computationally impossible if the RKHS is infinite dimensional or the feature map cannot be computed. However, the model given in (5) requires not the e.d.r. directions but only the projection onto these directions, that is, the  $d$  summary statistics

$$v_1 = \langle \beta_1, \phi(X) \rangle, \dots, v_d = \langle \beta_d, \phi(X) \rangle,$$

which we call the KSIR variates. The KSIR variates can be efficiently computed and require only the kernel  $k(\cdot, \cdot)$ , not the map  $\phi$ .

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

$$\begin{aligned} K_{ij} &= \langle \phi(x_i) - \bar{\phi}, \phi(x_j) - \bar{\phi} \rangle \\ &= k(x_i, x_j) - \frac{1}{n} \sum_{j=1}^n k(x_i, x_j) - \frac{1}{n} \sum_{i=1}^n k(x_i, x_j) + \frac{1}{n^2} \sum_{i,j=1}^n k(x_i, x_j). \end{aligned}$$

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

Given the centered gram matrix  $K$ , the following generalized eigen-decomposition problem can be used to compute the KSIR variates

$$KJKc = \lambda K^2c, \tag{10}$$

where  $c$  denotes the  $n$ -dimensional generalized eigenvector, and  $J$  denotes a  $n \times n$  matrix with  $J_{ij} = 1/n_m$  if  $i, j$  are in the  $m$ -th group consisting of  $n_m$  observations and zero otherwise. The following proposition states that two eigen-decomposition problems, (10) and (9), are equivalent in the recover the same KSIR covariates  $(v_1, \dots, v_d)$ .

**Proposition 2.** *Given observations  $\{(x_1, y_1), \dots, (x_n, y_n)\}$ , let  $(\hat{\beta}_1, \dots, \hat{\beta}_d)$  and  $(\hat{c}_1, \dots, \hat{c}_d)$  denote the generalized eigenvectors of (10) and (9), respectively. Then for any  $x \in \mathcal{X}$  and  $j = 1, \dots, d$ , the following holds*

$$\langle \hat{\beta}_j, \phi(x) \rangle = \hat{c}_j^T K_x, \quad K_x = (k(x, x_1), \dots, k(x, x_n))^T,$$

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

This result was proven in Wu [2008], in which the algorithm was further reduced to solving

$$JKc = \lambda Kc$$

by canceling  $K$  from both sides of (10). This may cause some problems since  $K$  is singular so we work with the original symmetric form (10), which also results in a simple interpretation of the regularization approach we now introduce.

## 3.2 Regularization and stability

The eigen-decomposition in equation (10) will often be ill-conditioned resulting in over-fitting 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  $\hat{\Sigma}$  or by adding a regularization term to (10). We introduce the following regularization:

$$KJKc = \lambda(K^2 + n^2sI)c, \quad (11)$$

where  $s$  is a tuning parameter. This results in robust estimates of the e.d.r. space and improved predictive accuracy. The following proposition, see Appendix B for proof, states that solving the generalized eigen-decomposition problem (11) is equivalent to finding the eigenvectors of

$$(\hat{\Sigma}^2 + sI)^{-1}\hat{\Sigma}\hat{\Gamma}. \quad (12)$$

**Proposition 3.** *Let  $\hat{c}_j$  be the eigenvectors of (11) and  $\hat{\beta}_j$  be the eigenvectors of (12), then the following holds for the regularized KSIR covariates*

$$\hat{v}_j = \hat{c}_j^T [k(x, x_1), \dots, k(x, x_n)] = \langle \hat{\beta}_j, \phi(x) \rangle.$$

Ridge type regularization is an alternative widely used in both linear SIR and functional SIR [Zhong *et al.* 2005, Ferré and Villa 2006, Li and Yin 2008b] and solves the eigen-decomposition

$$\hat{\Gamma}\beta = \lambda(\hat{\Sigma} + sI)\beta. \quad (13)$$

It was shown in Bernard-Michel *et al.* [2008] that Tikhonov regularization is more efficient.

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 eigen-decomposition problem is for  $n \times n$  matrices rather than the  $p \times p$  matrices in the standard SIR formulation.

## 3.3 Least square formulation

The generalized eigen-decomposition for linear SIR can be written as a least square formulation [Cook 2004, Li and Yin 2008b]. The advantage of this formulation is that a greater variety of regularization or shrinkage models can be applied such as  $\ell_1$  penalties. In addition, classical techniques such as generalized cross validation can be used to select the regularization parameter. We can also formulate regularized KSIR as a least square problem.

Denote  $z_i$  as the  $i$ -th column of  $K$ . This corresponds to thinking of the matrix  $K$  as a data matrix of  $n$  observations in a  $n$ -dimensional space. For this new data matrix, we can estimate the covariance of the inverse regression as

$$\hat{\Gamma} = \sum_{h=1}^H \frac{n_h}{n} \bar{Z}_h \bar{Z}_h^T = \frac{1}{n} KJK,$$

where  $\bar{Z}_h = \frac{1}{n_h} \sum_{i \in G_h} z_i$  and the e.d.r. directions can be estimated by solving the eigen-decomposition problem in equation (10). Informally, this perspective views KSIR as linear SIR in an  $n$ -dimensional space. This formulation and arguments in Cook [2004], Li and Yin [2008b] lead to the following least square penalty

$$\mathbf{G}(C, A) = \sum_{h=1}^H \frac{n_h}{n} \left\| \bar{Z}_h - \frac{1}{n} K^2 C a_h \right\|^2,$$

where  $C = (c_1, c_2, \dots, c_d) \in \mathbb{R}^{n \times d}$  and  $A = (a_1, \dots, a_H) \in \mathbb{R}^{d \times H}$ . By arguments from Li and Yin [2008b] the top  $d$  eigenvectors for the KSIR algorithm (10) can be computed by minimizing  $\mathbf{G}(C, A)$  with respect to  $C$  and  $A$ . By the same argument regularized KSIR (11) can be solved by minimizing the following penalized least square loss functional with respect to  $C$  and  $A$

$$\mathbf{G}_s(C, A) = \mathbf{G}(C, A) + n_s \text{vec}(CA)^T (D \otimes \frac{1}{n} K^2) \text{vec}(CA)$$

where  $D = \text{diag}(\frac{n_1}{n}, \dots, \frac{n_H}{n})$  and  $\text{vec}(\cdot)$  is a matrix operation that stacks all columns of a matrix. Once either  $A$  or  $C$  is fixed the above functional is a regularized least square problem, suggesting iteratively solving for  $A$  and  $C$ . Given an estimate  $\hat{C} = (\hat{c}_1, \dots, \hat{c}_d)$  the solution coincides with the formula as given by equation (11).

One motivation for the least square formulation in Li and Yin [2008a] is selection of the regularization parameter by a generalized cross-validation criterion. The criterion suggested by Li and Yin [2008a] can be written as follows for the regularized KSIR model

$$GCV = \frac{\|(I_{nH} - Q_s)W^{1/2}\tilde{Z}\|}{nH[1 - \text{trace}(Q_s)/nH]^2}$$

where  $W = D \otimes I_n$ ,  $\tilde{Z} = \text{vec}(\bar{Z}_1, \dots, \bar{Z}_H)$ , and

$$Q_s = \left( D^{1/2} \hat{A}^T (\hat{A} D \hat{A}^T)^{-1} \hat{A} D^{1/2} \right) \otimes \left( \frac{1}{n^2} K (K^2 + n^2 s I_n) K \right).$$

For linear SIR there are many methods to estimate the correct number e.d.r. directions in terms of a testing problem [Schott 1994, Ferré 1998, Bura and Cook 2001, Zhu *et al.* 2006]. Considering KSIR as linear in the gram matrix might allow us to select the correct number of e.d.r. directions using the same tests.

## 4 Consistency of regularized KSIR

In this section, we prove the asymptotic consistency of the e.d.r. directions estimated by KSIR and provide conditions under which the rate of consistency is  $O_p(n^{-1/4})$ . Our result also implies consistency for functional SIR with Tikhonov regularization. An important observation

from the proof is that the rate of convergence of the e.d.r. directions depends on the contribution of the smaller 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 [Hsing and Carroll 1992, Saracco 1997, Zhu and Ng 1995]. 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 Ferré and Yao [2003] 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 Ferré and Villa [2006], but it is of a weaker form than our result. This suggests that either there is a theoretical advantage for Tikhonov regularization over ridge regression or the previous consistency results for functional SIR can be improved.

The following theorem states the formal consistency result.

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

$$|\langle \hat{\beta}_j, \phi(\cdot) \rangle - \langle \beta_j, \phi(\cdot) \rangle| = o_p(1), \quad j = 1, \dots, d_\Gamma,$$

where  $d_\Gamma$  is the rank of  $\Gamma$ ,  $\langle \beta_j, \phi(\cdot) \rangle$  is the projection onto the  $j$ -th e.d.r., and  $\langle \hat{\beta}_j, \phi(\cdot) \rangle$  is the projection onto the  $j$ -th 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  $\Sigma$  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 3.** First define for  $N \geq 1$  the projection operator and its complement

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

where  $\{e_i\}_{i=1}^{\infty}$  are the eigenvectors of the covariance operator  $\Sigma$  as defined in (4), with the corresponding eigenvalues denoted by  $w_i$ .

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

$$\|(\hat{\Sigma} + sI)^{-1} \hat{\Sigma} \hat{\Gamma} - T\|_{HS} = O_p\left(\frac{1}{s\sqrt{n}}\right) + \sum_{j=1}^{d_\Gamma} \left(\frac{s}{w_N^2} \|\Pi_N(\tilde{u}_j)\| + \|\Pi_N^\perp(\tilde{u}_j)\|\right) \quad (14)$$

where  $\tilde{u}_j = \Sigma^{-1} u_j$  and  $\{u_j\}_{j=1}^{d_\Gamma}$  are the eigenvectors of  $\Gamma$  as defined in (8).

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

$$\|(\hat{\Sigma} + sI)^{-1} \hat{\Sigma} \hat{\Gamma} - T\|_{HS} = o_p(1).$$

## 5 Application to simulated and real data

In this section we compare classical sliced inverse regression (SIR), regularized sliced inverse regression (RSIR) as in (13), kernel sliced inverse regression (KSIR), and regularized kernel sliced inverse regression (RKSIR). The comparisons are used to address two questions: (1) does regularization improve the performance of kernel sliced inverse regression, and (2) in real data does the nonlinearity of kernel sliced inverse regression improve predictive performance.

We show using three examples that regularization does help with estimating the e.d.r. directions and the nonlinearity introduced by the kernel versions can make a dramatic difference in terms of predictive accuracy.

### 5.1 Importance of nonlinearity and regularization

This example illustrates that both the nonlinearity and regularization of RKSIR can significantly improve prediction accuracy.

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

$$Y = \left(1 + \frac{1}{5} \sum_{i=1}^5 \sin(\pi X_i)\right) \left(1 + \frac{1}{5} \sum_{i=6}^{10} \sin(\pi X_i)\right) + \varepsilon,$$

where each of the ten dimensions are uniform in  $[-1, 1]$ ,  $X_i \sim U[-1, 1]$  for  $i = 1, \dots, 10$ , and the noise is normal  $\varepsilon \sim N(0, .1^2)$ .

We examined the predictive accuracy as a function of the number of e.d.r. directions used in a Gaussian kernel regression model the bandwidth parameter of the Gaussian was set to the median of pairwise distances. We compared the performance of using SIR, KSIR, and RKSIR to compute the e.d.r. directions in this procedure. We used a training set of 100 observations to compute the e.d.r. directions and fit the nonlinear regression model. We then used an independent test set of 2000 observations to compute the mean square error of the regression model estimated. For both KSIR and RKSIR we used an additive Gaussian kernel

$$k(x, z) = \sum_{j=1}^d \exp(-(x_j - z_j)^2 / 2\sigma^2).$$

The regularization parameter in RKSIR was set by cross-validation.

Figure 1 displays the accuracy of the procedure as a function of the number of e.d.r. directions for the three methods. The result for SIR shows that linear directions do not capture the predictive structure in this data this structure is distributed across all ten e.d.r. directions. The result for KSIR illustrates that by adding the nonlinearity most of the predictive structure is captured in the first e.d.r. direction. However, this direction is far from optimal. For RKSIR

the first e.d.r. direction contains almost all of the predictive structure and adding regularization has greatly improved accuracy.

## 5.2 Effect of regularization

This example illustrates how regularization has an effect on the performance of KSIR as a function of the anisotropy of the predictors.

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

$$Y = X_1 + X_2^2 + \varepsilon, \quad \varepsilon \sim N(0, 0.1^2), \quad (15)$$

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

For this model it is known that SIR will not accurately infer the e.d.r. space since only the first variable  $X_1$  will be identified. For this reason we focus on the comparison of KSIR and RKSIR in this example.

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

$$\Phi(X) = \{1, X_i, (X_i X_j)_{i \leq j}\} \quad i, j = 1, \dots, 10.$$

In this feature space  $X_1 + X_2^2$  can be captured in one e.d.r. direction. Thus using the polynomial kernel a one dimensional e.d.r. space should contain all the predictive information.

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

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

So for this example given estimates of KSIR variates at the  $n$  data points  $\{\hat{v}_i\}_{i=1}^n = \{\langle \hat{\beta}_1, \phi(x_i) \rangle\}_{i=1}^n$  the error of the first e.d.r. direction can be measured from the following optimization problem

$$\begin{aligned} \text{error} &= \min_{a \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n (\hat{v}_i - \mathbb{E}v - a(x_{i,1} + x_{i,2}^2 - \mathbb{E}(X_1 + X_2^2)))^2 \\ &= \min_{a, b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n (\hat{v}_i - (a(x_{i,1} + x_{i,2}^2) + b))^2. \end{aligned}$$

We drew 200 observations from the model specified in (15). We then 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  $\theta$  increase and regularization helps to reduce this instability.

### 5.3 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) [Baudata and Anouar 2000] have been used to construct nonlinear discriminant surfaces and regularization has improved performance of KFDA [Kurita and Taguchi 2005]. In this example we show that this idea of adding a nonlinearity and a regularization term improves predictive accuracy in a real multi-class 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  $\{0, 1, 2, \dots, 9\}$  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.

We compared RSIR, 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 bandwidth 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 is reported in Table 1. The first interesting observation is that 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.

## 6 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 these methods are developed in the unsupervised learning framework. Therefore the estimated dimensions may not be optimal for regression models. Incorporating nonlinearity and regularization to inverse regression approaches results in a robust response driven nonlinear dimension reduction method. A least square formulation is also provided for parameter selection.

There are some interesting relations between KSIR and functional SIR. 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 computations involved in functional SIR can be simplified by a parametrization with respect to a RKHS or using a linear kernel in the parametrized function space. From a theoretical point of view, KSIR can be viewed as a special case of functional SIR as developed by Ferré and his coauthors in a series of papers [Ferré and Yao 2003, 2005, Ferré and Villa 2006].

## Acknowledgments

We acknowledge 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

- BACH, F. R. and JORDAN, M. I. (2002). Kernel independent component analysis. *Journal of Machine Learning Research* **3** 1–48.
- BAUDATA, G. and ANOUAR, F. (2000). Generalized discriminant analysis using a kernel approach. *Neural Computation* **12** 2385–2404.
- BERNARD-MICHEL, C., GARDES, L. and GIRARD, S. (2008). Gaussian regularized sliced inverse regression. Preprint.
- BLANCHARD, G., BOUSQUET, O. and ZWALD, L. (2007). Statistical properties of kernel principal component analysis. *Mach. Learn.* **66** 259–294.
- BURA, E. and COOK, R. D. (2001). Extending sliced inverse regression: the weighted chi-squared test. *J. Amer. Statist. Assoc.* **96** 996–1003.
- CHATELIN, F. (1983). *Spectral Approximation of Linear Operators*. Academic Press.
- COOK, R. and WEISBERG, S. (1991). Discussion of li (1991). *J. Amer. Statist. Assoc.* **86** 328–332.
- COOK, R. D. (2004). Testing predictor contributions in sufficient dimension reduction. *Ann. Statist.* **32** 1062–1092.

- DUAN, N. and LI, K. (1991). Slicing regression: a link-free regression method. *Ann. Stat.* **19** 505–530.
- FERRÉ, L. (1998). Determining the dimension in sliced inverse regression and related methods. *J. Amer. Statist. Assoc.* **93** 132–140.
- FERRÉ, L. and VILLA, N. (2006). Multilayer perceptron with functional inputs: an inverse regression approach. *Scandinavian Journal of Statistics* **33** 807–823.
- FERRÉ, L. and YAO, A. (2003). Functional sliced inverse regression analysis. *Statistics* **37** 475–488.
- FERRÉ, L. and YAO, A. (2005). Smoothed functional inverse regression. *Statist. Sinica* **15** 665–683.
- HALL, P. and LI, K.-C. (1993). On almost linearity of low-dimensional projections from high-dimensional data. *Ann. Statist.* **21** 867–889.
- HE, G., MÜLLER, H. and WANG, J. (2003). Functional canonical analysis for square integrable stochastic processes. *J. Multivariate Anal.* **85** 54–77.
- HSING, T. and CARROLL, R. J. (1992). An asymptotic theory for sliced inverse regression. *Ann. Statist.* **20** 1040–1061.
- KATO, T. (1966). *Perturbation Theory for Linear Operators*. Springer-Verlag, Berlin, Heidelberg, New York.
- KÖNIG, H. (1986). *Eigenvalue distribution of compact operators, Operator Theory: Advances and Applications*, vol. 16. Birkhäuser, Basel, CH.
- KURITA, T. and TAGUCHI, T. (2005). A kernel-based fisher discriminant analysis for face detection. *IEICE TRANS. INF. & SYST.* **E88CD** 628–635.
- LI, K. (1991). Sliced inverse regression for dimension reduction (with discussion). *J. Amer. Statist. Assoc.* **86** 316–342.
- LI, L., COOK, R. and TSAI, C.-L. (2007). Partial inverse regression. *Biometrika* **94** 615–625.
- LI, L. and YIN, X. (2008a). Rejoinder to “a note on sliced inverse regression with regularizations” Preprint.
- LI, L. and YIN, X. (2008b). Sliced inverse regression with regularizations. *Biometrics* **64** 124–131.

- MERCER, J. (1909). Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society, London A* **209** 415–446.
- SARACCO, J. (1997). An asymptotic theory for sliced inverse regression. *Comm. Statist. Theory Methods* **26** 2141–2171.
- SCHÖLKOPF, B. and SMOLA, A. J. (2002). *Learning with kernels*. MIT Press, MA.
- SCHÖLKOPF, B., SMOLA, A. J. and MÜLLER, K. (1997). Kernel principal component analysis. In *Artificial Neural Networks ICANN'97* (W. Gerstner, A. Germond, M. Hasler and J.-D. Nicoud, eds.), *Springer Lecture Notes in Computer Science*, vol. 1327, 583–588. Berlinpp.
- SCHOTT, J. R. (1994). Determining the dimensionality in sliced inverse regression. *J. Amer. Statist. Assoc.* **89** 141–148.
- WU, H.-M. (2008). Kernel sliced inverse regression with applications on classification. *Journal of Computational and Graphical Statistics* Accepted.
- ZHONG, W., ZENG, P., MA, P., LIU, J. S. and ZHU, Y. (2005). RSIR: regularized sliced inverse regression for motif discovery. *Bioinformatics* **21** 4169–4175.
- ZHU, L., MIAO, B. and PENG, H. (2006). On sliced inverse regression with high-dimensional covariates. *J. Amer. Statist. Assoc.* **101** 630–643.
- ZHU, L. X. and NG, K. W. (1995). Asymptotics of sliced inverse regression. *Statist. Sinica* **5** 727–736.

## Appendix A: proof of Theorem 1

Under the assumption of Proposition 1, for each  $Y = y$ ,

$$\mathbb{E}[\phi(X)|Y = y] \in \text{span}\{\Sigma\beta_i, i = 1, \dots, d\}. \quad (16)$$

So the rank of  $\Gamma$  (i.e., the dimension of the image of  $\Gamma$ ) is less than  $d$ . Since this implies  $\Gamma$  is compact, together with the fact it is symmetric and positive, there exist  $d_\Gamma$  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  $f \in \mathcal{H}$ ,

$$\Gamma f = \langle \mathbb{E}[\phi(X)|Y], f \rangle \mathbb{E}[\phi(X)|Y]$$

also belongs to

$$\text{span}\{\Sigma\beta_i, i = 1, \dots, d\} \subset \text{range}(\Sigma)$$

because of (16), so

$$u_i = \frac{1}{\tau_i} \Gamma u_i \in \text{range}(\Sigma).$$

This proves (i).

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

$$Tf = \Sigma \left( \sum_{i=1}^{d_\Gamma} \langle u_i, f \rangle u_i \right) = \sum_{i=1}^{d_\Gamma} \langle u_i, f \rangle \Sigma(u_i) = \left( \sum_{i=1}^{d_\Gamma} \Sigma(u_i) \otimes u_i \right) f.$$

This proves (ii).

## Appendix B: Proof of Proposition 3

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

Suppose  $\Phi$  has the following SVD decomposition

$$\Phi = UDV^T = [u_1 \dots u_{d_K}] \begin{bmatrix} \bar{D}_{d \times d} & \mathbf{0}_{(n-d) \times (d_K-d)} \\ \mathbf{0}_{(n-d) \times d} & \mathbf{0}_{(n-d) \times (d_K-d)} \end{bmatrix} \begin{bmatrix} v_1^T \\ \vdots \\ v_n^T \end{bmatrix} = \bar{U} \bar{D} \bar{V}^T, \quad (17)$$

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

We need to show the KSIR variates

$$\hat{v}_j = c_j^T [k(x, x_1), \dots, k(x, x_n)] = c_j^T \Phi^T \Phi(x) = \langle \Phi c_j, \phi(x) \rangle = \langle \beta_j, \phi(x) \rangle.$$

It suffices to prove that if  $(\lambda, c)$  is a solution to (11), then  $(\lambda, \beta)$  is also a pair of eigenvalue and eigenvector of  $(\hat{\Sigma}^2 + \gamma I)^{-1} \hat{\Sigma} \hat{\Gamma}$  and vice versa, where  $c$  and  $\beta$  is related by

$$\beta = \Phi c \text{ and } c = \bar{V} \bar{D} \bar{U}^T \beta.$$

Noting that facts  $\hat{\Sigma} = \frac{1}{n} \Phi \Phi^T$ ,  $\hat{\Gamma} = \frac{1}{n} \Phi J \Phi^T$ , and  $K = \Phi^T \Phi = \bar{V} \bar{D}^2 \bar{V}^T$ , the argument may be made as follows:

$$\begin{aligned} \hat{\Sigma} \hat{\Gamma} \beta = \lambda (\hat{\Sigma}^2 + \gamma I) \beta &\iff \Phi \Phi^T \Phi J \Phi^T \Phi c = \lambda (\Phi \Phi^T \Phi \Phi^T \Phi c + n^2 \gamma \Phi c) \\ &\iff \Phi K J K c = \lambda \Phi (K^2 + n^2 \gamma) c \\ &\iff \bar{V} \bar{V}^T K J K c = \lambda \bar{V} \bar{V}^T (K^2 + n^2 \gamma I) c \\ &\iff K J K c = \lambda (K^2 + n^2 \gamma I) c. \end{aligned}$$

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

$$\bar{V} \bar{V}^T K = \bar{V} \bar{V}^T \bar{V} \bar{D}^2 \bar{V}^T = \bar{V} \bar{D}^2 \bar{V}^T = K$$

and

$$\bar{V} \bar{V}^T c = \bar{V} \bar{V}^T \bar{V} \bar{D}^{-1} \bar{U}^T \beta = \bar{V} \bar{D}^{-1} \bar{U}^T \beta = c.$$

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

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

$$\hat{\Sigma} = \sum_{i=1}^{d_K} \hat{\sigma}_i u_i \otimes u_i = \sum_{i=1}^d \sigma_i u_i \otimes u_i$$

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

## Appendix C: Proof of Consistency

### C.1 Preliminaries

In order to prove Theorems 2 and Theorem 3 we use properties of Hilbert-Schmidt operators, covariance operators for Hilbert space valued random variables, and perturbation theory for linear operators. In this subsection we provide a brief introduction to them. For details see Kato [1966], Chatelin [1983], Blanchard *et al.* [2007] and references therein.

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

$$\|L\|_{HS}^2 = \sum_{i=1}^{p_{\mathcal{H}}} \|Le_i\|_{\mathcal{H}}^2 < \infty,$$

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

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

$$\|SL\|_{HS} \leq \|S\| \|L\|_{HS}, \quad \|LS\|_{HS} \leq \|L\|_{HS} \|S\|$$

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

$$\|L\|^2 = \sup_{f \in \mathcal{H}} \frac{\|Lf\|^2}{\|f\|^2}.$$

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

$$\Sigma = \mathbb{E}[(Z - \mathbb{E}Z) \otimes (Z - \mathbb{E}Z)],$$

is self-adjoint, positive, compact, and belongs to 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 3.

We will use the following result from Ferré and Yao [2003]

$$\|\hat{\Sigma} - \Sigma\|_{HS} = O_p(1/\sqrt{n}) \quad \text{and} \quad \|\hat{\Gamma} - \Gamma\|_{HS} = O_p(1/\sqrt{n}).$$

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

$$T_1 = (\hat{\Sigma}^2 + sI)^{-1} \hat{\Sigma} \hat{\Gamma} \quad \text{and} \quad T_2 = (\Sigma^2 + sI)^{-1} \Sigma \Gamma.$$

Then

$$\|\hat{T}_s - T\|_{HS} \leq \|\hat{T}_s - T_1\|_{HS} + \|T_1 - T_2\|_{HS} + \|T_2 - T\|_{HS}.$$

For the first term observe that

$$\|\hat{T}_s - T_1\|_{HS} \leq \|(\hat{\Sigma}^2 + sI)^{-1}\| \|\hat{\Sigma}\hat{\Gamma} - \Sigma\Gamma\|_{HS} = O_p\left(\frac{1}{s\sqrt{n}}\right).$$

For the second term note that

$$T_1 = \sum_{j=1}^{d_\Gamma} \tau_j \left( (\hat{\Sigma}^2 + sI)^{-1} \Sigma u_j \right) \otimes u_j \quad \text{and} \quad T_2 = \sum_{j=1}^{d_\Gamma} \tau_j \left( (\Sigma^2 + sI)^{-1} u_j \right) \otimes u_j.$$

Therefore

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

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

$$\left( (\Sigma^2 + sI)^{-1} - (\hat{\Sigma}^2 + sI)^{-1} \right) \Sigma u_j = (\hat{\Sigma}^2 + sI)^{-1} \left( \hat{\Sigma}^2 - \Sigma^2 \right) (\Sigma^2 + sI)^{-1} \Sigma^2 \tilde{u}_j.$$

which implies

$$\begin{aligned} \|T_1 - T_2\| &\leq \sum_{j=1}^{d_\Gamma} \tau_j \left\| (\hat{\Sigma}^2 + sI)^{-1} \right\| \left\| \hat{\Sigma}^2 - \Sigma^2 \right\|_{HS} \left\| (\hat{\Sigma}^2 + sI)^{-1} \Sigma^2 \right\| \|\tilde{u}_j\| \\ &= O_p\left(\frac{1}{s\sqrt{n}}\right). \end{aligned}$$

For the third term the following holds

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

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

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

Combining these terms results in (14).

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

$$\|\hat{T}_s - T\|_{HS} = o_p(1)$$

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

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

$$\tilde{u}_j = \Sigma^{-1}u_j \in \Sigma^{-1}(\mathcal{S}^*) \subset \text{span}\{e_i, i = 1, \dots, N\}.$$

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

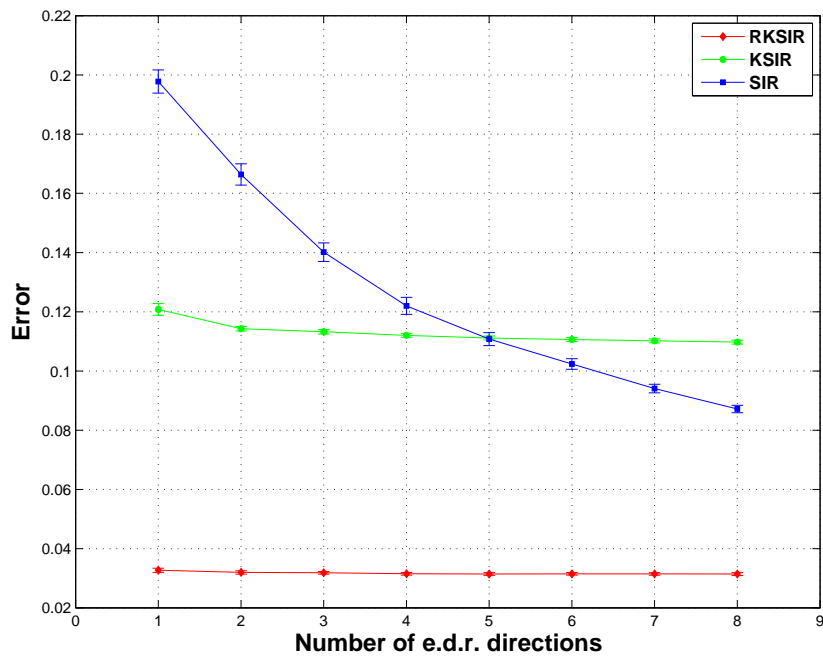


Figure 1: Mean square error on test data versus number of e.d.r. directions for the product of sines. The blue curve is a plot of the mean and standard error for SIR, the red curve is the same for RKSIR, and the green curve is for KSIR.

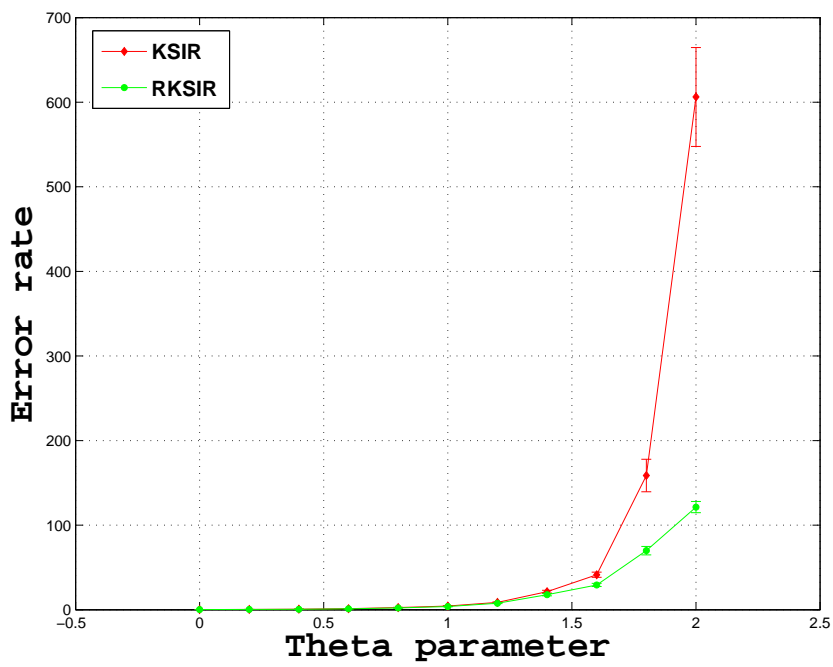


Figure 2: Error in e.d.r. as a function of  $\theta$ .

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)

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