Gradient-based kernel dimension reduction for regression

Kenji Fukumizu*and Chenlei Leng†

August 23, 2013

Abstract

This paper proposes a novel approach to linear dimension reduction for regression using nonparametric estimation with positive definite kernels or reproducing kernel Hilbert spaces. The purpose of the dimension reduction is to find such directions in the explanatory variables that explain the response sufficiently: this is called *sufficient dimension reduction*. The proposed method is based on an estimator for the gradient of regression function considered for the feature vectors mapped into reproducing kernel Hilbert spaces. It is proved that the method is able to estimate the directions that achieve sufficient dimension reduction. In comparison with other existing methods, the proposed one has wide applicability without strong assumptions on the

*The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562 Japan

†Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK, and Department of Statistics and Applied Probability, National University of Singapore, 6 Science Drive 2, Singapore, 117546

distributions or the type of variables, and needs only eigendecomposition for estimating the projection matrix. The theoretical analysis shows that the estimator is consistent with certain rate under some conditions. Experimental results demonstrate that the proposed method successfully finds effective directions with efficient computation even for high dimensional explanatory variables.

1 Introduction

Recent data analysis often handles high dimensional data, which may be given by images, texts, genomic expressions, and so on. Dimension reduction is almost always involved in such data analysis for avoiding various problems caused by the high dimensionality; they are known as *curse of dimensionality*. The purpose of dimension reduction thus includes preprocessing for another data analysis aiming at less expensive computation in later processing, noise reduction by suppressing noninformative directions, and construction of readable low dimensional expressions such as visualization.

This paper discusses dimension reduction for regression, where X is an explanatory variable in $\mathbb{R}^m$ and Y is a response variable. The domain of Y is arbitrary, either continuous or discrete. The purpose of dimension reduction in this setting is to find such features of X that explain Y as effectively as possible. This paper focuses on linear dimension reduction, in which linear combinations of the components of X are used to make effective features.

Beyond the classical approaches such as reduced rank regression and canonical correlation analysis, which can be used for extracting linear features straightforwardly, a modern approach to this problem is based on

the *sufficient dimension reduction* (Cook, 1994, 1998), which formulates the problem by conditional independence. More precisely, assuming

$$p(Y|X) = \tilde{p}(Y|B^T X) \qquad \text{or equivalently} \qquad Y \perp\!\!\!\perp X \,|\, B^T X \tag{1}$$

for the distribution, where $p(Y|X)$ and $\tilde{p}(Y|B^T X)$ are respective conditional probability density functions, and B is a projection matrix ($B^T B = I_d$, where I_d is the unit matrix) onto a d-dimensional subspace ($d < m$) in $\mathbb{R}^m$, we wish to estimate B with a finite sample from that distribution. The subspace spanned by the column vectors of B is called the *effective dimension reduction (EDR) space* (Li, 1991). We consider nonparametric methods of estimating B without assuming any specific parametric models for $p(y|x)$.

The first method that aims at finding the EDR space is the *sliced inverse regression* (SIR, Li, 1991), which employs the fact that the inverse regression $E[X|Y]$ distributes in the EDR space under some assumptions. Many methods have been proposed in this vein of inverse regression such as SAVE (Cook and Weisberg, 1991), directional regression (Li and Wang, 2007) and contour regression (Li et al., 2005), which use statistics such as mean and variance in each slice or contour of Y. While many inverse regression methods are computationally simple, they often need some strong assumptions on the distribution of X such as elliptic symmetry. Additionally, many methods such as slice-based methods assume that Y is a real valued random variable, and thus are not suitable for multidimensional or discrete responses.

Other interesting approaches to the linear dimension reduction include the minimum average variance estimation (MAVE, Xia et al., 2002), in

which the conditional variance of the regression in the direction of $B^T X$, $E[(Y - E[Y|B^T X])^2|B^T X]$, is minimized with the conditional variance estimated by the local linear kernel smoothing method. The kernel smoothing method requires, however, careful choice of the bandwidth parameter in the kernel, and it is usually difficult to apply if the dimensionality is very high. Additionally, the iterative computation of MAVE is expensive for large data set. Another recent approach uses support vector machines for linear and nonlinear dimension reduction, which estimates the EDR directions by the normal direction of classifiers for the classification problems given by slicing the response variable (Li et al., 2011).

The most relevant to this paper is the methods based on the gradient of regressor $\varphi(x) = E[Y|X = x]$ (Samarov, 1993, Hristache et al., 2001). As detailed in Section 2.1, under Eq. (1) the gradient of $\varphi(x)$ is contained in the EDR space at each x. One can thus estimate B by nonparametric estimation of the gradients. There are, however, some limitations in this method: the nonparametric estimation of the gradient in high-dimensional spaces is challenging as in MAVE, and if the conditional variance of Y is dependent on X, the method is not able to extract that direction.

This paper proposes a novel approach to sufficient dimension reduction with positive definite kernels. Positive definite kernels or reproducing kernel Hilbert spaces have been widely used for data analysis (Wahba, 1990), especially since the success of the support vector machine (Boser et al., 1992, Hofmann et al., 2008). The methods, in short, extract nonlinear features or higher-order moments of data by transforming them into reproducing kernel Hilbert spaces (RKHSs) defined by positive definite kernels. Various methods for nonparametric inference also have been recently developed in this

discipline (Gretton et al., 2005b, 2008, 2009, Fukumizu et al., 2008).

A method for linear dimension reduction based on positive definite kernels has been already proposed to overcome various limitations of existing methods. The kernel dimension reduction (KDR, Fukumizu et al., 2004, 2009) uses conditional covariance on RKHSs to characterize the conditional independence relation in Eq. (1). The KDR is a general method applicable to a wide class of problems without requiring any strong assumptions on the distributions or types of the variable X or Y. The involved computation, however, requires the numerical optimization for the nonconvex objective function of the projection matrix B, and uses the gradient descent method which needs many inversions of Gram matrices. While KDR shows good estimation accuracy for small data sets, the difficulty in the optimization prohibits applications of KDR to very high-dimensional or large-size data. Another relevant method using RKHS is the kernel sliced inverse regression (Wu, 2008). This method, however, considers nonlinear extension of SIR with the feature map, and differs from linear dimension reduction, which is the focus of the current paper. Additionally, with RKHSs, Hsing and Ren (2009) discuss an extension of inverse regression from finite-dimensional to infinite-dimensional problems.

The method proposed in this paper uses the approach by the gradient of regression function. Unlike the existing ones (Samarov, 1993, Hristache et al., 2001), the gradient is estimated nonparametrically by the covariance operators on RKHS, which is based on the recent development in the kernel method (Fukumizu et al., 2009, Song et al., 2009). The proposed method solves the problems of existing ones: by virtue of the kernel method the sufficient dimension reduction is realized without any strong assumption

on the regressor and probability distributions, the response Y can be of arbitrary type, and the kernel estimation of the gradient is stable without elaborate tuning of bandwidth. It solves also the computational problem in the KDR: the estimator is given by the solution of an eigenproblem with no need of numerical optimization. The method is thus applicable to large and high-dimensional data, as we will demonstrate with numerical examples.

This paper is organized as follows. In Section 2, after giving a review on the gradient-based method for dimension reduction and a brief explanation of the statistical method with positive definite kernels, we will introduce the kernel method for gradient-based dimension reduction. Some discussions and theoretical results are also shown. Section 3 demonstrates the performance of the method with some artificial and real world data sets. Section 4 concludes this paper. The technical proofs of the theoretical results are shown in Appendix.

2 Gradient-based kernel dimension reduction

In this paper, it is assumed that all Hilbert spaces are separable. The range of an operator A is denoted by $\mathcal{R}(A)$, and the Frobenius norm of a matrix M by $\|M\|_F$.

2.1 Gradient of regression function and dimension reduction

We first review the basic idea of the gradient-based method for dimension reduction in regression, which has been used in Samarov (1993) and Hristache et al. (2001). Suppose Y is a real-valued random variable such that the regression function $E[Y|X = x]$ is differentiable with respect to x. We

assume that Eq. (1) holds, and wish to estimate the projection matrix B using i.i.d. sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. Under Eq. (1), it is easy to see

$$\frac{\partial}{\partial x} E[Y|X=x] = \frac{\partial}{\partial x} \int y p(y|x) dy$$
$$= \int y \frac{\partial \tilde{p}(y|B^T x)}{\partial x} dy = B \int y \left. \frac{\partial \tilde{p}(y|z)}{\partial z} \right|_{z=B^T x} dy,$$

where exchangeability of the differentiation and integration is assumed. The above equation implies that the gradient $\partial E[Y|X=x]/\partial x$ at any x is contained in the EDR space. Based on this necessary condition, the average derivative estimates (ADE, Samarov, 1993) has been proposed to use the average of the gradients at X_i for estimating B.

In the more recent method (Hristache et al., 2001), the EDR space is estimated by the principal component analysis for the gradient estimates, which are given by the standard local linear least squares with a smoothing kernel (Fan and Gijbels, 1996). Additionally, the contribution of the estimated projector is gradually increased in the iterative procedure so that the dimensionality of data is continuously reduced to a desired one. This iterative procedure is expected to alleviate the difficulty of estimating the gradients in a high dimensional space. We call the method in Hristache et al. (2001) the iterative average derivative estimates (IADE) in the sequel.

Note that the methods based on the gradient of the regression function $E[Y|X=x]$ use only a necessary condition of Eq. (1), and not sufficient in general. In fact, if the variables $X = (X^1, \ldots, X^m)$ and Y follow $Y = f(X^1) + N(0, \sigma(X^2)^2)$, where $f(x^1)$ and $\sigma(x^2)$ are some fixed functions, the conditional probability $p(y|x)$ depends on x^1 and x^2, while the regression $E[Y|X]$ depends only on x^1. The existing gradient-based methods fail to find the direction x^2. In contrast, the method proposed in this paper avoids

this obvious limitation of the gradient-based methods by virtue of nonlinear feature mapping given by positive definite kernels, while keeping tractable computational cost.

2.2 Kernel method for conditional mean

Positive definite kernels or reproducing kernel Hilbert spaces have been extensively applied to data analysis especially since the success of the support vector machine in classification problems (Wahba, 1990, Schölkopf and Smola, 2002, Hofmann et al., 2008). More recently, it has been revealed that kernel methods can be applied to statistical problems through representing distributions in the form of means and covariances in RKHS (Fukumizu et al., 2004, 2009, Song et al., 2009), which is briefly reviewed below.

For a set Ω, a ($\mathbb{R}$-valued) positive definite kernel k on Ω is a symmetric kernel $k : \Omega \times \Omega \to \mathbb{R}$ such that $\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \geq 0$ for any $x_1, \ldots, x_n$ in Ω and $c_1, \ldots, c_n \in \mathbb{R}$. It is known (Aronszajn, 1950) that a positive definite kernel on Ω is uniquely associated with a Hilbert space $\mathcal{H}$ consisting of functions on Ω such that (i) $k(\cdot, x)$ is in $\mathcal{H}$, (ii) the linear hull of $\{k(\cdot, x) \mid x \in \Omega\}$ is dense in $\mathcal{H}$, and (iii) for any $x \in \Omega$ and $f \in \mathcal{H}$, $\langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$, where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is the inner product of $\mathcal{H}$. The property (iii) is called *reproducing property*, and the Hilbert space $\mathcal{H}$ the *reproducing kernel Hilbert space* (RKHS) associated with k.

Let $(\mathcal{X}, \mathcal{B}_{\mathcal{X}}, \mu_{\mathcal{X}})$ and $(\mathcal{Y}, \mathcal{B}_{\mathcal{Y}}, \mu_{\mathcal{Y}})$ be measure spaces, and (X, Y) be a random variable on $\mathcal{X} \times \mathcal{Y}$ with probability distribution P. Let $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ be measurable positive definite kernels on $\mathcal{X}$ and $\mathcal{Y}$, respectively, with respective RKHS $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$. It is assumed that $E[k_{\mathcal{X}}(X, X)]$ and $E[k_{\mathcal{Y}}(Y, Y)]$ are finite. The (uncentered) *cross-covariance operator* $C_{YX} : \mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{Y}}$ is

defined as the operator such that

$$\langle g, C_{YX} f \rangle_{\mathcal{H}_\mathcal{Y}} = E[f(X)g(Y)] = E\big[\langle f, \Phi_\mathcal{X}(X)\rangle_{\mathcal{H}_\mathcal{X}} \langle \Phi_\mathcal{Y}(Y), g\rangle_{\mathcal{H}_\mathcal{Y}}\big] \qquad (2)$$

holds for all $f \in \mathcal{H}_\mathcal{X}, g \in \mathcal{H}_\mathcal{Y}$, where $\Phi_\mathcal{X} : \mathcal{X} \to \mathcal{H}_\mathcal{X}$ and $\Phi_\mathcal{Y} : \mathcal{Y} \to \mathcal{H}_\mathcal{Y}$ are defined by $x \mapsto k_\mathcal{X}(\cdot, x)$ and $y \mapsto k_\mathcal{Y}(\cdot, y)$, respectively. Similarly, C_{XX} denotes the operator on $\mathcal{H}_\mathcal{X}$ that satisfies $\langle f_2, C_{XX} f_1\rangle = E[f_2(X)f_1(X)]$ for any $f_1, f_2 \in \mathcal{H}_\mathcal{X}$. These definitions are straightforward extensions of the ordinary covariance matrices on Euclidean spaces, as C_{YX} is the covariance of the random vectors $\Phi_\mathcal{X}(X)$ and $\Phi_\mathcal{Y}(Y)$ on RKHSs. Although C_{YX} and C_{XX} depend on the kernels, we omit the dependence in the notation for simplicity.

With $g = k_\mathcal{Y}(\cdot, y)$ in Eq. (2), the reproducing property derives

$$(C_{YX} f)(y) = \int k_\mathcal{Y}(y, \tilde{y}) f(\tilde{x}) dP(\tilde{x}, \tilde{y})$$

and

$$(C_{XX} f)(x) = \int k_\mathcal{X}(x, \tilde{x}) f(\tilde{x}) dP_X(\tilde{x}),$$

where P_X is the marginal distribution of X. These equations show the explicit expressions of C_{YX} and C_{XX} as integral operators.

An important notion in statistical inference with positive definite kernels is the characteristic property. A bounded measurable positive definite kernel k (with RKHS $\mathcal{H}$) on a measurable space $(\Omega, \mathcal{B})$ is called *characteristic* if the mapping from a probability Q on $(\Omega, \mathcal{B})$ to the mean $E_{X\sim Q}[k(\cdot, X)] \in \mathcal{H}$ of the $\mathcal{H}$-valued random variable $\Phi(X) = k(\cdot, X)$ is injective (Fukumizu et al., 2004, 2009, Sriperumbudur et al., 2010). This is equivalent to assuming that $E_{X\sim P}[k(\cdot, X)] = E_{X'\sim Q}[k(\cdot, X')]$ implies $P = Q$, that is, probabilities are uniquely determined by their means on the associated RKHS. Intuitively,

with a characteristic kernel, the nonlinear function $x \mapsto E[k(x, X)]$ represents a variety of moments enough to determine the underlying probability. Popular examples of characteristic kernel on an Euclidean space are the Gaussian RBF kernel $k(x, y) = \exp(-\|x - y\|^2/(2\sigma^2))$ and Laplace kernel $k(x, y) = \exp(-\alpha \sum_{i=1}^m |x_i - y_i|)$. It is also known (Fukumizu et al., 2009) that a positive definite kernel on a measurable space $(\Omega, \mathcal{B})$ with corresponding RKHS $\mathcal{H}$ is characteristic if and only if $\mathcal{H} + \mathbb{R}$ is dense in the space of square integrable functions $L^2(P)$ for arbitrary probability P on $(\Omega, \mathcal{B})$, where $\mathcal{H} + \mathbb{R}$ is the direct sum of two RKHSs $\mathcal{H}$ and $\mathbb{R}$ (Aronszajn, 1950).

An advantage of using positive definite kernels is that many quantities can be estimated easily with finite sample by virtue of the reproducing property. Given i.i.d. sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ with law P, the covariance operator is estimated by the empirical covariance operator

$$\widehat{C}_{YX}^{(n)} f = \frac{1}{n} \sum_{i=1}^n k_{\mathcal{Y}}(\cdot, Y_i) \langle k_{\mathcal{X}}(\cdot, X_i), f \rangle_{\mathcal{H}_{\mathcal{X}}} = \frac{1}{n} \sum_{i=1}^n f(X_i) k_{\mathcal{Y}}(\cdot, Y_i). \quad (3)$$

The estimator $\widehat{C}_{XX}^{(n)}$ is given similarly. It is known that these estimators are $\sqrt{n}$-consistent in the Hilbert-Schmidt norm (Gretton et al., 2005a).

The fundamental result in discussing conditional probabilities with positive definite kernels is the following fact.

Theorem 1 (Fukumizu et al. (2004))**.** *If $E[g(Y)|X = \cdot] \in \mathcal{H}_{\mathcal{X}}$ holds for $g \in \mathcal{H}_{\mathcal{Y}}$, then*

$$C_{XX} E[g(Y)|X = \cdot] = C_{XY} g.$$

If C_{XX} is injective, the above relation can be thus expressed as

$$E[g(Y)|X = \cdot] = {C_{XX}}^{-1} C_{XY} g. \quad (4)$$

Noting $\langle C_{XX} f, f \rangle = E[f(X)^2]$, it is easy to see that C_{XX} is injective, if $k_{\mathcal{X}}$ is a continuous kernel on a topological space $\mathcal{X}$, and P_X is a Borel probability measure such that $P(U) > 0$ for any open set U in $\mathcal{X}$. The assumption $E[g(Y)|X = \cdot] \in \mathcal{H}_{\mathcal{X}}$, however, may not hold in general; we can easily make counterexamples with Gaussian RBF kernel and Gaussian distributions. We can nonetheless obtain a regularized empirical estimator of $E[g(Y)|X = \cdot]$ based on Eq. (4), namely,

$$(\widehat{C}_{XX}^{(n)} + \varepsilon_n I)^{-1} \widehat{C}_{XY}^{(n)} g, \tag{5}$$

where ε_n is a regularization coefficient in Thikonov-type regularization. We can prove that Eq. (5) is a consistent estimator of $E[g(Y)|X = \cdot]$ in $L^2(P_X)$-norm even if $E[g(Y)|X = \cdot]$ is not in $\mathcal{H}_{\mathcal{X}}$ but in $L^2(P_X)$, and under the assumption $E[g(Y)|X = \cdot] \in \mathcal{H}_{\mathcal{X}}$, it is consistent in $\mathcal{H}_{\mathcal{X}}$ norm. Furthermore, if $E[g(Y)|X = \cdot] \in \mathcal{R}(C_{XX}^{\nu})$ for $\nu > 0$, it is consistent in $\mathcal{H}_{\mathcal{X}}$ norm of the order $O\big(n^{-\min\{\frac{1}{4}, \frac{\nu}{2\nu+2}\}}\big)$ with $\varepsilon_n = n^{-\max\{\frac{1}{4}, \frac{1}{2\nu+2}\}}$. These facts have been proved in various contexts (e.g. Smale and Zhou, 2005, 2007, Caponnetto and De Vito, 2007, Bauer et al., 2007), so the proof is omitted. Also, this type of regularization has been recently used in combination with some dimension reduction techniques (Zhong et al., 2005, Bernard-Michel et al., 2008).

The estimator Eq. (5) is simply the same as the kernel ridge regression with $g(Y)$ as a response. Note, however, that the operator $(\widehat{C}_{XX}^{(n)} + \varepsilon_n I)^{-1} \widehat{C}_{XY}^{(n)}$ includes the information on the regression with various nonlinear transform of Y simultaneously. With a characteristic kernel, this will provide sufficient dimension reduction rigorously as we see in Section 2.3.3.

Beyond the estimation of regression functions, the dimension reduction

method discussed in Section 2.1 requires to estimate the gradient of the regression function. It is known (e.g., Steinwart and Christmann, 2008, Section 4.3) that if a positive definite kernel $k(x, y)$ on an open set in the Euclidean space is continuously differentiable with respect to x and y, every f in the corresponding RKHS is continuously differentiable, and if further $\partial k(\cdot, x)/\partial x \in \mathcal{H}_\mathcal{X}$, the relation

$$\frac{\partial f(x)}{\partial x} = \left\langle f, \frac{\partial}{\partial x} k(\cdot, x) \right\rangle_{\mathcal{H}_\mathcal{X}} \tag{6}$$

holds for any $f \in \mathcal{H}_\mathcal{X}$. Namely, the derivative of any function in that RKHS can be computed in the form of the inner product. This property combined with the estimator Eq. (5) provides our method for dimension reduction.

2.3 Gradient-based kernel dimension reduction

2.3.1 Method

Let (X, Y) be a random vector on $\mathbb{R}^m \times \mathcal{Y}$, where $\mathcal{Y}$ is a measurable space with measure $\mu_\mathcal{Y}$. We prepare positive definite kernels $k_\mathcal{X}$ and $k_\mathcal{Y}$ on $\mathbb{R}^m$ and $\mathcal{Y}$, respectively, with respective RKHS $\mathcal{H}_\mathcal{X}$ and $\mathcal{H}_\mathcal{Y}$. We assume that Eq. (1) holds for some $m \times d$ matrix B with $B^T B = I_d$. It is then easy to see that for any $g \in \mathcal{H}_\mathcal{Y}$ there exists a function $\varphi_g(z)$ on $\mathbb{R}^d$ such that

$$E[g(Y) \mid X] = \varphi_g(B^T X). \tag{7}$$

In fact, we can simply set $\varphi_g(z) = \int g(y)\tilde{p}(y|z)d\mu_\mathcal{Y}$. Note that $g \mapsto \varphi_g(B^T X)$ is a linear functional of $\mathcal{H}_\mathcal{Y}$ for any value of X.

Recall we make the following assumptions

(i) $\mathcal{H}_\mathcal{X}$ and $\mathcal{H}_\mathcal{Y}$ are separable.

(ii) $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ are measurable, and $E[k_{\mathcal{X}}(X,X)] < \infty, E[k_{\mathcal{Y}}(Y,Y)] < \infty$.

In deriving an estimator for B, we further make the following technical assumptions.

(iii) $k_{\mathcal{X}}(\tilde{x}, x)$ is continuously differentiable and $\partial k_{\mathcal{X}}(\cdot, x)/\partial x^i \in \mathcal{R}(C_{XX})$ for $i = 1, \ldots, m$.

(iv) $E[k_{\mathcal{Y}}(y, Y)|X = \cdot] \in \mathcal{H}_{\mathcal{X}}$ for any $y \in \mathcal{Y}$.

(v) $\varphi_g(z)$ in Eq. (7) is differentiable with respect to z, and the linear functional

$$g \mapsto \frac{\partial \varphi_g(z)}{\partial z^a}$$

is continuous for any $z \in \mathbb{R}^d$ and $a = 1, \ldots, d$.

The assumption (iv) implies that $E[g(Y)|X = \cdot] \in \mathcal{H}_{\mathcal{X}}$ for any $g \in \mathcal{H}_{\mathcal{Y}}$. Under Eq. (1), the assumption (v) is true if $C := \int \sqrt{k_{\mathcal{Y}}(y,y)} |\partial \tilde{p}(y|z)/\partial z^a| d\mu_{\mathcal{Y}}(y)$ is finite for any z and the differentiation and integration are exchangeable: in fact, it is easy to see

$$\Big|\frac{\partial \varphi_g(z)}{\partial z^a}\Big| \leq \int \big|\langle g, k_{\mathcal{Y}}(\cdot, y)\rangle\big| \Big|\frac{\partial \tilde{p}(y|z)}{\partial z^a}\Big| d\mu_{\mathcal{Y}}(y) \leq C\|g\|_{\mathcal{H}_{\mathcal{Y}}}.$$

By Riesz' theorem, the assumption (v) implies that there is $\Psi_a(z) \in \mathcal{H}_{\mathcal{X}}$ such that for $a = 1, \ldots, d$,

$$\langle g, \Psi_a(z)\rangle_{\mathcal{H}_{\mathcal{Y}}} = \frac{\partial \varphi_g(z)}{\partial z^a}.$$

We write $\nabla_a \varphi(z)$ for $\Psi_a(z)$, because it is the derivative of the $\mathcal{H}_{\mathcal{Y}}$-valued function $z \mapsto E[k_{\mathcal{Y}}(\cdot, Y)|B^T X = z]$. The relation Eq. (7) then implies that

$$\frac{\partial}{\partial x^i} E[g(Y)|X = x] = \frac{\partial \varphi_g(B^T x)}{\partial x^i} = \sum_{a=1}^{d} B_{ia} \langle g, \nabla_a \varphi(B^T x)\rangle_{\mathcal{H}_{\mathcal{Y}}} \qquad (8)$$

holds for any $g \in \mathcal{H}_\mathcal{Y}$. On the other hand, letting $C_{XX}^{-1}\big(\partial k_\mathcal{X}(\cdot,x)/\partial x^i\big)$ denote the inverse element guaranteed by the assumption (iii), Theorem 1 and Eq. (6) show that for any $g \in \mathcal{H}_\mathcal{Y}$

$$\frac{\partial}{\partial x^i} E[g(Y)|X=x] = \Big\langle C_{XY} g, C_{XX}^{-1} \frac{\partial k_\mathcal{X}(\cdot,x)}{\partial x^i} \Big\rangle = \Big\langle g, C_{YX} C_{XX}^{-1} \frac{\partial k_\mathcal{X}(\cdot,x)}{\partial x^i} \Big\rangle. \tag{9}$$

From Eqs. (8) and (9), we have $C_{YX}C_{XX}^{-1}(k_\mathcal{X}(\cdot,x)/\partial x^i) = \sum_{a=1}^d B_{ia}\nabla_a \varphi(B^T x)$ and thus

$$\begin{aligned}\Big\langle C_{YX} C_{XX}^{-1} \frac{\partial k_\mathcal{X}(\cdot,x)}{\partial x^i}, C_{YX} C_{XX}^{-1} \frac{\partial k_\mathcal{X}(\cdot,x)}{\partial x^j} \Big\rangle_{\mathcal{H}_\mathcal{Y}} \\ = \sum_{a,b=1}^d B_{ia} B_{jb} \langle \nabla_a \varphi(B^T x), \nabla_b \varphi(B^T x) \rangle_{\mathcal{H}_\mathcal{Y}}\end{aligned}$$

for $i,j = 1,\ldots,m$. This means that the eigenvectors for the non-trivial eigenvalues of the $m \times m$ matrix $M(x)$, which is defined by

$$M_{ij}(x) = \Big\langle C_{YX} C_{XX}^{-1} \frac{\partial k_\mathcal{X}(\cdot,x)}{\partial x^i}, C_{YX} C_{XX}^{-1} \frac{\partial k_\mathcal{X}(\cdot,x)}{\partial x^j} \Big\rangle_{\mathcal{H}_\mathcal{Y}}, \tag{10}$$

are contained in the EDR space.

Given i.i.d. sample $(X_1,Y_1),\ldots,(X_n,Y_n)$ from the true distribution, the estimator of $M(x)$ is easily obtained based on Eq. (5):

$$\begin{aligned}\widehat{M}_n(x) &= \Big\langle \frac{\partial k_\mathcal{X}(\cdot,x)}{\partial x}, \big(\widehat{C}_{XX}^{(n)} + \varepsilon_n I\big)^{-1} \widehat{C}_{XY}^{(n)} \widehat{C}_{YX}^{(n)} \big(\widehat{C}_{XX}^{(n)} + \varepsilon_n I\big)^{-1} \frac{\partial k_\mathcal{X}(\cdot,x)}{\partial x} \Big\rangle \\ &= \nabla \mathbf{k}_X(x)^T (G_X + n\varepsilon_n I)^{-1} G_Y (G_X + n\varepsilon_n I)^{-1} \nabla \mathbf{k}_X(x),\end{aligned} \tag{11}$$

where G_X and G_Y are Gram matrices $(k_\mathcal{X}(X_i,X_j))$ and $(k_\mathcal{Y}(Y_i,Y_j))$, respectively, and $\nabla \mathbf{k}_X(x) = (\partial k_\mathcal{X}(X_1,x)/\partial x, \cdots, \partial k_\mathcal{X}(X_n,x)/\partial x)^T \in \mathbb{R}^{n\times m}$. In the case of Gaussian RBF kernel, for example, the j-th row of $\nabla \mathbf{k}_X(X_i)$ is given by $(1/\sigma^2)(X_i - X_j)\exp(-\|X_i - X_j\|^2/(2\sigma^2))$, which is simply the

Hadamard product between the Gram matrix G_X and $(X_i^a - X_j^a)_{ij=1}^n$ ($a = 1, \ldots, m$).

As the eigenvectors of $M(x)$ are contained in the EDR space for any x, we propose to use the eigenvectors of the $m \times m$ symmetric matrix

$$\begin{aligned}\tilde{M}_n &:= \frac{1}{n}\sum_{i=1}^{n} \widehat{M}_n(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \nabla \mathbf{k}_X(X_i)^T (G_X + n\varepsilon_n I_n)^{-1} G_Y (G_X + n\varepsilon_n I_n)^{-1} \nabla \mathbf{k}_X(X_i), \quad (12)\end{aligned}$$

the average of $\widehat{M}_n(X_i)$ over all the data points X_i. The projection matrix B in Eq. (1) is then estimated by the eigenvectors corresponding to the d largest eigenvalues of the $\tilde{M}_n$. We call this method the *gradient-based kernel dimension reduction* (gKDR). As shown in Section 2.3.3, the empirical average $\tilde{M}_n$ converges to the population mean $E[M(X)]$ at some rate.

2.3.2 Discussions and extensions

As an advantage of the kernel methods, the gKDR method can handle any type of variable for Y including multivariate or non-vectorial one in the same way, once a kernel is defined on the space. Also, the nonparametric nature of the kernel method avoids making strong assumptions on the distribution of X, Y, or the conditional probability, which are often needed in many famous dimension reduction methods such as SIR, pHd, contour regression, and so on.

As shown in Introduction, the previous gradient-based methods ADE and IADE are not necessarily able to find the EDR space, since they do not consider the conditional probability but only regressor. In contrast, by incorporating various nonlinear functions given by the nonlinear feature map

$k_{\mathcal{Y}}(\tilde{y}, \cdot)$, the gKDR method is able to find the EDR space with a characteristic kernel, as shown in Theorem 2 later.

The KDR method (Fukumizu et al., 2004, 2009) also provides a method for sufficient dimension reduction with no strong assumptions on the distribution. The computation of KDR, however, requires a gradient method with expensive matrix inversion, as discussed in Introduction. This makes it infeasible to apply KDR to large dimensionality more than hundreds. In contrast, the gKDR uses only the eigendecomposition after Gram matrix manipulation. As we see in Section 3, the gKDR approach can be used for data sets of ten thousand dimension.

The results of gKDR depend in practice on the choice of kernels and regularization coefficients as in all kernel methods. We use the cross-validation (CV) for choosing kernels and parameters, combined with some regression or classification method. In this paper, the simple k-nearest neighbor (kNN) regression / classification is used in the CV; for each candidate of kernel or parameter, we compute the CV error by the kNN method with the input data projected on the subspace given by gKDR, and choose the one that gives the least error.

The selection of appropriate dimensionality d is also an important issue. While many methods have been developed for the choice of dimensionality in respective dimension reduction methods (Schott, 1994, Ferré, 1998, Cook and Lee, 1999, Bura and Cook, 2001, Yin and Seymour, 2005, Li and Wang, 2007, Li et al., 2011, to list some), they are derived from asymptotic analysis of some test statistics, which may not be practical in situations of large dimensionality and small samples encountered often in current data analysis. In this paper, we do not discuss asymptotics of test statistics to select the

dimensionality, but consider the cross-validation with kNN, as discussed for parameter selection above, for estimating the optimum dimensionality.

The time complexity of the matrix inversions and the eigendecomposition required for gKDR are $O(n^3)$, which may be prohibitive for large data. We can apply, however, low-rank approximation of Gram matrices, such as incomplete Cholesky factorization (Fine and Scheinberg, 2001), which is a standard method for reducing time complexity in handling Gram matrices. It is known that the eigenspectrum of Gram matrices with Gaussian kernel decays fast for some typical data distributions (Widom, 1963, 1964) so that the low-rank approximation can give good approximation accuracy with significant saving of the computational cost. The complexity of incomplete Cholesky factorization for a matrix of size n is $O(nr^2)$ in time and $O(nr)$ in space, where r is the rank. The space complexity may be also a problem of gKDR, since $(\nabla \mathbf{k}_X(X_i))_{i=1}^n$ has $n^2 \times m$ dimension. In the case of Gaussian RBF kernel, the necessary memory can be reduced by low rank approximation of the Gram matrices. Recall that $\partial k_{\mathcal{X}}(X_j, x)/\partial x^a|_{x=X_i}$ for Gaussian RBF kernel is given by $(1/\sigma^2)(X_j^a - X_i^a)\exp(-\|X_j - X_i\|^2/(2\sigma^2))$ $(a = 1, \ldots, m)$. Let $G_X \approx RR^T$ and $G_Y \approx HH^T$ be the low rank approximation with $r_x = \mathrm{rk}R$ and $r_y = \mathrm{rk}H$ $(r_x, r_y < \min\{n, m\})$. With the notation $F := (G_X + n\varepsilon_n I_n)^{-1}H$ and $\Theta_i^{as} = (1/\sigma^2)X_i^a R_{is}$, we have

$$\tilde{M}_{n,ab} \approx \sum_{i=1}^{n}\sum_{t=1}^{r_y} \Gamma_{ia}^t \Gamma_{ib}^t \qquad (1 \le a, b \le m),$$

where

$$\begin{aligned}\Gamma^t_{ia} &= \sum_{j=1}^{n}\sum_{s=1}^{r_x}\frac{1}{\sigma^2}(X^a_j - X^a_i)R_{js}R_{is}F_{jt} \\ &= \sum_{s=1}^{r_x} R_{is}\Big(\sum_{j=1}^{n}\Theta^{as}_j F_{jt}\Big) - \sum_{s=1}^{r_x}\Theta^{as}_i\Big(\sum_{j=1}^{n} R_{js}F_{jt}\Big).\end{aligned}$$

With this approximation, the complexity is $O(nmr)$ in space and $O(nm^2r)$ in time ($r = \max\{r_x, r_y\}$), which is much more efficient in space than straightforward implementation.

We introduce two variants of gKDR. First, as discussed in Hristache et al. (2001), accurate nonparametric estimation for the derivative of regression function with high-dimensional X may not be easy in general. We propose a method for decreasing the dimensionality iteratively in a similar idea to IADE, but more directly. Using gKDR, we first find a projection matrix B_1 of a larger dimension d_1 than the target dimensionality d, project data X_i onto the subspace as $Z^{(1)}_i = B_1^T X_i$, and find the projection matrix B_2 ($d_1 \times d_2$ matrix) for $Z^{(1)}_i$ onto a d_2 ($d_2 < d_1$) dimensional subspace. After repeating this process to the dimensionality d, the final result is given by $\hat{B} = B_1 B_2 \cdots B_\ell$. In this way, we can expect the later projector is more accurate by the low dimensionality of the data $Z^{(s)}_i$. We call this method gKDR-i.

The iterative approach taken in gKDR-i is much simper than the method used by IADE, in which the data is projected by the matrix $(I+\rho^{-2}BB^T)^{-1/2}$ where BB^T is the projector estimated in the previous step and ρ is the parameter decreasing in the iteration. While IADE can continuously increase the contribution of the projector in the iterative procedure, the choice of the parameter ρ is arbitrary, and not easy to control.

Second, we see from Eq. (12) that the rank of $\tilde{M}_n$ is at most that of G_Y. This is a strong limitation of gKDR, since in classification problems, where the L classes are encoded as L different points, the Gram matrix G_Y is of rank L at most. Note that this problem is shared by some other linear dimension reduction methods including SIR and canonical correlation analysis (CCA). To solve this problem, we propose to use the variants of $\widehat{M}_n(x)$ over all points $x = X_i$ instead of the average $\tilde{M}_n$. After partitioning $\{1, \ldots, n\}$ into $T_1, \ldots, T_\ell$, we compute the $m \times d$ matrices $\widehat{B}_{[a]}$ $(a = 1, \ldots, \ell)$ given by the eigenvectors of $\widehat{M}_{[a]} = \sum_{i \in T_a} \widehat{M}(X_i)$, and make the final estimator $\widehat{B} \in \mathbb{R}^{m \times d}$ by the eigenvectors corresponding to the largest d eigenvalues of the matrix $\widehat{P} = \frac{1}{\ell} \sum_{a=1}^{\ell} \widehat{B}_{[a]} \widehat{B}_{[a]}^T$. We call this method gKDR-v. While we can use the same technique as the one in IADE, where orthonormal basis functions with respect to $(X_i)_{i=1}^n$ are employed in making a larger dimensional space than m, we take a simpler approach of partitioning the data points.

2.3.3 Theoretical properties of gKDR

We have derived the gKDR method based on the necessary condition of EDR space. The following theorem shows that the condition is sufficient also, if $k_{\mathcal{Y}}$ is characteristic. In the sequel, $\mathrm{Span}(B)$ denotes the subspace spanned by the column vectors of matrix B.

Theorem 2. *In addition to the assumptions (i)-(v), assume that the kernel $k_{\mathcal{Y}}$ is characteristic. If the eigenvectors of $M(x)$ is contained in* $\mathrm{Span}(B)$ *almost surely, then Y and X are conditionally independent given $B^T X$.*

Proof. First note that, from Eqs. (9) and (10), the eigenvectors of $M(x)$ is

contained in $\mathrm{Span}(B)$ if and only if $\partial E[g(Y)|X=x]/\partial x \in \mathrm{Span}(B)$ for any $g \in \mathcal{H}_{\mathcal{Y}}$. Let C be an $m \times (m-d)$ matrix such that $C^T C = I_{m-d}$ and the column vectors of C are orthogonal to those of B, and write $(U, V) = (B^T X, C^T X)$. Then, the condition $\partial E[g(Y)|X=x]/\partial x \in \mathrm{Span}(B)$ is equivalent to $E[g(Y)|(U,V)=(u,v)] = E[g(Y)|U=u]$ for any $g \in \mathcal{H}_{\mathcal{Y}}$. Since $k_{\mathcal{Y}}$ is characteristic, this implies that the conditional probability of Y given (U, V) is equal to that of Y given U, which means the desired conditional independence. □

The above theorem implies that the gKDR method estimates the sufficient dimension reduction space, which gives the conditional independence of Y and X given $B^T X$, assuming the existence of such a matrix B. While there may not exist such a subspace rigorously in practice, the ratio of the sum of the top d-eigenvalues $\sum_{i=1}^{d} \lambda_i / \sum_{j=1}^{m} \lambda_j$, where $\lambda_1 \geq \cdots \geq \lambda_m \geq 0$ are the eigenvalues of $\tilde{M}_n$, may be used for quantifying the degree of conditional independence. To see this possibility, we made a simple experiment using $Y = X_1 + \eta \cos(X_2) + Z$, where $X = (X_1, \ldots, X_5) \sim \mathrm{Unif}[-\pi, \pi]^5$ is five dimensional explanatory variables and $Z \sim N(0, 10^{-2})$ is an independent noise. With $n = 400$ and $d = 1$, we evaluated the ratio over 100 runs with different samples, and observed that the means of the ratio decrease monotonically $(0.893, 0.830, 0.722, 0.654, 0.590, 0.521)$, as the deviation from the conditional independence with $d = 1$ increases ($\eta = 0.0, 0.2, 0.4, 0.6, 0.8, 1.0$). This illustrates that the ratio can be a useful indicator for evaluating the conditional independence assumption. More theoretical discussions on this measure for conditional dependence will be an interesting and important problem, but it is not within the scope of this paper.

The next theorems show the consistency and its rate of gKDR estimator under some conditions on the smoothness. Theorem 3 shows the consistency with the total dimension m fixed, and Theorem 4 discusses the situation where the dimensionality m grows as sample size n increases. While the former is a corollary to the latter, for simplicity we show the result independently. The proofs are shown in Appendix.

Theorem 3. *Assume that* $\partial k_{\mathcal{X}}(\cdot, x)/\partial x^a \in \mathcal{R}(C_{XX}^{\beta+1})$ $(a = 1, \ldots, m)$ *for some* $\beta \geq 0$ *and* $E[k_{\mathcal{Y}}(y, Y)|X = \cdot] \in \mathcal{H}_{\mathcal{X}}$ *for every* $y \in \mathcal{Y}$. *Then, for the choice*

$$\varepsilon_n = n^{-\max\{\frac{1}{3}, \frac{1}{2\beta+2}\}},$$

we have

$$\widehat{M}_n(x) - M(x) = O_p\Big(n^{-\min\{\frac{1}{3}, \frac{2\beta+1}{4\beta+4}\}}\Big)$$

for every $x \in \mathcal{X}$ *as* $n \to \infty$. *If further* $E[\|M(X)\|_F^2] < \infty$ *and* $\partial k_{\mathcal{X}}(\cdot, x)/\partial x^a = C_{XX}^{\beta+1} h_x^a$ *for some* $h_x^a \in \mathcal{H}_{\mathcal{X}}$ *with* $E\|h_X^a\|_{\mathcal{H}_{\mathcal{X}}} < \infty$ $(a = 1, \ldots, m)$, *then* $\tilde{M}_n$ *converges in probability to* $E[M(X)]$ *in the same order as above.*

In considering dimension reduction for high dimensional X, it is important to consider the case where the dimension m grows as sample size n increases. In such cases, the positive definite kernel for X must be dependent on m. We assume that the response variable Y is fixed, and use $k^{(m)}$ for the positive definite kernel on $\mathbb{R}^m$ with the associated RKHS $\mathcal{H}_{\mathcal{X}}{}^{(m)}$. In discussing the convergence with the series of kernels, it is reasonable to assume $E[k^{(m)}(X, X)^2] = 1$ for any m, which normalizes the scale of the kernels. This is satisfied if the kernel has the form $k^{(m)}(x, \tilde{x}) = \varphi(\|x - \tilde{x}\|_{\mathbb{R}^m})$ with $\varphi(0) = 0$ such as Gaussian and Laplace kernel. In the following theorem the dimension m depends on n so that $m = m_n$. For notational simplicity,

however, the dependence of m on n is not explicitly shown in the symbols below.

As many quantities depend on the dimensionality m, we make the following assumptions in addition to (i)-(v).

(vi) For each $m = m_n$ there is $\beta_m \geq 0$ and $L_m \geq 0$ such that some $h_{a,x}^{(m)} \in {\mathcal{H}_{\mathcal{X}}}^{(m)}$ satisfies

$$\frac{\partial k^{(m)}(\cdot, x)}{\partial x^a} = C_{XX}^{\beta_m+1} h_{a,x}^{(m)} \qquad (a = 1, \ldots, m),$$

and $\|h_{a,x}^{(m)}\|_{{\mathcal{H}_{\mathcal{X}}}^{(m)}} \leq L_m$ irrespective to a and x.

(vii) Let

$$\alpha_m := (E[k^{(m)}(X, X)^2] - E[k^{(m)}(X, \tilde{X})^2])^{1/2},$$

where $\tilde{X}$ is an independent copy of X. Then,

$$\frac{\alpha_m}{\sqrt{n}} \to 0 \quad (n \to \infty).$$

Theorem 4. *Under the assumptions* (i)-(vii)*, for the choice*

$$\varepsilon_n = \Big(\frac{\alpha_m^2}{n}\Big)^{\max\{\frac{1}{3}, \frac{1}{2\beta_m+2}\}},$$

we have

$$\Big\|\widehat{M}_n(x) - M(x)\Big\|_F = O_p\Big(mL_m^2\Big(\frac{\alpha_m^2}{n}\Big)^{\min\{\frac{1}{3}, \frac{2\beta_m+1}{4\beta_m+4}\}}\Big)$$

for every $x \in \mathcal{X}$ as $n \to \infty$. If further $mL_m^2/\sqrt{n} \to 0$ $(n \to \infty)$, then $\tilde{M}_n$ converges in probability to $E[M(X)]$ of the order $O_p(mL_m^2/\sqrt{n} + mL_m^2(\frac{\alpha_m^2}{n})^{\min\{\frac{1}{3}, \frac{2\beta_m+1}{4\beta_m+4}\}})$ in Frobenius norm.

Note that, assuming that the d-th largest eigenvalues of $M(x)$ or $E[M(X)]$ is strictly larger than $(d+1)$-th largest one, the convergence of the matrices in

Theorems 3 and 4 implies the convergence of the corresponding eigenspaces (e.g., Stewart and Sun, 1990, Sec. V.2). This means that the estimator of gKDR is consistent to the subspace given by the top d eigenvectors of $E[M(X)]$. From Theorems 2, 3, and 4, under the assumptions, the gKDR gives a consistent method for sufficient dimension reduction.

To illustrate implications of Theorem 4, consider the case where $k^{(m)}(x, y) = \exp(x^T y/(2\sigma_m^2))$ and $X \sim N(0, \tau_m^2 I_m)$ with $\sigma_m > \sqrt{2}\tau_m$. It is easy to see that $\alpha_m^2 = 1/(1 - 2\delta_m^2)^m - 1/(1 - \delta_m^4)^m$ with $\delta_m = \tau_m/\sigma_m$. Suppose $\delta_m \to 0$. Then, from $(1 - 2\delta_m^2)^m = (1 - 2\delta_m^2)^{(1/2\delta_m^2)(2m\delta_m^2)} \approx e^{-2m\delta_m^2}$ and $1 - \left(\frac{1-2\delta_m^2}{1-\delta_m^4}\right)^m = 1 - (1 - \gamma_m\delta_m^2)^m \approx 1 - e^{-2m\delta_m^2}$ with $\gamma_m = (2 - \delta_m^2)/(1 - \delta_m^4)$, if $m\delta_m^2 \to \beta \in [0, \infty]$ as $m \to \infty$, we have $\alpha_m^2 \to e^{2\beta} - 1$, and in the case $m\delta_m^2 \to 0$, we further obtain $\alpha_m^2 \approx 2m\delta_m^2$. This shows τ_m/σ_m controls the convergence rate. On the other hand, the choice of σ_m is related to the assumption on L_m, for which the analysis is not straightforward. The above example on the order of α_m suggests that the convergence order may depend much on the kernel or kernel parameter. More detailed analysis of the high-dimensional kernel methods is an important future research direction.

The above consistency results assume the use of full Gram matrices, and thus the low-rank approximation discussed in Section 2.3.2 is not incorporated. Some consistency results can be proved without difficulty, if we set the rank sufficiently large, as sample size increases, so that the approximation errors can be negligibly small. The computational cost is higher, however, if the rank is larger. The method with low-rank approximation then has trade-off between estimation accuracy and computational cost, and the optimal choice is not straightforward.

3 Numerical examples

In the kernel methods of this section, the Gaussian RBF kernel $k(x, \tilde{x}) = \exp(-\|x - \tilde{x}\|^2/(2\sigma^2))$ is always used even for discrete variables.

3.1 Synthesized data

First we use the following four types of synthesized data to verify the basic performance of gKDR and the two variants:

$$\begin{aligned}
\text{(A):} \quad & Y = Z\sin(Z) + W, \quad Z = \tfrac{1}{\sqrt{5}}(X_1 + 2X_2), \\
& X \sim \mathrm{Unif}[-1,1]^{10}, \quad W \sim N(0, 10^{-2}),
\end{aligned}$$

$$\begin{aligned}
\text{(B):} \quad & Y = (Z_1^3 + Z_2)(Z_1 - Z_2^3) + W, \\
& Z_1 = \tfrac{1}{\sqrt{2}}(X_1 + X_2), \quad Z_2 = \tfrac{1}{\sqrt{2}}(X_1 - X_2), \\
& X \sim \mathrm{Unif}[-1,1]^{10}, \quad W \sim \Gamma(1,2).
\end{aligned}$$

$$\begin{aligned}
\text{(C):} \quad & Y = (X_1 - a)^4 E, \\
& X \sim (N(0,1/4) * I_{[-1,1]})^{10}, \quad E \sim N(0,1).
\end{aligned}$$

$$\begin{aligned}
\text{(D):} \quad & Y = \sum_{j=1}^{5} (Z_{2j-1}^3 + Z_{2j})(Z_{2j-1} - Z_{2j}^3) + W, \\
& Z_{2j-1} = \tfrac{1}{\sqrt{2}}(X_{2j-1} + X_{2j}), \quad Z_{2j} = \tfrac{1}{\sqrt{2}}(X_{2j-1} - X_{2j}), \\
& X \sim \mathrm{Unif}[-1,1]^{50}, \quad W \sim \mathrm{Laplace}(2).
\end{aligned}$$

The model (A) includes the additive Gaussian noise, while (B) has a skewed noise, which follows the Gamma distribution. The model (C) has multiplicative noise. In (A), (B) and (C), X is 10 dimensional, while (D) uses 50

dimensional X. Except (C), X is uniformly distributed, while in (C) X is generated by the truncated normal distribution. The model (A) is the same as the ones used in Hristache et al. (2001). The sample size is $n = 100, 200$ for (A)(B), $n = 200, 400$ for (C), and $n = 1000, 2000$ for (D). The discrepancy between the estimator B and the true projector B_0 is measured by $\|B_0 B_0^T (I_m - BB^T)\|_F / \sqrt{d}$. For choosing the parameter σ in Gaussian RBF kernel and the regularization parameter ε_n, the CV in Section 2.3.2 with kNN (k = 5) is used with 8 different values given by $c\sigma_{med}$ ($0.5 \leq c \leq 10$), where σ_{med} is the median of pairwise distances of data (Gretton et al., 2008), and $\ell = 4, 5, 6, 7$ for $\varepsilon_n = 10^{-\ell}$ (a similar strategy is used for the CV in all the experiments below). For gKDR-i, the dimensionality is reduced one by one in the case of (A)–(C), and 10 dimensions are reduced at one iteration for (D). For gKDR-v, the data is partitioned into 50 groups.

We compare the results with those of IADE, SIR II (Li, 1991), MAVE, and KDR. In IADE there are seven parameters: h_1 and ρ_1 for the initial value of the bandwidth h_k in the smoothing kernel $K(x/h_k)$ and the coefficient in the projection matrix $(I + B_k B_k / \rho_k^2)^{1/2}$, respectively; a_h and a_ρ for the increase / decay rate of h_k and ρ_k, respectively; $h_{\max}$ and $\rho_{\min}$ for the maximum / minimum values for the parameters; C_w for the threshold of the minimum eigenvalue of the weighted covariance matrix. We use the following setting:

$$h_1 = \gamma_h n^{-1/\max(4,m)}, \qquad h_{\max} = 2\sqrt{d}, \quad a_h = e^{1/2\max(4,m)},$$
$$\rho_1 = 1, \qquad \rho_{\min} = \gamma_\rho n^{-1/3}, \quad a_\rho = e^{-1/6}, \qquad C_w = 1/4$$

and optimize γ_h, γ_ρ manually for each data set so that we can obtain optimum results. Although Hristache et al. (2001) use $\gamma_h = \gamma_\rho = 1$, we observed

	gKDR	gKDR-i	gKDR-v	IADE	SIR II	MAVE	KDR	gKDR +KDR
(A) $n = 100$	0.1989	0.1639	0.2002	0.1372	0.2986	0.0748	0.2807	0.0883
	(0.0553)	(0.0479)	(0.0555)	(0.0552)	(0.1021)	(0.0934)	(0.3364)	(0.1473)
(A) $n = 200$	0.1264	0.0995	0.1287	0.0857	0.2077	0.0410	0.1175	0.0501
	(0.0321)	(0.0352)	(0.0351)	(0.0258)	(0.0554)	(0.0108)	(0.2184)	(0.0964)
(B) $n = 200$	0.2999	0.2743	0.3040	0.3972	0.3627	0.3306	0.3418	0.2643
	(0.1047)	(0.0796)	(0.0930)	(0.1319)	(0.0781)	(0.1332)	(0.2004)	(0.1105)
(B) $n = 400$	0.1763	0.1725	0.1833	0.2382	0.2361	0.1939	0.2587	0.1606
	(0.0373)	(0.0426)	(0.0369)	(0.0646)	(0.0457)	(0.0681)	(0.2228)	(0.0348)
(C-a) $n = 200$	0.1919	0.2322	0.1930	0.7724	0.7326	0.6216	0.1479	0.1285
	(0.0791)	(0.1512)	(0.0763)	(0.1665)	(0.0153)	(0.2402)	(0.1307)	(0.0483)
(C-a) $n = 400$	0.1346	0.1372	0.1369	0.7863	0.7167	0.4951	0.0897	0.0893
	(0.0472)	(0.0644)	(0.0499)	(0.1846)	(0.0470)	(0.2578)	(0.0294)	(0.0294)
(C-b) $n = 200$	0.2819	0.2949	0.2942	0.8212	0.9476	0.6222	0.1925	0.1897
	(0.1158)	(0.1722)	(0.1383)	(0.1369)	(0.0459)	(0.2206)	(0.0686)	(0.0632)
(C-b) $n = 400$	0.1794	0.1903	0.1849	0.8169	0.9094	0.5273	0.1216	0.1241
	(0.0728)	(0.1380)	(0.0844)	(0.1654)	(0.0729)	(0.1998)	(0.0372)	(0.0373)
(D) $n = 1000$	0.4321	0.4485	0.4366	$--$	0.6236	0.5269	0.9638	0.3126
	(0.0292)	(0.0367)	(0.0317)		(0.0255)	(0.0364)	(0.0117)	(0.0385)
(D) $n = 2000$	0.2323	0.2291	0.2327	$--$	0.4250	0.2517	0.9532	0.1830
	(0.0097)	(0.0121)	(0.00976)		(0.0159)	(0.0457)	(0.0057)	(0.0088)

Table 1: Results for the synthesized data. Means and standard errors (in brackets) over 100 samples are shown. (C-a) and (C-b) use $a = 0$ and 0.5, respectively.

this setting may not necessarily give good results in our simulations. For the smoothing kernel in IADE, the biweight kernel $K(z) = (1 - |Z|^2)_+^2$ is used as in Hristache et al. (2001). The choice of these parameters in IADE is not easy: if h_k is too small, only a small number of X_i lie in the support of the biweight kernel, which makes the weighted variance used in the method unstable. For SIR II, we tried several numbers of slices, and chose the one that gave the best result. For MAVE, we used the rMAVE Matlab code provided by Y. Xia (`http://www.stat.nus.edu.sg/~staxyc/`).

From Table 1, we see that gKDR, gKDR-i, and gKDR-v show much better results than SIR II for all the cases. The IADE and MAVE work better than these methods for the data (A); in particular, for additive Gaussian noise (A), MAVE shows much better results than all the other methods. For the multiplicative noise (C), IADE, SIR-II, and MAVE do not give meaningful estimation. The gKDR and gKDR-v show similar errors in all cases, and gKDR-i improves them for (A) and (B). The KDR method attains higher accuracy for (C), but is less accurate for (A) and (B) with $n = 100$; the undesired results in (A) and (B) are caused by failure of optimization in some cases, as the large standard deviations indicate. For the large dimensional case (D), IADE did not end after 2 days so we omit experiments, and the optimization of KDR does not seem to work well. We also use the results of gKDR as the initial state for KDR. As we can see from the table, KDR improves the accuracy significantly for all the cases with small standard errors, showing best results among the compared methods except MAVE for (A). Note again, however, that the data sets used here are very small in size and dimensionality, and it is not feasible to apply the KDR method to large data sets used in the following subsection.

3.2 Real world data

One way of evaluating dimension reduction methods in regression is to evaluate the regression or classification accuracy after projecting data onto the estimated subspaces. In this subsection, we use real world data sets for classification tasks. We compare gKDR (-i, -v) method with KDR, SIR-II, IADE, MAVE, CCA, and linear discriminant analysis (LDA), if they are applicable to the problems. Since IADE and MAVE assume one-dimensional response, we use those two methods only for binary classification by encoding the response 0 and 1. Except these cases, L classes are represented by the binary vectors $(1, 0, \ldots, 0), \ldots, (0, \ldots, 0, 1)$. Note that CCA and LDA can find at most $L - 1$ dimensional subspaces, and thus we do not apply them to binary classification. The wide applicability of the gKDR approach contrast with these limitations of the previous methods. In the following examples, when applicable, the gKDR-i reduces the dimension 10 times with almost equal amount, and the gKDR-v partitions data into 50 groups. For SIR-II, the slices are always given by the class labels. The support vector machine (SVM) with Gaussian RBF kernel is always used to evaluate the classification accuracy with the projected data. The one-vs-one method is applied for the multiclass cases, and the parameters in SVM (bandwidth in the kernel and the trade-off parameter C) are chosen by 10-fold CV. The experiments is performed with the standard package `libsvm` (Chang and Lin, 2011).

	Dim.	Train	Test
heart-disease	13	149	148
ionosphere	34	151	200
breast-cancer	30	200	369

Table 2: Summary of data sets: dimensionality of X and the number of data.

3.2.1 Small data sets

We first use three data sets for binary classification, *heart-disease*, *ionosphere*, and *breast-cancer-Wisconsin*. To evaluate the classification ability, each data set is divided into a training set and test set: the former is used to estimate an EDR space and make a classifier on that space, and the latter to test the classifier. This simple split method is adopted to reduce the computational cost needed for some methods. The configuration of the data sets are listed in Table 2. As explained in Section 2.3.2, the dimensionality of the subspace that can be found by gKDR and gKDR-i is at most the rank of the Gram matrix G_Y. Since the rank is at most two for binary classification, only gKDR-v is applied among the proposed methods. For the CV of the parameters, only the variance parameter of the Gaussian RBF kernel for X is changed at 9 values, and ε_n is fixed as 10^{-5}. For KDR, the number of iterations for optimization is 150, and the bandwidth parameters are reduced during the iterations, as described in Fukumizu et al. (2009). The bandwidth parameter for MAVE is chosen by CV among 8 values.

The classification rates of the SVM are shown in Table 3. We can see that gKDR-v, KDR, and IADE show competitive results, and the classifi-

cation rates are similar to the ones given by the original variables without dimension reduction in most cases. This implies that those methods found subspaces containing sufficient information for the classification tasks. The results of the MAVE are slightly worse than those three methods in many cases; this may be reasonable since MAVE assumes the additive noise model. The SIR-II gives worse results than the others. In classification problems, it is unlikely that the linearity or elliptic assumption required for the method holds. Table 4 shows the computational time needed for each parameter set in CV. The methods are implemented with MATLAB on Intel® Xeon® X5677 (3.47GHz). Since gKDR-v and SIR-II do not need iterative optimization, the required computational cost is significantly smaller than the other three methods. IADE sometimes shows very slow convergence with some parameter setting.

3.2.2 Larger data sets

In this subsection, we use three multiclass classification data sets, which are much larger in dimensionality and sample size than the ones used in the previous subsection. Since the tasks are multiclass classification, MAVE and IADE are not used for the experiments. Also, the optimization of KDR is not feasible to this size of data. In addition to SIR-II, we use LDA and CCA as baseline linear methods in comparison with the gKDR, gKDR-i, and gKDR-v.

USPS2007. The first data set is *USPS2007*, which is 2007 images of USPS handwritten digit data set used in Song et al. (2008). Each image has 256 gray-scale pixels used for X, and the class label for ten digits assigned for Y. We divide the data set into two; 1000 data are used for estimating

		$d=3$	$d=5$	$d=7$	$d=9$	$d=11$	All (13)
Heart-disease	gKDR-v	79.05	80.41	82.43	79.73	79.05	81.08
	KDR	76.35	78.38	79.05	77.70	81.08	
	MAVE	77.70	73.65	72.97	74.32	79.05	
	SIR-II	61.49	63.51	60.14	68.24	63.51	
	IADE	79.73	78.38	78.38	78.38	78.38	
		$d=3$	$d=5$	$d=10$	$d=15$	$d=20$	All (34)
Ionosphere	gKDR-v	75.50	87.50	88.00	86.00	89.00	89.00
	KDR	86.50	87.50	85.00	92.00	94.00	
	MAVE	88.00	81.00	85.00	83.00	85.50	
	SIR-II	40.50	49.50	72.50	76.50	76.00	
	IADE	84.50	94.00	91.50	88.50	90.00	
		$d=3$	$d=5$	$d=10$	$d=15$	$d=20$	All (30)
Breast-cancer	gKDR-v	90.79	93.77	91.87	92.14	92.41	92.14
	KDR	90.79	91.33	91.33	91.33	95.12	
	MAVE	85.64	87.26	84.28	87.80	92.14	
	SIR-II	77.51	85.37	81.57	81.84	80.22	
	IADE	89.97	91.60	90.51	91.60	94.04	

Table 3: Classification accuracy (%) for small binary classification data sets.

the subspace and training SVM, and the remaining 1007 for evaluating the classification errors of the SVM.

From Table 5, the three gKDR methods give significantly better classification performance than the other three methods. The SIR-II does not show meaningful results, since it is likely that the distribution of the explanatory variables does not satisfy the linearity or ellipticity condition required for SIR-II. This unfavored results of SIR-II are seen in all the large multiclass cases.

ISOLET. The second data set is ISOLET taken from UCI repository (Frank and Asuncion, 2010). The task is to classify 26 alphabets from 617 dimensional continuous features of speech signals. In addition to 6238 train-

	gKDR-v	KDR	MAVE	SIR-II	IADE
Heart-disease ($d = 11$)	0.044	622	16.7	0.000817	3.78
Ionosphere ($d = 20$)	0.l03	84.8	47.6	0.00849	6.62
Breast-cancer ($d = 20$)	0.116	615	61.0	0.0115	1345

Table 4: Computational time in seconds for binary classification data sets.

ing data, 1559 test data are separately provided. Table 6 shows the classification errors of SVM with data projected on the estimated subspaces. The classifier with gKDR (-i, -v) outperforms CCA, LDA, and SIR-II, if the dimensionality is larger than 10. For this data set also, the SIR-II does not provide meaningful results. From the information on the data at the UCI repository, the best performance with neural networks and C4.5 with ECOC are 3.27% and 6.61%, respectively. It is remarkable that the gKDR and gKDR-i combined with SVM are competitive with the best known results only with 20-25 dimensional linear features, and gKDR-v + SVM with 50 dimensional features outperforms them. This implies that the gKDR methods find very effective subspaces from the high dimensional variable for the classification task.

Amazon Commerce Reviews. The next example uses the data set for author identification of Amazon commerce reviews, which is taken from UCI repository. The explanatory variable is 10000 dimensional, consisting of authors' linguistic style such as usage of digit, punctuation, word and sentence length, usage frequency of words and so on. The total number of authors is 50, and 30 reviews have been collected for each author; the total size of data is thus 1500. we varied the used number of authors (L) to make different levels of difficulty for the tasks. It is known that the task

Dim.	3	5	7	9	15	20	25
gKDR	36.94	17.58	12.41	9.73	–	–	–
gKDR-i	40.42	23.54	17.18	10.63	–	–	–
gKDR-v	36.94	17.58	12.41	12.21	8.54	7.94	7.75
SIR-II	70.21	70.71	67.53	57.50	46.97	46.97	42.70
CCA	34.66	20.66	16.68	15.99	–	–	–
LDA	34.76	21.15	15.39	14.60	–	–	–

Table 5: USPS2007: classification errors (%) of SVM for test data.

Dim.	10	15	20	25	30	35	40	45	50
gKDR	14.43	7.50	5.00	4.75	–	–	–	–	–
gKDR-i	11.74	6.03	4.04	4.23	–	–	–	–	–
gKDR-v	16.87	7.57	4.75	4.30	3.85	3.85	3.59	3.53	3.08
SIR-II	74.41	69.53	66.07	60.81	57.47	51.31	48.88	46.18	42.53
CCA	13.09	8.66	6.54	6.09	–	–	–	–	–
LDA	13.21	8.15	6.61	6.67	–	–	–	–	–

Table 6: ISOLET: classification errors (%) of SVM for test data.

# Authors	10	20	30	40	50
gKDR	12.0	16.2	18.0	21.8	19.5
SIR-II	86.3	94.8	96.0	97.1	98.1
CCA	82.0	85.8	86.9	89.6	90.2
Corr (500)	15.7	30.2	29.2	35.4	41.1
Corr (2000)	8.3	18.0	24.0	25.0	29.0

Table 7: Amazon Commerce Reviews: 10-fold CV errors (%) of SVM.

becomes much more difficult if the number of classes (authors) is larger. The dimensionality of the projection given by gKDR is set to the same as the number of authors, and the 10-fold CV classification errors of SVM classifiers are evaluated with data projected on the estimated EDR space. For computational reason, we apply only the gKDR method among the proposed ones. Since the dimensionality is much larger than the sample size, LDA does not work by the singularity of the covariance matrix, and CCA provides very poor results by strong overfitting. Instead, we have applied the squared sum of variable-wise Pearson correlations, $\sum_{\ell=1}^{L} \mathrm{Corr}[X^a, Y^\ell]^2$ ($a = 1, \ldots, 10000$), to choose explanatory variables. This variable-wise method with Pearson Correlation is popularly used to select effective variables among a very high dimensional explanatory variables. The variables with top 500 and 2000 correlations are used to make SVM classifiers.

As we can see from Table 7, the gKDR gives much more effective subspaces for regression than the Pearson correlation method, when the number of authors is large. The creator of the data set has also reported the classification result with a neural network model (Liu et al., 2011); for 50 authors, the 10-fold cross-validation error with selected 2000 variables is 19.51%, which is similar to the gKDR result with the subspace of only 50 dimensions. This confirms again that the gKDR works well for finding an informative subspace in the large dimensional explanatory variables.

4 Conclusion

This paper has proposed a new method for linear dimension reduction in regression, using nonparametric estimation with positive definite kernels. By

estimating the gradient of regression function considered for feature vectors mapped into reproducing kernel Hilbert spaces, the proposed method is able to find the directions that achieve sufficient dimension reduction, that is, the directions to make the response and explanatory conditionally independent given the projection on those directions.

The proposed gKDR methods have several advantages over existing linear dimension reduction methods. The first one is the wide applicability: no strong assumptions on the distributions or the type of variables are needed. They can be used to continuous or discrete response in the same manner, and applicable to any noise model including multiplicative noise without knowing the noise model. The experimental results show that the gKDR methods achieve competitive results for additive noise models, and show much better accuracy for multiplicative noise than the existing methods. Second, their computational cost is inexpensive: they do not need iterative optimization procedures but need only eigendecomposition for estimating the projection matrix. Third, the methods have a theoretical guarantee of the sufficient dimension reduction. This can not be realized in general setting by many existing methods. Additionally, by virtue of the above properties, the methods work preferably for large dimensional data. This has been observed both theoretically and empirically; Theorem 4 shows the consistency of the estimator under the condition that the dimensionality of the explanatory variables may grow to infinity, and the experimental results indicate that the proposed methods successfully find effective directions with efficient computation for data sets up to 10000 dimension, for which many conventional sufficient dimension reduction methods have difficulty in finding effective directions.

There are many theoretical results on high-dimensional data in the literatures, where the dimensionality is assumed to be at most exponential order of the sample size. In the current work, as discussed after Theorem 4, the conditions of the convergence in the kernel method is linked with the dimensionality of the space only implicitly. It is an interesting future work to connect the conditions with more popular ones for other theoretical results and to clarify connections with the diverging order of the dimensionality.

The Matlab codes implementing the algorithms are provided at the first author's home page (`http://www.ism.ac.jp/~fukumizu/`).

Acknowledgements

KF has been supported in part by JSPS KAKENHI (B) 22300098.

Appendix

A Proof of Theorems 3 and 4

Since Theorem 3 can be shown as a corollary to Theorem 4 by setting α_m and L_m as constants, we show only the proof of Theorem 4 below. For notational simplicity, we omit the dependence on m in writing $k^{(m)}$ and associated covariance operators.

Let $g_a = \partial k_{\mathcal{X}}(\cdot, x)/\partial x^a$ $(a = 1, \ldots, m)$ (we omit to write the dependence on x). Since

$$\begin{aligned} M_{ab}(x) &= \Big\langle \big\langle E[k_{\mathcal{Y}}(*, Y)|X = \cdot], g_a \big\rangle_{\mathcal{H}_{\mathcal{X}}}, \big\langle E[k_{\mathcal{Y}}(*, Y)|X = \cdot], g_b \big\rangle_{\mathcal{H}_{\mathcal{X}}} \Big\rangle_{\mathcal{H}_{\mathcal{Y}}} \\ &= \big\langle E[k_{\mathcal{Y}}(*, Y)|g_a(X)], E[k_{\mathcal{Y}}(*, Y)|g_b(X)] \big\rangle_{\mathcal{H}_{\mathcal{Y}}} \end{aligned} \tag{13}$$

and

$$\widehat{M}_{n,ab}(x) = \langle \widehat{C}_{YX}^{(n)} \big(\widehat{C}_{XX}^{(n)} + \varepsilon_n I\big)^{-1} g_a, \widehat{C}_{YX}^{(n)} \big(\widehat{C}_{XX}^{(n)} + \varepsilon_n I\big)^{-1} g_b \rangle_{\mathcal{H}_\mathcal{Y}},$$

we have

$$\begin{aligned}
&\big|\widehat{M}_{n,ab}(x) - M_{ab}(x)\big| \\
\leq &\big|\big\langle \widehat{C}_{YX}^{(n)} \big(\widehat{C}_{XX}^{(n)} + \varepsilon_n I\big)^{-1} g_a, \widehat{C}_{YX}^{(n)} \big(\widehat{C}_{XX}^{(n)} + \varepsilon_n I\big)^{-1} g_b - E[k_\mathcal{Y}(*, Y)|g_b(X)]\big\rangle_{\mathcal{H}_\mathcal{Y}}\big| \\
&+ \big|\big\langle \widehat{C}_{YX}^{(n)} \big(\widehat{C}_{XX}^{(n)} + \varepsilon_n I\big)^{-1} g_a - E[k_\mathcal{Y}(*, Y)|g_a(X)], E[k_\mathcal{Y}(*, Y)|g_b(X)]\big\rangle_{\mathcal{H}_\mathcal{Y}}\big|.
\end{aligned}$$

From Assumption (vii), $\|C_{XX}\|_{HS}^2 = \|E[k_\mathcal{X}(\cdot, X) \otimes k_\mathcal{X}(\cdot, X)]\|_{\mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\mathcal{X}}^2 = E[k_\mathcal{X}(X, \tilde{X})^2] \leq E[k_\mathcal{X}(X, X)^2] = 1$, where $\tilde{X}$ is an independent copy of X. It also follows from Lemma 6 shown below that $\|\big(C_{XX} - \widehat{C}_{XX}^{(n)}\big)(C_{XX} + \varepsilon_n I)^{-1}\|_{HS} = O_p(\varepsilon_n^{-1} \alpha_m n^{-1/2})$. Noting $\alpha_m n^{-1/2} \varepsilon_n^{-1} \to 0$ by the choice of ε_n and the assumption (vii), from the expression

$$\big(\widehat{C}_{XX}^{(n)} + \varepsilon_n I\big)^{-1} = (C_{XX} + \varepsilon_n I)^{-1} \big\{ I - \big(C_{XX} - \widehat{C}_{XX}^{(n)}\big)(C_{XX} + \varepsilon_n I)^{-1} \big\}^{-1},$$

we obtain

$$\big\| C_{XX} \big(\widehat{C}_{XX}^{(n)} + \varepsilon_n I\big)^{-1} \big\| = O_p(1).$$

From $g_a = C_{XX}^{\beta_m + 1} h_a$, we have $\|\widehat{C}_{YX}^{(n)} \big(\widehat{C}_{XX}^{(n)} + \varepsilon_n I\big)^{-1} g_a\| = \|\widehat{C}_{YX}^{(n)} \big(\widehat{C}_{XX}^{(n)} + \varepsilon_n I\big)^{-1} C_{XX}^{\beta_m + 1} h_a\| = O_p(L_m)$. For the proof of the first assertion of Theorem 4, it is then sufficient to prove the following theorem.

Theorem 5. *Under the assumptions (i)-(vii), where $g_{a,x} = \partial k(\cdot, x)/\partial x^a$ and $h_{a,x}$ in (vi) are replaced by g and h, respectively, for $\varepsilon_n = (\alpha_m^2/n)^{\max\{1/3, 1/2(\beta_m+1)\}}$, we have*

$$\big\| \widehat{C}_{YX}^{(n)} \big(\widehat{C}_{XX}^{(n)} + \varepsilon_n I_n\big)^{-1} g - E[k_\mathcal{Y}(\cdot, Y)|g(X)] \big\|_{\mathcal{H}_\mathcal{Y}} = O_p\Big(L_m \Big(\frac{\alpha_m^2}{n} \Big)^{\min\{\frac{1}{3}, \frac{2\beta_m + 1}{4\beta_m + 4}\}} \Big)$$

as $n \to \infty$.

Proof. It suffices to show

$$\left\|\widehat{C}_{YX}^{(n)}\left(\widehat{C}_{XX}^{(n)}+\varepsilon_n I\right)^{-1}g - C_{YX}\left(C_{XX}+\varepsilon_n I\right)^{-1}g\right\|_{\mathcal{H}_\mathcal{Y}} = O_p\left(L_m\varepsilon_n^{-1/2}\alpha_m n^{-1/2}\right) \tag{14}$$

and

$$\left\|C_{YX}\left(C_{XX}+\varepsilon_n I\right)^{-1}g - E[k_\mathcal{Y}(\cdot,Y)|g(X)]\right\|_{\mathcal{H}_\mathcal{Y}} = O\left(L_m\varepsilon_n^{\min\{1,(2\beta_m+1)/2\}}\right) \tag{15}$$

as $n\to\infty$. In fact, balancing the rates easily derives the assertion of the theorem.

Since $B^{-1}-A^{-1}=B^{-1}(A-B)A^{-1}$ for any invertible operators A and B, the left hand side of Eq. (14) is upper bounded by

$$\begin{aligned}&\left\|\widehat{C}_{YX}^{(n)}\left(\widehat{C}_{XX}^{(n)}+\varepsilon_n I\right)^{-1}(C_{XX}-\widehat{C}_{XX}^{(n)})\left(C_{XX}+\varepsilon_n I\right)^{-1}C_{XX}^{\beta_m+1}h\right\|_{\mathcal{H}_\mathcal{Y}}\\ &\qquad+\left\|(\widehat{C}_{YX}^{(n)}-C_{YX})\left(C_{XX}+\varepsilon_n I\right)^{-1}C_{XX}^{\beta_m+1}h\right\|_{\mathcal{H}_\mathcal{Y}}.\end{aligned}$$

By the decomposition $\widehat{C}_{YX}^{(n)}=\widehat{C}_{YY}^{(n)1/2}\widehat{W}_{YX}\widehat{C}_{XX}^{(n)1/2}$ with $\|\widehat{W}_{YX}\|\le 1$ (Baker, 1973), we have $\|\widehat{C}_{YX}^{(n)}\left(\widehat{C}_{XX}^{(n)}+\varepsilon_n I\right)^{-1}\| = O(\varepsilon_n^{-1/2})$. From Lemma 6, $\|C_{XX}-\widehat{C}_{XX}^{(n)}\| = O_p(\alpha_m n^{-1/2})$. It follows from these two facts and assumption (vi) that the first term of the above expression is of $O_p(L_m\alpha_m\varepsilon_n^{-1/2}n^{-1/2})$. Eq. (14) is then obtained, since the second term is of $O_p(L_m\alpha_m n^{-1/2})$.

For Eq. (15), first note that by using Theorem 1 for each y

$$\begin{aligned}E[k_\mathcal{Y}(y,Y)|g(X)] &= \langle E[k_\mathcal{Y}(y,Y)|X=\cdot], g\rangle = \langle E[k_\mathcal{Y}(y,Y)|X=\cdot], C_{XX}^{\beta_m+1}h\rangle\\ &= \langle C_{XX}E[k_\mathcal{Y}(y,Y)|X=\cdot], C_{XX}^{\beta_m}h\rangle = \langle C_{XY}k_\mathcal{Y}(y,\cdot), C_{XX}^{\beta_m}h\rangle\\ &= \langle k_\mathcal{Y}(y,\cdot), C_{YX}C_{XX}^{\beta_m}h\rangle = (C_{YX}C_{XX}^{\beta_m}h)(y),\end{aligned}$$

which means $E[k_\mathcal{Y}(\cdot,Y)|g(X)] = C_{YX}C_{XX}^{\beta_m}h$. Let $C_{YX}=C_{YY}^{1/2}W_{YX}C_{XX}^{1/2}$ be

the decomposition with $\|W_{YX}\| \leq 1$. Then, we have

$$\begin{aligned}&\left\|C_{YX}\left(C_{XX}+\varepsilon_n I\right)^{-1} g - E[k_{\mathcal{Y}}(\cdot, Y)|g(X)]\right\|_{\mathcal{H}_{\mathcal{Y}}} \\ &\qquad = \|C_{YY}^{1/2} W_{YX}\| \left\|C_{XX}^{\beta_m+3/2}\left(C_{XX}+\varepsilon_n I\right)^{-1} h - C_{XX}^{\beta_m+1/2} h\right\|_{\mathcal{H}_{\mathcal{Y}}}.\end{aligned}$$

From

$$\|C_{XX}^{\beta_m+3/2}\left(C_{XX}+\varepsilon_n I\right)^{-1} - C_{XX}^{\beta_m+1/2}\| = \varepsilon_n \|C_{XX}^{\beta_m+1/2}\left(C_{XX}+\varepsilon_n I\right)^{-1}\|,$$

this norm is upper bounded by ε_n for $\beta_m \geq 1/2$. For $0 < \beta_m < 1/2$, it follows from $\varepsilon_n C_{XX}^{\beta_m+1/2}\left(C_{XX}+\varepsilon_n I\right)^{-1} = \varepsilon_n^{\beta_m+1/2}\{\varepsilon_n (C_{XX}+\varepsilon_n I)^{-1}\}^{1/2-\beta_m}\{(C_{XX}+\varepsilon_n I)^{-1} C_{XX}\}^{\beta_m+1/2}$ that the above norm is upper bounded by $\varepsilon_n^{\beta_m+1/2}$. We have thus Eq. (15), which completes the proof of Theorem 5

□

For the second assertion of Theorem 3, note

$$\begin{aligned}&\left\|\frac{1}{n}\sum_{i=1}^{n} \widehat{M}_n(X_i) - E[M(X)]\right\|_F \\ &\quad \leq \left\|\frac{1}{n}\sum_{i=1}^{n} \widehat{M}_n(X_i) - \frac{1}{n}\sum_{i=1}^{n} M(X_i)\right\|_F + \left\|\frac{1}{n}\sum_{i=1}^{n} M(X_i) - E[M(X)]\right\|_F.\end{aligned}$$

From Eq. (13) and $E[k_{\mathcal{Y}}(\cdot, Y)|g(X)] = C_{YX} C_{XX}^{\beta_m} h$, we have $E[M_{ab}(X)^2] = O(L_m^4)$, The second term in the right hand side is thus of $O_p(m L_m^2 n^{-1/2})$ by the central limit theorem. By replacing g and h in the proof of Theorem 5 by $\sum_{i=1}^n g_{a,X_i}/n$ and $\sum_{i=1}^n h_{a,X_i}/n$, respectively, where $g_{a,x} = \partial k_{\mathcal{X}}(\cdot, x)/\partial x^a$ and $g_{a,x} = C_{XX}^{\beta_m+1} h_{a,x}$, the first term can be bounded by

$$O_p\Big(m L_m \Big(\frac{\alpha_m^2}{n}\Big)^{\min\{\frac{1}{3}, \frac{2\beta_m+1}{4\beta_m+4}\}}\Big)$$

in the same manner.

The proof of the following Lemma, which is used in the above proof, can given by direct computation, and we omit it.

Lemma 6. *Let X be a random variable on a measurable space $(\Omega, \mathcal{B})$, and k be a measurable positive definite kernel such that $E[k_{\mathcal{X}}(X,X)^2] < \infty$. Then,*

$$E\big[\|\widehat{C}_{XX}^{(n)} - C_{XX}\|_{HS}^2\big] = \frac{1}{n}\big(E[k_{\mathcal{X}}(X,X)^2] - E[k_{\mathcal{X}}(X,\tilde{X})^2]\big),$$

where $\tilde{X}$ is an independent copy of X.

References

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

C.R. Baker. Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186:273–289, 1973.

F. Bauer, S. Pereverzev, and L. Rosasco. On regularization algorithms in learning theory. *Journal of Complexity*, 23(1):52–72, 2007.

C. Bernard-Michel, L. Gardes, and S. Girard. A note on sliced inverse regression with regularizations. *Biometrics*, 64:982–986, 2008.

B.E. Boser, I.M. Guyon, and V.N. Vapnik. A training algorithm for optimal margin classifiers. In *Fifth Annual ACM Workshop on Computational Learning Theory*, 144–152, ACM Press, 1992.

E. Bura and R.D. Cook. Extending sliced inverse regression: The weighted chi-squared test. *Journal of the American Statistical Association*, 96(455): 996–1003, 2001.

A. Caponnetto and E. De Vito. Optimal rates for regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2: 27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

R. D. Cook. On the interpretation of regression plots. *Journal of the American Statistical Association*, 89(425):177–189, 1994.

R. D. Cook. *Regression Graphics*. Wiley Inter-Science, 1998.

R. D. Cook and H. Lee. Dimension reduction in regression with a binary response. *Journal of the American Statistical Association*, 94:1187–1200, 1999.

R. D. Cook and S. Weisberg. Discussion of Li (1991). *Journal of the American Statistical Association*, 86:328–332, 1991.

J. Fan and I. Gijbels. *Local Polynomial Modelling and its Applications*. Chapman and Hall, 1996.

L. Ferré. Determining the dimension in sliced inverse regression and related methods. *Journal of the American Statistical Association*, 93(441):132–140, 1998.

S. Fine and K. Scheinberg. Efficient SVM training using low-rank kernel representations. *Journal of Machine Learning Research*, 2:243–264, 2001.

A. Frank and A. Asuncion. UCI machine learning repository, 2010. URL `http://archive.ics.uci.edu/ml`.

K. Fukumizu, F.R. Bach, and M.I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5:73–99, 2004.

K. Fukumizu, F.R. Bach, and M.I. Jordan. Kernel dimension reduction in regression. *Annals of Statistics*, 37(4):1871–1905, 2009. ISSN 0090-5364. doi: 10.1214/08-AOS637.

K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems 20*, pages 489–496. MIT Press, 2008.

A. Gretton, A. J. Smola, O. Bousquet, R. Herbrich, A. Belitski, M.A. Augath, Y. Murayama, J. Pauls, B. Schölkopf, and N.K. Logothetis. Kernel constrained covariance for dependence measurement. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, 2005a.

A. Gretton, R. Herbrich A. J. Smola, O. Bousquet, and B. Schölkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6:2075–2129, 2005b.

A. Gretton, K. Fukumizu, C.H. Teo, L. Song, B. Schölkopf, and Alex Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems 20*, pages 585–592. MIT Press, 2008.

A. Gretton, K. Fukumizu, and B. K. Sriperumbudur. Discussion of: Brownian distance covariance. *Annals of Applied Statistics*, 3(4):1285–1294, 2009.

T. Hofmann, B. Schölkopf, and A. J. Smola. Kernel methods in machine learning. *Annals of Statistics*, 36(3):1171–1220, 2008.

M. Hristache, A. Juditsky, J. Polzehl, and V. Spokoiny. Structure adaptive approach for dimension reduction. *Annals of Statistics*, 29(6):1537–1566, 2001.

T. Hsing and H. Ren. An rkhs formulation of the inverse regression dimension-reduction problem. *Annals of Statistics*, 37(2):726–755, 2009.

B. Li and S. Wang. On directional regression for dimension reduction. *Journal of the American Statistical Association*, 102:997–1008, 2007.

B. Li, H. Zha, and F. Chiaromonte. Contour regression: A general approach to dimension reduction. *Annals of Statistics*, 33(4):1580–1616, 2005.

B. Li, A. Artemiou, and L. Li. Principal support vector machine for linear and nonlinear sufficient dimension reduction. *Annals of Statistics*, 39(6): 3182-3210, 2011

K.-C. Li. Sliced inverse regression for dimension reduction (with discussion). *Journal of American Statistical Association*, 86:316–342, 1991.

K.-C. Li. On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *Journal of American Statistical Association*, 87:1025–1039, 1992.

S. Liu, Z. Liu, J. Sun, and L. Liu. Application of synergetic neural network in online writeprint identification. *International Journal of Digital Content Technology and its Applications*, 5(3):126–135, 2011.

A. M. Samarov. Exploring regression structure using nonparametric functional estimation. *Journal of the American Statistical Association*, 88 (423):836–847, 1993.

B. Schölkopf and A.J. Smola. *Learning with Kernels*. MIT Press, 2002.

J. R. Schott. Determining the dimensionality in sliced inverse regression. *Journal of the American Statistical Association*, 89:141–148, 1994.

S. Smale and D.-X. Zhou. Shannon sampling II: Connections to learning theory. *Applied and Computational Harmonic Analysis*, 19(3):285–302, 2005.

S. Smale and D.-X. Zhou. Learning theory estimates via integral operators and their approximation. *Constructive Approximation*, 26:153–172, 2007.

L. Song, J. Huang, A.J. Smola, and K. Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th International Conference on Machine Learning (ICML2009)*, pages 961–968. 2009.

L. Song, A. Smola, K. Borgwardt, and A. Gretton. Colored maximum variance unfolding. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1385–1392. MIT Press, Cambridge, MA, 2008.

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R.G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, 2008.

G.W. Stewart and J.-G. Sun. *Matrix Perturbation Theory*. Academic Press, 1990.

G. Wahba. *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics, 59. SIAM, 1990.

H. Widom. Asymptotic behavior of the eigenvalues of certain integral equations. *Transactions of the American Mathematical Society*, 109:278–295, 1963.

H. Widom. Asymptotic behavior of the eigenvalues of certain integral equations II. *Archive for Rational Mechanics and Analysis*, 17:215–229, 1964.

H.-M. Wu. Kernel sliced inverse regression with applications to classification. *Journal of Computational and Graphical Statistics*, 17(3):590–610, 2008.

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

X. Yin and L. Seymour. Asymptotic distributions for dimension reduction in the SIR-II method. *Statistica Sinica*, 15:1069–1079, 2005.

W. Zhong, P. Zeng, P. Ma, J.S. Liu, and Y. Zhu. RSIR : Regularized sliced inverse regression for motif discovery. *Bioinformatics*, 21(22):4169–4175, 2005.