



Gradient-Based Kernel Dimension Reduction for Regression

Kenji Fukumizu & Chenlei Leng

To cite this article: Kenji Fukumizu & Chenlei Leng (2014) Gradient-Based Kernel Dimension Reduction for Regression, Journal of the American Statistical Association, 109:505, 359-370, DOI: [10.1080/01621459.2013.838167](https://doi.org/10.1080/01621459.2013.838167)

To link to this article: <https://doi.org/10.1080/01621459.2013.838167>



Accepted author version posted online: 26 Sep 2013.
Published online: 26 Sep 2013.



Submit your article to this journal [↗](#)



Article views: 792



View Crossmark data [↗](#)



Citing articles: 15 View citing articles [↗](#)

Gradient-Based Kernel Dimension Reduction for Regression

Kenji FUKUMIZU and Chenlei LENG

This article proposes a novel approach to linear dimension reduction for regression using nonparametric estimation with positive-definite kernels or reproducing kernel Hilbert spaces (RKHSs). **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 the regression function considered for the feature vectors mapped into RKHSs. 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 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. The experimental results demonstrate that the proposed method successfully finds **effective directions with efficient computation even for high-dimensional explanatory variables.**

KEY WORDS: Conditional independence; Kernel method; Reproducing kernel Hilbert space; Sufficient dimension reduction.

1. INTRODUCTION

Modern data analysts are often confronted with high-dimensional data, which may come from images, texts, genomic expressions, and so on. Dimension reduction is almost always involved in such data analysis to avoid various problems caused by the high dimensionality; these problems are known as the *curse of dimensionality*. The purpose of dimension reduction thus includes preprocessing prior to 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 article 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 article 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 (CCA), **which can be used for extracting linear features straightforwardly, a modern approach to this problem is based on *sufficient dimension reduction* (Cook 1994, 1998), which formulates the problem by conditional independence.** More precisely, assuming

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

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, Zha, and Chiaromonte 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 datasets. Another recent approach uses support vector machines (SVM) 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).

Most relevant to this article is the method based on the gradient of regressor $\varphi(x) = E[Y|X = x]$ (Samarov 1993; Hristache et al. 2001). As detailed in Section 2.1, in Equation (1) the gradient of $\varphi(x)$ is contained in the EDR space at each x . One

Kenji Fukumizu is PhD, The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan (E-mail: fukumizu@ism.ac.jp). Chenlei Leng is PhD, 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, Singapore (E-mail: C.Leng@warwick.ac.uk). KF has been supported in part by JSPS KAKENHI (B) 22300098.

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 article proposes a novel approach to sufficient dimension reduction with positive definite kernels. Positive-definite kernels or reproducing kernel Hilbert spaces (RKHSs) have been widely used for data analysis (Wahba 1990), especially since the success of the SVM (Boser, Guyon, and Vapnik 1992; Hofmann, Schölkopf, and Smola 2008). The methods, in short, extract nonlinear features or higher-order moments of data by transforming them into RKHSs defined by positive-definite kernels. Various methods for nonparametric inference also have been recently developed in this discipline (Gretton et al. 2005a, 2008; Gretton, Fukumizu, and Sriperumbudur 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, Bach, and Jordan 2004, 2009) uses conditional covariance on RKHSs to characterize the conditional independence relation in Equation (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 datasets, the difficulty in the optimization prohibits application 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 article. Additionally, with RKHSs, Hsing and Ren (2009) discussed an extension of inverse regression from finite-dimensional to infinite-dimensional problems.

The method proposed in this article 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, Bach, and Jordan 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 article 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 datasets. Section 4 concludes this article. The technical proofs of the theoretical results are shown in the Appendix.

2. GRADIENT-BASED KERNEL DIMENSION REDUCTION

In this article, 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 was 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 Equation (1) holds and wish to estimate the projection matrix B using iid sample $(X_1, Y_1), \dots, (X_n, Y_n)$. In Equation (1), it is easy to see

$$\begin{aligned} \frac{\partial}{\partial x} E[Y|X = x] &= \frac{\partial}{\partial x} \int y p(y|x) dy \\ &= \int y \frac{\partial \tilde{p}(y|B^T x)}{\partial x} dy \\ &= B \int y \frac{\partial \tilde{p}(y|z)}{\partial z} \Big|_{z=B^T x} dy, \end{aligned}$$

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 following.

Note that the methods based on the gradient of the regression function $E[Y|X = x]$ use only a necessary condition of Equation (1) and not sufficient in general. In fact, if the variables $X = (X^1, \dots, X^m)$ 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 article 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 RKHSs have been extensively applied to data analysis especially since the success of the SVM in classification problems (Wahba 1990; Schölkopf and Smola 2002; Hofmann, Schölkopf, and Smola 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, Bach, and Jordan 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 $k : \Omega \times \Omega \rightarrow \mathbb{R}$ such that $\sum_{i,j=1}^n c_i c_j k(x_i, x_j) \geq 0$ for any x_1, \dots, x_n in Ω and $c_1, \dots, c_n \in \mathbb{R}$. 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 (1) $k(\cdot, x)$ is in \mathcal{H} , (2) the linear hull of $\{k(\cdot, x) \mid x \in \Omega\}$ is dense in \mathcal{H} , and (3) 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 (3) is called *reproducing property*, and the Hilbert space \mathcal{H} the 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 $C_{YX} : \mathcal{H}_{\mathcal{X}} \rightarrow \mathcal{H}_{\mathcal{Y}}$ is defined as the operator such that

$$\begin{aligned} \langle g, C_{YX} f \rangle_{\mathcal{H}_{\mathcal{Y}}} &= E[f(X)g(Y)] \\ &= E[\langle f, \Phi_{\mathcal{X}}(X) \rangle_{\mathcal{H}_{\mathcal{X}}} \langle \Phi_{\mathcal{Y}}(Y), g \rangle_{\mathcal{H}_{\mathcal{Y}}}] \end{aligned} \quad (2)$$

holds for all $f \in \mathcal{H}_{\mathcal{X}}$, $g \in \mathcal{H}_{\mathcal{Y}}$, where $\Phi_{\mathcal{X}} : \mathcal{X} \rightarrow \mathcal{H}_{\mathcal{X}}$ and $\Phi_{\mathcal{Y}} : \mathcal{Y} \rightarrow \mathcal{H}_{\mathcal{Y}}$ 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 Equation (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

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

where $P_{\mathcal{X}}$ 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 (Ω, \mathcal{B}) is called *characteristic* if the mapping from a probability Q on (Ω, \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, Bach, and Jordan 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, prob-

abilities 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, Bach, and Jordan 2009) that a positive-definite kernel on a measurable space (Ω, \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 (Ω, \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 iid sample $(X_1, Y_1), \dots, (X_n, Y_n)$ with law P , the covariance operator is estimated by the empirical covariance operator

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

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

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

Theorem 1 (Fukumizu, Bach, and Jordan 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 $P_{\mathcal{X}}$ 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 Equation (4), namely,

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

where ε_n is a regularization coefficient in Thikonov-type regularization. We can prove that Equation (5) is a consistent estimator of $E[g(Y)|X = \cdot]$ in $L^2(P_{\mathcal{X}})$ -norm even if $E[g(Y)|X = \cdot]$ is not in $\mathcal{H}_{\mathcal{X}}$ but in $L^2(P_{\mathcal{X}})$, 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 $O(n^{-\min\{\frac{1}{4}, \frac{\nu}{2\nu+2}\}})$ 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; Bauer, Pereverzev, and Rosasco 2007; Caponnetto and De Vito 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, Gardes, and Girard 2008).

The estimator (Equation (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 $\partial k(\cdot, x)/\partial x \in \mathcal{H}_X$, the relation

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

holds for any $f \in \mathcal{H}_X$. 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 Equation (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 μ_Y . We prepare positive definite-kernels k_X and k_Y on \mathbb{R}^m and \mathcal{Y} , respectively, with respective RKHS \mathcal{H}_X and \mathcal{H}_Y . We assume that Equation (1) holds for some $m \times d$ matrix B with $B^T B = I_d$. It is then easy to see that for any $g \in \mathcal{H}_Y$, there exists a function $\varphi_g(z)$ on \mathbb{R}^d such that

$$E[g(Y) | X] = \varphi_g(B^T X). \quad (7)$$

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

Recall we make the following assumptions.

- (i) \mathcal{H}_X and \mathcal{H}_Y are separable.
- (ii) k_X and k_Y are measurable, and $E[k_X(X, X)] < \infty$, $E[k_Y(Y, Y)] < \infty$.

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

- (iii) $k_X(\tilde{x}, x)$ is continuously differentiable and $\partial k_X(\cdot, x)/\partial x^i \in \mathcal{R}(C_{XX})$ for $i = 1, \dots, m$.
- (iv) $E[k_Y(y, Y) | X = \cdot] \in \mathcal{H}_X$ for any $y \in \mathcal{Y}$.
- (v) $\varphi_g(z)$ in Equation (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 $a = 1, \dots, d$.

Assumption (iv) implies that $E[g(Y) | X = \cdot] \in \mathcal{H}_X$ for any $g \in \mathcal{H}_Y$. Under Equation (1), Assumption (v) is true if $C := \int \sqrt{k_Y(y, \bar{y})} |\partial \tilde{p}(y|z)/\partial z^a| d\mu_Y(y)$ is finite for any z and the differentiation and integration are exchangeable: in fact, it is easy

to see

$$\left| \frac{\partial \varphi_g(z)}{\partial z^a} \right| \leq \int |\langle g, k_Y(\cdot, y) \rangle| \left| \frac{\partial \tilde{p}(y|z)}{\partial z^a} \right| d\mu_Y(y) \leq C \|g\|_{\mathcal{H}_Y}.$$

By Riesz' theorem, Assumption (v) implies that there is $\Psi_a(z) \in \mathcal{H}_X$ such that for $a = 1, \dots, d$,

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

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

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

holds for any $g \in \mathcal{H}_Y$. On the other hand, letting $C_{XX}^{-1}(\partial k_X(\cdot, x)/\partial x^i)$ denote the inverse element guaranteed by Assumption (iii), Theorem 1 and Equation (6) show that for any $g \in \mathcal{H}_Y$

$$\begin{aligned} \frac{\partial}{\partial x^i} E[g(Y) | X = x] &= \left\langle C_{YX} g, C_{XX}^{-1} \frac{\partial k_X(\cdot, x)}{\partial x^i} \right\rangle \\ &= \left\langle g, C_{YX} C_{XX}^{-1} \frac{\partial k_X(\cdot, x)}{\partial x^i} \right\rangle. \end{aligned} \quad (9)$$

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

$$\begin{aligned} &\left\langle C_{YX} C_{XX}^{-1} \frac{\partial k_X(\cdot, x)}{\partial x^i}, C_{YX} C_{XX}^{-1} \frac{\partial k_X(\cdot, x)}{\partial x^j} \right\rangle_{\mathcal{H}_Y} \\ &= \sum_{a,b=1}^d B_{ia} B_{jb} \langle \nabla_a \varphi(B^T x), \nabla_b \varphi(B^T x) \rangle_{\mathcal{H}_Y} \end{aligned}$$

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

$$M_{ij}(x) = \left\langle C_{YX} C_{XX}^{-1} \frac{\partial k_X(\cdot, x)}{\partial x^i}, C_{YX} C_{XX}^{-1} \frac{\partial k_X(\cdot, x)}{\partial x^j} \right\rangle_{\mathcal{H}_Y}, \quad (10)$$

are contained in the EDR space.

Given iid sample $(X_1, Y_1), \dots, (X_n, Y_n)$ from the true distribution, the estimator of $M(x)$ is easily obtained based on Equation (5):

$$\begin{aligned} \widehat{M}_n(x) &= \left\langle \frac{\partial k_X(\cdot, x)}{\partial x}, (\widehat{C}_{XX}^{(n)} + \varepsilon_n I)^{-1} \widehat{C}_{XY}^{(n)} \widehat{C}_{YX}^{(n)} \right. \\ &\quad \times \left. (\widehat{C}_{XX}^{(n)} + \varepsilon_n I)^{-1} \frac{\partial k_X(\cdot, x)}{\partial x} \right\rangle \\ &= \nabla \mathbf{k}_X(x)^T (G_X + n\varepsilon_n I)^{-1} G_Y (G_X + n\varepsilon_n I)^{-1} \nabla \mathbf{k}_X(x), \end{aligned} \quad (11)$$

where G_X and G_Y are Gram matrices $(k_X(X_i, X_j))$ and $(k_Y(Y_i, Y_j))$, respectively, and $\nabla \mathbf{k}_X(x) = (\partial k_X(X_1, x)/\partial x, \dots, \partial k_X(X_n, x)/\partial x)^T \in \mathbb{R}^{n \times m}$. 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 $(X_i^a - X_j^a)_{i,j=1}^n$ ($a = 1, \dots, m$).

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

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

the average of $\hat{M}_n(X_i)$ over all the data points X_i . The projection matrix B in Equation (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 variables for Y including multivariate or nonvectorial one in the same way, once a kernel is defined on the space. Also, the non-parametric 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 (Li 1992), contour regression, and so on.

As shown in Section 1, 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 $k_Y(\tilde{y}, \cdot)$, the gKDR method is able to find the EDR space with a characteristic kernel, as shown in Theorem 2.

The KDR method (Fukumizu, Bach, and Jordan 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 Section 1. This makes it infeasible to apply KDR to datasets with more than hundreds of dimensions. 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 datasets of 10,000 dimensions.

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 article, 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 encoun-

tered often in current data analysis. In this article, 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 also be 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 $\partial k_X(X_j, x)/\partial x^a|_{x=X_i}$ for Gaussian RBF kernel is given by $(1/\sigma^2)(X_j^a - X_i^a) \exp(-\|X_j - X_i\|^2/(2\sigma^2))$ ($a = 1, \dots, m$). Let $G_X \approx RR^T$ and $G_Y \approx HH^T$ be the low-rank approximation with $r_x = \text{rk} R$ and $r_y = \text{rk} H$ ($r_x, r_y < \min\{n, m\}$). 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

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

where

$$\begin{aligned}\Gamma_{ia}^t &= \sum_{j=1}^n \sum_{s=1}^{r_x} \frac{1}{\sigma^2} (X_j^a - X_i^a) R_{js} R_{is} F_{jt} \\ &= \sum_{s=1}^{r_x} R_{is} \left(\sum_{j=1}^n \Theta_j^{as} F_{jt} \right) - \sum_{s=1}^{r_x} \Theta_i^{as} \left(\sum_{j=1}^n R_{js} F_{jt} \right).\end{aligned}$$

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 $Z_i^{(1)} = B_1^T X_i$, and find the projection matrix B_2 ($d_1 \times d_2$ matrix) for $Z_i^{(1)}$ onto a d_2 ($d_2 < d_1$) dimensional subspace. After repeating this process to the dimensionality d , the final result is given by $\hat{B} = B_1 B_2 \dots B_\ell$. In this way, we can expect that the later projector is more accurate by the low dimensionality of the data $Z_i^{(s)}$. We call this method gKDR-i.

The iterative approach taken in gKDR-i is much simpler than the method used by IADE, in which the data are projected by the matrix $(I + \rho^{-2} B B^T)^{-1/2}$, where $B B^T$ 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 Equation (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 CCA. To solve this problem, we propose to use the variants of $\hat{M}_n(x)$ over all points $x = X_i$ instead of the average \tilde{M}_n . After partitioning $\{1, \dots, n\}$ into T_1, \dots, T_ℓ , we compute the $m \times d$ matrices $\hat{B}_{[a]}$ ($a = 1, \dots, \ell$) given by the eigenvectors of $\hat{M}_{[a]} = \sum_{i \in T_a} \hat{M}(X_i)$, and make the final estimator $\hat{B} \in \mathbb{R}^{m \times d}$ by the eigenvectors corresponding to the largest d eigenvalues of the matrix $\hat{P} = \frac{1}{\ell} \sum_{a=1}^{\ell} \hat{B}_{[a]} \hat{B}_{[a]}^T$. 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_Y is characteristic. In the following, $\text{Span}(B)$ denotes the subspace spanned by the column vectors of matrix B .

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

Proof. First note that, from Equations (9) and (10), the eigenvectors of $M(x)$ is contained in $\text{Span}(B)$ if and only if $\partial E[g(Y)|X = x]/\partial x \in \text{Span}(B)$ for any $g \in \mathcal{H}_Y$. 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 $\partial E[g(Y)|X = x]/\partial x \in \text{Span}(B)$ is equivalent to $E[g(Y)|(U, V) = (u, v)] = E[g(Y)|U = u]$ for any $g \in \mathcal{H}_Y$. Since k_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. \square

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 $\lambda_1 \geq \dots \geq \lambda_m \geq 0$ 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 $X = (X_1, \dots, X_5) \sim \text{Unif}[-\pi, \pi]^5$ 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 article.

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 the Appendix.

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

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

we have

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

for every $x \in \mathcal{X}$ as $n \rightarrow \infty$. If further $E[\|M(X)\|_F^2] < \infty$ and $\partial k_X(\cdot, x)/\partial x^a = C_{XX}^{\beta+1} h_x^a$ for some $h_x^a \in \mathcal{H}_X$ with $E\|h_x^a\|_{\mathcal{H}_X} < \infty$ ($a = 1, \dots, m$), 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 $\mathcal{H}_X^{(m)}$. 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 Assumptions (i)–(v).

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

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

and $\|h_{a,x}^{(m)}\|_{\mathcal{H}_X^{(m)}} \leq L_m$ 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,

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

Theorem 4. Under Assumptions (i)–(vii), for the choice

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

we have

$$\|\hat{M}_n(x) - M(x)\|_F = O_p \left(m L_m^2 \left(\frac{\alpha_m^2}{n} \right)^{\min\{\frac{1}{3}, \frac{2\beta_m+1}{4\beta_m+4}\}} \right)$$

for every $x \in \mathcal{X}$ as $n \rightarrow \infty$. If further $m L_m^2 / \sqrt{n} \rightarrow 0$ ($n \rightarrow \infty$), then \hat{M}_n converges in probability to $E[M(X)]$ of the order $O_p(m L_m^2 / \sqrt{n} + m L_m^2 (\frac{\alpha_m^2}{n})^{\min\{\frac{1}{3}, \frac{2\beta_m+1}{4\beta_m+4}\}})$ 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, Section 5.2). 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 $\delta_m \rightarrow 0$. 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 $1 - (\frac{1-2\delta_m^2}{1-\delta_m^4})^m = 1 - (1 - \gamma_m\delta_m^2)^m \approx 1 - e^{-2m\delta_m^2}$ with $\gamma_m = (2 - \delta_m^2)/(1 - \delta_m^4)$, if $m\delta_m^2 \rightarrow \beta \in [0, \infty]$ as $m \rightarrow \infty$, we have $\alpha_m^2 \rightarrow e^{2\beta} - 1$, and in the case $m\delta_m^2 \rightarrow 0$, 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:

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

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

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

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

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 models (A), (B), and (C), X is 10-dimensional, while (D) uses 50-dimensional X . Except model (C), X is uniformly distributed, while in model (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 $\|B_0 B_0^T (I_m - B B^T)\|_F / \sqrt{d}$. 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 eight different values given by $c\sigma_{\text{med}}$ ($0.5 \leq c \leq 10$), 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 models (A)–(C), and 10 dimensions are reduced at one iteration for model (D). For gKDR-v, the data are 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 ρ_{\min} for the maximum/minimum values for the parameters; and C_w for the threshold of the minimum eigenvalue of the weighted covariance matrix. We use the following setting:

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

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

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

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

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 datasets used here are very small in size and dimensionality, and it is not feasible to apply the KDR method to large datasets 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 section, we use real-world datasets for classification tasks. We compare gKDR (-i, -v) method with KDR, SIR-II, IADE, MAVE, CCA, and linear discriminant analysis (LDA), if they are ap-

plicable 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 $(1, 0, \dots, 0), \dots, (0, \dots, 0, 1)$. 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 SVM with Gaussian RBF kernel is always used to evaluate the classification accuracy with the projected data. The one-to-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).

3.2.1 Small Datasets. We first use three datasets for binary classification, *heart disease*, *ionosphere*, and *breast-cancer Wisconsin*. To evaluate the classification ability, each dataset 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 datasets is 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 nine values, and ε_n is fixed as 10^{-5} . For KDR, the number of iterations for optimization

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

	Dim.	Train	Test
Heart disease	13	149	148
Ionosphere	34	151	200
Breast cancer	30	200	369

is 150, and the bandwidth parameters are reduced during the iterations, as described in Fukumizu, Bach, and Jordan (2009). The bandwidth parameter for MAVE is chosen by CV among eight 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 classification rates are similar to those 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 (3.47 GHz). 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 Datasets. In this section, we use three multiclass classification datasets, which are much larger in dimensionality and sample size than those 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,

Table 4. Computational time in seconds for binary classification datasets

	gKDR-v	KDR	MAVE	SIR-II	IADE
Heart disease ($d = 11$)	0.044	622	16.7	0.000817	3.78
Ionosphere ($d = 20$)	0.103	84.8	47.6	0.00849	6.62
Breast cancer ($d = 20$)	0.116	615	61.0	0.0115	1345

we use LDA and CCA as baseline linear methods in comparison with the gKDR, gKDR-i, and gKDR-v.

USPS2007. The first dataset is *USPS2007*, which is 2007 images of USPS handwritten digit dataset used in Song et al. (2008). Each image has 256 grayscale pixels used for X , and the class label for 10 digits assigned for Y . We divide the dataset into two; 1000 data are used for estimating 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 dataset 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 training 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 dataset 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

Table 3. Classification accuracy (%) for small binary classification datasets

		$d = 3$	$d = 5$	$d = 7$	$d = 9$	$d = 11$	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	
Ionosphere		$d = 3$	$d = 5$	$d = 10$	$d = 15$	$d = 20$	All (34) 89.00
	gKDR-v	75.50	87.50	88.00	86.00	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	
Breast cancer	IADE	84.50	94.00	91.50	88.50	90.00	All (30) 92.14
		$d = 3$	$d = 5$	$d = 10$	$d = 15$	$d = 20$	
	gKDR-v	90.79	93.77	91.87	92.14	92.41	
	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 5. USPS2007: classification errors (%) of SVM for test data

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	—	—	—

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 dataset for author identification of Amazon commerce reviews, which is taken from UCI repository. **The explanatory variable is 10,000 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 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 variablewise Pearson correlations, $\sum_{\ell=1}^L \text{corr}[X^a, Y^\ell]^2$ ($a = 1, \dots, 10,000$), to choose explanatory variables. This variablewise 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 dataset 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

Table 7. Amazon commerce reviews: 10-fold CV errors (%) of SVM

# 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

for finding an informative subspace in the large dimensional explanatory variables.

4. CONCLUSION

This article 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 RKHSs, 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 cannot be realized in general setting by many existing methods. Additionally, by virtue of the above properties, the methods work well 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 datasets up to 10,000 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 literature, where the dimensionality is assumed to be at

Table 6. ISOLET: 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	—	—	—	—	—

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/>).

APPENDIX A

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. For notational simplicity, we omit the dependence on m in writing $k^{(m)}$ and associated covariance operators.

Let $g_a = \partial k_X(\cdot, x)/\partial x^a$ ($a = 1, \dots, m$) (we omit to write the dependence on x). Since

$$\begin{aligned} M_{ab}(x) &= \langle \langle E[k_Y(\cdot, Y)|X = \cdot], g_a \rangle_{\mathcal{H}_X}, \langle E[k_Y(\cdot, Y)|X = \cdot], g_b \rangle_{\mathcal{H}_X} \rangle_{\mathcal{H}_Y} \\ &= \langle E[k_Y(\cdot, Y)|g_a(X)], E[k_Y(\cdot, Y)|g_b(X)] \rangle_{\mathcal{H}_Y} \end{aligned} \quad (\text{A.1})$$

and

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

we have

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

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

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

we obtain

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

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

Theorem A.1. Under 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

$$\begin{aligned} &\|\widehat{C}_{YX}^{(n)}(\widehat{C}_{XX}^{(n)} + \varepsilon_n I)^{-1} g - E[k_Y(\cdot, Y)|g(X)]\|_{\mathcal{H}_Y} \\ &= O_p\left(L_m \left(\frac{\alpha_m^2}{n}\right)^{\min\{\frac{1}{3}, \frac{2\beta_m+1}{4\beta_m+4}\}}\right) \end{aligned}$$

as $n \rightarrow \infty$.

Proof. It suffices to show

$$\begin{aligned} &\|\widehat{C}_{YX}^{(n)}(\widehat{C}_{XX}^{(n)} + \varepsilon_n I)^{-1} g - C_{YX}(C_{XX} + \varepsilon_n I)^{-1} g\|_{\mathcal{H}_Y} \\ &= O_p(L_m \varepsilon_n^{-1/2} \alpha_m n^{-1/2}) \end{aligned} \quad (\text{A.2})$$

and

$$\begin{aligned} &\|C_{YX}(C_{XX} + \varepsilon_n I)^{-1} g - E[k_Y(\cdot, Y)|g(X)]\|_{\mathcal{H}_Y} \\ &= O(L_m \varepsilon_n^{\min\{1, (2\beta_m+1)/2\}}) \end{aligned} \quad (\text{A.3})$$

as $n \rightarrow \infty$. 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 Equation (A.2) is upper bounded by

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

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

For Equation (A.3), first note that by using Theorem 1 for each y

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

which means $E[k_Y(\cdot, Y)|g(X)] = C_{YX} C_{XX}^{\beta_m} h$. Let $C_{YX} = C_{YY}^{1/2} W_{YX} C_{XX}^{1/2}$ be the decomposition with $\|W_{YX}\| \leq 1$. Then, we have

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

From

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

this norm is upper bounded by ε_n for $\beta_m \geq 1/2$. For $0 < \beta_m < 1/2$, it follows from $\varepsilon_n C_{XX}^{\beta_m+1/2} (C_{XX} + \varepsilon_n I)^{-1} = \varepsilon_n^{1/2} \{ \varepsilon_n (C_{XX} + \varepsilon_n I)^{-1} \}^{1/2-\beta_m} \{ (C_{XX} + \varepsilon_n I)^{-1} C_{XX} \}^{\beta_m+1/2}$ that the above norm is upper bounded by $\varepsilon_n^{\beta_m+1/2}$. We have thus Equation (A.3), which completes the proof of Theorem A.1 \square

For the second assertion of Theorem 3, note

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

From Equation (A.1) and $E[k_Y(\cdot, Y)|g(X)] = C_{YX} C_{XX}^{\beta_m} h$, 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 A.1 by $\sum_{i=1}^n g_{a,X_i}/n$ and $\sum_{i=1}^n h_{a,X_i}/n$, respectively, where $g_{a,x} = \partial k_X(\cdot, x)/\partial x^a$ and $g_{a,x} = C_{XX}^{\beta_m+1} h_{a,x}$, the first term can be bounded by

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

in the same manner.

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

Lemma A.1. Let X be a random variable on a measurable space (Ω, \mathcal{B}) , and k be a measurable positive-definite kernel such that $E[k_X(X, X)^2] < \infty$. Then,

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

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

[Received April 2012. Revised July 2013]

REFERENCES

- Aronszajn, N. (1950), "Theory of Reproducing Kernels," *Transactions of the American Mathematical Society*, 68, 337–404. [361]
- Baker, C. R. (1973), "Joint Measures and Cross-Covariance Operators," *Transactions of the American Mathematical Society*, 186, 273–289. [369]
- Bauer, F., Pereverzev, S., and Rosasco, L. (2007), "On Regularization Algorithms in Learning Theory," *Journal of Complexity*, 23, 52–72. [361]
- Bernard-Michel, C., Gardes, L., and Girard, S. (2008), "A Note on Sliced Inverse Regression With Regularizations," *Biometrics*, 64, 982–986. [362]
- Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992), "A Training Algorithm for Optimal Margin Classifiers," in *Fifth Annual ACM Workshop on Computational Learning Theory*, Pittsburgh, PA: ACM Press, pp. 144–152. [360]
- Bura, E., and Cook, R. D. (2001), "Extending Sliced Inverse Regression: The Weighted Chi-squared Test," *Journal of the American Statistical Association*, 96, 996–1003. [363]
- Caponnetto, A., and De Vito, E. (2007), "Optimal Rates for Regularized Least-Squares Algorithm," *Foundations of Computational Mathematics*, 7, 331–368. [361]
- Chang, C.-C., and Lin, C.-J. (2011), "LIBSVM: A Library for Support Vector Machines," *ACM Transactions on Intelligent Systems and Technology*, 2, 27:1–27:27. Available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>. [366]
- Cook, R. D. (1994), "On the Interpretation of Regression Plots," *Journal of the American Statistical Association*, 89, 177–189. [359]
- (1998), *Regression Graphics*, New York: Wiley Inter-Science. [359]
- Cook, R. D., and Lee, H. (1999), "Dimension Reduction in Regression With a Binary Response," *Journal of the American Statistical Association*, 94, 1187–1200. [363]
- Cook, R. D., and Weisberg, S. (1991), Discussion of "Sliced Inverse Regression for Dimension Reduction" by K.-C. Li, *Journal of the American Statistical Association*, 86, 328–332. [359]
- Fan, J., and Gijbels, I. (1996), *Local Polynomial Modelling and its Applications*, London: Chapman and Hall. [360]
- Ferré, L. (1998), "Determining the Dimension in Sliced Inverse Regression and Related Methods," *Journal of the American Statistical Association*, 93, 132–140. [363]
- Fine, S., and Scheinberg, K. (2001), "Efficient SVM Training Using Low-rank Kernel Representations," *Journal of Machine Learning Research*, 2, 243–264. [363]
- Frank, A., and Asuncion, A. (2010), "UCI Machine Learning Repository," Available at <http://archive.ics.uci.edu/ml> [367]
- Fukumizu, K., Bach, F. R., and Jordan, M. I. (2004), "Dimensionality Reduction for Supervised Learning With Reproducing Kernel Hilbert Spaces," *Journal of Machine Learning Research*, 5, 73–99. [360,361,363]
- (2009), "Kernel Dimension Reduction in Regression," *The Annals of Statistics*, 37, 1871–1905. doi:10.1214/08-AOS637. [360,361,363,367]
- Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. (2008), "Kernel Measures of Conditional Dependence," in *Advances in Neural Information Processing Systems 20*, Cambridge, MA: MIT Press, pp. 489–496. [360]
- Gretton, A., Fukumizu, K., and Sriperumbudur, B. K. (2009), "Discussion of Brownian Distance Covariance," *Annals of Applied Statistics*, 3, 1285–1294. [360]
- Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., and Smola, A. (2008), "A Kernel Statistical Test of Independence," in *Advances in Neural Information Processing Systems 20*, Cambridge, MA: MIT Press, pp. 585–592. [360,365]
- Gretton, A., Herbrich, R., Smola, A. J., Bousquet, O., and Schölkopf, B. (2005a), "Kernel Methods for Measuring Independence," *Journal of Machine Learning Research*, 6, 2075–2129. [360]
- Gretton, A., Smola, A. J., Bousquet, O., Herbrich, R., Belitski, A., Augath, M. A., Murayama, Y., Pauls, J., Schölkopf, B., and Logothetis, N. K. (2005b), "Kernel Constrained Covariance for Dependence Measurement," in *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, 112–119. [361]
- Hofmann, T., Schölkopf, B., and Smola, A. J. (2008), "Kernel Methods in Machine Learning," *The Annals of Statistics*, 36, 1171–1220. [360,361]
- Hristache, M., Juditsky, A., Polzehl, J., and Spokoiny, V. (2001), "Structure Adaptive Approach for Dimension Reduction," *The Annals of Statistics*, 29, 1537–1566. [359,360,363,365]
- Hsing, T., and Ren, H. (2009), "An RKHS Formulation of the Inverse Regression Dimension-reduction Problem," *The Annals of Statistics*, 37, 726–755. [360]
- Li, B., Artemiou, A., and Li, L. (2011), "Principal Support Vector Machine for Linear and Nonlinear Sufficient Dimension Reduction," *The Annals of Statistics*, 39, 3182–3210. [359,363]
- Li, B., and Wang, S. (2007), "On Directional Regression for Dimension Reduction," *Journal of the American Statistical Association*, 102, 997–1008. [359,363]
- Li, B., Zha, H., and Chiaromonte, F. (2005), "Contour Regression: A General Approach to Dimension Reduction," *The Annals of Statistics*, 33, 1580–1616. [359]
- Li, K.-C. (1991), "Sliced Inverse Regression for Dimension Reduction" (with discussion), *Journal of American Statistical Association*, 86, 316–342. [359,365]
- Li, K.-C. (1992), "On Principal Hessian Directions for Data Visualization and Dimension Reduction: Another Application of Stein's Lemma," *Journal of American Statistical Association*, 87, 1025–1039. [363]
- Liu, S., Liu, Z., Sun, J., and Liu, L. (2011), "Application of Synergetic Neural Network in Online Writeprint Identification," *International Journal of Digital Content Technology and its Applications*, 5, 126–135. [368]
- Samarov, A. M. (1993), "Exploring Regression Structure Using Nonparametric Functional Estimation," *Journal of the American Statistical Association*, 88, 836–847. [359,360]
- Schölkopf, B., and Smola, A. J. (2002), *Learning With Kernels*, Cambridge, MA: MIT Press. [361]
- Schott, J. R. (1994), "Determining the Dimensionality in Sliced Inverse Regression," *Journal of the American Statistical Association*, 89, 141–148. [363]
- Smale, S., and Zhou, D.-X. (2005), "Shannon Sampling. II: Connections to Learning Theory," *Applied and Computational Harmonic Analysis*, 19, 285–302. [361]
- (2007), "Learning Theory Estimates Via Integral Operators and Their Approximation," *Constructive Approximation*, 26, 153–172. [361]
- Song, L., Huang, J., Smola, A. J., and Fukumizu, K. (2009), "Hilbert Space Embeddings of Conditional Distributions With Applications to Dynamical Systems," in *Proceedings of the 26th International Conference on Machine Learning (ICML2009)*, pp. 961–968. [360,361]
- Song, L., Smola, A., Borgwardt, K., and Gretton, A. (2008), "Colored Maximum Variance Unfolding," in *Advances in Neural Information Processing Systems*, Vol. 20, eds. J. C. Platt, D. Koller, Y. Singer, and S. Roweis, Cambridge, MA: MIT Press, pp. 1385–1392. [367]
- Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. (2010), "Hilbert Space Embeddings and Metrics on Probability Measures," *Journal of Machine Learning Research*, 11, 1517–1561. [361]
- Steinwart, I., and Christmann, A. (2008), *Support Vector Machines*, New York: Springer. [362]
- Stewart, G. W., and Sun, J.-G. (1990), *Matrix Perturbation Theory*, New York: Academic Press. [365]
- Wahba, G. (1990), *Spline Models for Observational Data*, CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 59, Philadelphia, PA: SIAM. [360,361]
- Widom, H. (1963), "Asymptotic Behavior of the Eigenvalues of Certain Integral Equations," *Transactions of the American Mathematical Society*, 109, 278–295. [363]
- (1964), "Asymptotic Behavior of the Eigenvalues of Certain Integral Equations. II," *Archive for Rational Mechanics and Analysis*, 17, 215–229. [363]
- Wu, H.-M. (2008), "Kernel Sliced Inverse Regression With Applications to Classification," *Journal of Computational and Graphical Statistics*, 17, 590–610. [360]
- Xia, Y., Tong, H., Li, W. K., and Zhu, L.-X. (2002), "An Adaptive Estimation of Dimension Reduction Space," *Journal of Royal Statistical Society, Series B*, 64, 363–410. [359]
- Yin, X., and Seymour, L. (2005), "Asymptotic Distributions for Dimension Reduction in the SIR-II Method," *Statistica Sinica*, 15, 1069–1079. [363]
- 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. [362]