Kernel Dimension Reduction in Regression

Kenji Fukumizu
Institute of Statistical Mathematics
4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569, Japan
fukumizu@ism.ac.jp

Francis R. Bach
Centre de Morphologie Mathématique
Ecole des Mines de Paris
77300 Fontainebleau, France
francis.bach@mines.org

Michael I. Jordan
Department of Computer Science and Electrical Engineering
Department of Statistics
University of California, Berkeley, CA 04720, USA
jordan@stat.berkeley.edu

September 12, 2006

Abstract

We present a new methodology for sufficient dimension reduction (SDR). Our methodology derives directly from a formulation of SDR in terms of the conditional independence of the covariate X from the response Y, given the projection of X on the central subspace (cf. Li, 1991; Cook, 1998). We show that this conditional independence assertion can be characterized in terms of conditional covariance operators on reproducing kernel Hilbert spaces and we show how this characterization leads to an M-estimator for the central subspace. The resulting estimator is shown to be consistent under weak conditions; in particular, we do not have to impose linearity or ellipticity conditions of the kinds that are generally invoked for SDR methods. We also present empirical results showing that the new methodology is competitive in practice.

1 Introduction

The problem of *sufficient dimension reduction* (SDR) for regression is that of finding a subspace S such that the projection of the covariate vector X onto S captures the statistical dependency of the response Y on X. More formally, let us characterize a *dimension-reduction subspace* S in terms of the following conditional independence assertion:

$$Y \perp\!\!\!\perp X \mid \Pi_S X, \tag{1}$$

where $\Pi_S X$ denotes the orthogonal projection of X onto S. It is possible to show that under weak conditions the intersection of dimension reduction subspaces is itself a dimension reduction subspace, in which case the intersection is referred to as a *central subspace* (Cook, 1998; Chiaromonte and Cook, 2002). As suggested in a seminal paper by Li (1991), it is of great interest to develop procedures for estimating this subspace, quite apart from any interest in the conditional distribution $P(Y \mid X)$ or the conditional mean $E(Y \mid X)$. Once the central subspace is identified, subsequent analysis can attempt to infer a conditional distribution or a regression function using the (low-dimensional) coordinates $\Pi_S X$.

The line of research on SDR initiated by Li is to be distinguished from the large and heterogeneous collection of methods for dimension reduction in regression in which specific modeling assumptions are imposed on the conditional distribution $P(Y \mid X)$ or the regression $E(Y \mid X)$. These methods include ordinary least squares, partial least squares, canonical correlation analysis, ACE, projection pursuit regression and neural networks. These methods can be effective if the modeling assumptions that they embody are met, but if these assumptions do not hold there is no guarantee of finding the central subspace.

Li's paper not only provided a formulation of SDR as a semiparametric inference problem—with subsequent contributions by Cook and others bringing it to its elegant expression in terms of conditional independence—but also suggested a specific inferential methodology that has had significant influence on the ensuing literature. Specifically, Li suggested approaching the SDR problem as an *inverse* regression problem. Roughly speaking, the idea is that if the conditional distribution $P(Y \mid X)$ concentrates on a subspace of the covariate space, then the inverse regression $E(X \mid Y)$ should lie in that same subspace. Moreover, it should be easier to regress X on Y than vice versa, given that Y is generally low-dimensional (indeed, one-dimensional in the majority of applications) while X is high-dimensional. Li (1991) proposed a particularly simple instantiation of this idea—known

as *sliced inverse regression* (SIR)—in which $E(X \mid Y)$ is estimated as a constant vector within each slice of the response variable Y, and principal component analysis is used to aggregate these constant vectors into an estimate of the central subspace. The past decade has seen a number of further developments in this vein, including principal Hessian directions (pHd, Li, 1992), sliced average variance estimation (SAVE, Cook and Weisberg, 1991; Cook and Yin, 2001) and contour regression (Li et al., 2005). A particular focus of these more recent developments has been the exploitation of second moments within an inverse regression framework.

While the inverse regression perspective has been quite useful, it is not without its drawbacks. In particular, performing a regression of X on Y generally requires making assumptions with respect to the probability distribution of X, assumptions that can be difficult to justify. In particular, most of the inverse regression methods make the assumption of linearity of the conditional mean of the covariate along the central subspace (or make a related assumption for the conditional covariance). These assumptions hold in particular if the distribution of X is elliptic. In practice, however, we do not necessarily expect that the covariate vector will follow an elliptic distribution, nor is it easy to assess departures from ellipticity in a high-dimensional setting. In general it seems unfortunate to have to impose probabilistic assumptions on X in the setting of a regression methodology.

Inverse regression methods can also exhibit some additional limitations depending on the specific nature of the response variable Y. In particular, pHd and contour regression are applicable only to a one-dimensional response. Also, if the response variable takes its values in a finite set of p elements, SIR yields a subspace of dimension at most $p - 1$; thus, for the important problem of binary classification SIR yields only a one-dimensional subspace. Finally, in the binary classification setting, if the covariance matrices of the two classes are the same, SAVE and pHd also provide only a one-dimensional subspace (Cook and Lee, 1999). The general problem in these cases is that the estimated subspace is smaller than the central subspace.

In this paper we present a new methodology for SDR that is rather different from the approaches considered thus far in the literature. Rather than focusing on first and second moments, and thereby engaging the machinery of classical regression, we focus instead on the criterion of conditional independence in terms of which the SDR problem is defined. We develop a contrast function for evaluating subspaces that is minimized precisely when the conditional independence assertion in Eq. (1) is realized. As befits a criterion that measures departure from conditional independence, our contrast

function is not based solely on first and second moments.

Our approach involves the use of conditional covariance operators on reproducing kernel Hilbert spaces (RKHS's). Our use of RKHS's is related to their use in nonparametric regression and classification; in particular, the RKHS's given by some positive definite kernels are Hilbert spaces of smooth functions that are "small" enough to yield computationally-tractable procedures, but are rich enough to capture nonparametric phenomena of interest (Wahba, 1990), and this computational focus is an important aspect of our work. On the other hand, whereas in nonparametric regression and classification the role of RKHS's is to provide basis expansions of regression functions and discriminant functions, in our case the RKHS plays a different role. Our interest is not in the functions in the RKHS per se, but rather in conditional covariance operators defined on the RKHS. We show that these operators can be used to measure departures from conditional independence. We also show that these operators can be estimated from data and that these estimates are functions of Gram matrices. Thus our approach—which we refer to as *kernel dimension reduction* (KDR)—involves computing Gram matrices from data and optimizing a particular functional of these Gram matrices to yield an estimate of the central subspace.

This approach makes no strong assumptions on either the conditional distribution $p_{Y|\Pi_S X}(y \mid \Pi_S x)$ or the marginal distribution $p_X(x)$. As we show, KDR is consistent as an estimator of the central subspace under weak conditions.

There are alternatives to the inverse regression approach in the literature that have some similarities to KDR. In particular, minimum average variance estimation (MAVE, Xia et al., 2002) is based on nonparametric estimation of the conditional covariance of Y given X, an idea related to KDR. This method explicitly estimates the regressor, however, assuming an additive noise model $Y = f(X) + Z$, where Z is independent of X. KDR does not make such an assumption, and does not estimate the regressor explicitly. Other related approaches include methods that estimate the derivative of the regression function; these are based on the fact that the derivative of the conditional expectation $g(x) = E[y \mid B^T x]$ with respect to x belongs to a dimension reduction subspace (Samarov, 1993; Hristache et al., 2001). These methods again assume an additive noise model, however, and impose the condition $E[g'(B^T X)] \neq 0$; a condition that is violated if g and the distribution of X exhibit certain symmetries. In general, we are aware of no method that attacks SDR directly by assessing departures from conditional independence.

We presented an earlier kernel dimension reduction method in Fukumizu

et al. (2004). The contrast function presented in that paper, however, was not derived as an estimator of a conditional covariance operator, and it was not possible to establish a consistency result for that approach. The contrast function that we present here is derived directly from the conditional covariance perspective; moreover, it is simpler than the earlier estimator and it is possible to establish consistency for the new formulation. We should note, however, that the empirical performance of the earlier KDR method was shown by Fukumizu et al. (2004) to yield a significant improvement on SIR and pHd in the case of non-elliptic data, and these empirical results motivated us to pursue the general approach further.

While KDR has advantages over other SDR methods because of its generality and its directness in capturing the semiparametric nature of the SDR problem, it also reposes on a more complex mathematical framework that presents new theoretical challenges. Thus, while consistency for SIR and related methods follows from a straightforward appeal to the central limit theorem (under ellipticity assumptions), more effort is required to study the statistical behavior of KDR theoretically. This effort is of some general value, however; in particular, to establish the consistency of KDR we prove the uniform $O(n^{-1/2})$ convergence of an empirical process that takes values in a reproducing kernel Hilbert space. This result, which accords with the order of uniform convergence of an ordinary real-valued empirical process, may be of independent theoretical interest.

It should be noted at the outset that we do not attempt to provide distribution theory for KDR in this paper, and in particular we do not address the problem of inferring the dimensionality of the central subspace.

The paper is organized as follows. In Section 2 we show how conditional independence can be characterized by cross-covariance operators on an RKHS and use this characterization to derive the KDR method. Section 3 presents numerical examples of the KDR method. We present a consistency theorem and its proof in Section 4. Section 5 provides concluding remarks. Some of the details of the proof of consistency are provided in the Appendix.

2 Kernel Dimension Reduction for Regression

The method of kernel dimension reduction is based on a characterization of conditional independence using operators on RKHS's. We present this characterization in Section 2.1 and show how it yields a population criterion for SDR in Section 2.2. This population criterion is then turned into a

finite-sample estimation procedure in Section 2.3.

In this paper, a Hilbert space means a separable Hilbert space, and an operator always means a linear operator. The operator norm of a bounded operator T is denoted by $\|T\|$. The null space and the range of an operator T are denoted by $\mathcal{N}(T)$ and $\mathcal{R}(T)$, respectively.

2.1 Characterization of conditional independence

Let $(\mathcal{X}, \mathcal{B}_\mathcal{X})$ and $(\mathcal{Y}, \mathcal{B}_\mathcal{Y})$ denote measurable spaces. When the base space is a topological space, the Borel σ-field is always assumed. Let $(\mathcal{H}_\mathcal{X}, k_\mathcal{X})$ and $(\mathcal{H}_\mathcal{Y}, k_\mathcal{Y})$ be RKHS's of functions on $\mathcal{X}$ and $\mathcal{Y}$, respectively, with measurable positive definite kernels $k_\mathcal{X}$ and $k_\mathcal{Y}$ (Aronszajn, 1950). We consider a random vector $(X, Y) : \Omega \to \mathcal{X} \times \mathcal{Y}$ with the law P_{XY}. The marginal distribution of X and Y are denoted by P_X and P_Y, respectively. It is always assumed that the positive definite kernels satisfy

$$E_X[k_\mathcal{X}(X, X)] < \infty \quad \text{and} \quad E_Y[k_\mathcal{Y}(Y, Y)] < \infty. \tag{2}$$

Under this assumption, $\mathcal{H}_\mathcal{X}$ and $\mathcal{H}_\mathcal{Y}$ are included in $L^2(P_X)$ and $L^2(P_Y)$, respectively, where $L^2(\mu)$ denotes the Hilbert space of square integrable functions with respect to the measure μ, and the inclusions $J_\mathcal{X} : \mathcal{H}_\mathcal{X} \to L^2(P_X)$ and $J_\mathcal{Y} : \mathcal{H}_\mathcal{Y} \to L^2(P_Y)$ are continuous, because $E_X[f(X)^2] = E_X[\langle f, k_\mathcal{X}(\cdot, X)\rangle^2_{\mathcal{H}_\mathcal{X}}] \le \|f\|^2_{\mathcal{H}_\mathcal{X}} E_X[k_\mathcal{X}(X, X)]$ for $f \in \mathcal{H}_\mathcal{X}$.

The *cross-covariance operator* of (X, Y) is an operator from $\mathcal{H}_\mathcal{X}$ to $\mathcal{H}_\mathcal{Y}$ so that

$$\langle g, \Sigma_{YX} f\rangle_{\mathcal{H}_\mathcal{Y}} = E_{XY}\big[(f(X) - E_X[f(X)])(g(Y) - E_Y[g(Y)])\big] \tag{3}$$

holds for all $f \in \mathcal{H}_\mathcal{X}$ and $g \in \mathcal{H}_\mathcal{Y}$ (Baker, 1973; Fukumizu et al., 2004). Obviously, $\Sigma_{YX} = \Sigma^*_{XY}$, where T^* denotes the adjoint of an operator T. If Y is equal to X, the positive self-adjoint operator Σ_{XX} is called the *covariance operator*.

For a random variable $X : \Omega \to \mathcal{X}$, the *mean element* $m_X \in \mathcal{H}_\mathcal{X}$ is defined by the element that satisfies

$$\langle f, m_X\rangle_{\mathcal{H}_\mathcal{X}} = E_X[f(X)] \tag{4}$$

for all $f \in \mathcal{H}_\mathcal{X}$; that is, $m_X = J^*_X 1$, where 1 is the constant function. Using the mean elements, Eq. (3), which characterizes Σ_{YX}, can be written as

$$\langle g, \Sigma_{YX} f\rangle_{\mathcal{H}_\mathcal{Y}} = E_{XY}[\langle f, k_\mathcal{X}(\cdot, X) - m_X\rangle_{\mathcal{H}_\mathcal{X}} \langle k_\mathcal{Y}(\cdot, Y) - m_Y, g\rangle_{\mathcal{H}_\mathcal{Y}}].$$

Let Q_X and Q_Y be the orthogonal projections which map $\mathcal{H}_\mathcal{X}$ onto $\overline{\mathcal{R}(\Sigma_{XX})}$ and $\mathcal{H}_\mathcal{Y}$ onto $\overline{\mathcal{R}(\Sigma_{YY})}$, respectively. It is known (Baker, 1973, Theorem 1) that Σ_{YX} has a representation of the form

$$\Sigma_{YX} = \Sigma_{YY}^{1/2} V_{YX} \Sigma_{XX}^{1/2}, \tag{5}$$

where $V_{YX} : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Y}$ is a unique bounded operator such that $\|V_{YX}\| \leq 1$ and $V_{YX} = Q_Y V_{YX} Q_X$.

A cross-covariance operator on an RKHS can be represented explicitly as an integral operator. For arbitrary $\varphi \in L^2(P_X)$ and $y \in \mathcal{Y}$, the integral

$$G_\varphi(y) = \int_{\mathcal{X}\times\mathcal{Y}} k_\mathcal{Y}(y, \tilde{y})(\varphi(\tilde{x}) - E_X[\varphi(X)]) dP_{XY}(\tilde{x}, \tilde{y}) \tag{6}$$

always exists and G_φ is an element of $L^2(P_Y)$. It is not difficult to see that

$$S_{YX} : L^2(P_X) \to L^2(P_Y), \qquad \varphi \mapsto G_\varphi$$

is a bounded linear operator with $\|S_{YX}\| \leq E_Y[k_\mathcal{Y}(Y, Y)]$. If f is a function in $\mathcal{H}_\mathcal{X}$, we have for any $y \in \mathcal{Y}$

$$G_f(y) = \langle k_\mathcal{Y}(\,\cdot\,, y), \Sigma_{YX} f \rangle_{\mathcal{H}_\mathcal{Y}} = \big(\Sigma_{YX} f\big)(y),$$

which implies the following proposition:

Proposition 1. *The covariance operator $\Sigma_{YX} : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Y}$ is the restriction of the integral operator S_{YX} to $\mathcal{H}_\mathcal{X}$. More precisely,*

$$J_\mathcal{Y} \Sigma_{YX} = S_{YX} J_\mathcal{X}.$$

Conditional variance can be also represented by covariance operators. Define the *conditional covariance operator* $\Sigma_{YY|X}$ by

$$\Sigma_{YY|X} = \Sigma_{YY} - \Sigma_{YY}^{1/2} V_{YX} V_{XY} \Sigma_{YY}^{1/2},$$

where V_{YX} is the bounded operator in Eq. (5). For convenience we sometimes write $\Sigma_{YY|X}$ as

$$\Sigma_{YY|X} = \Sigma_{YY} - \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XY},$$

which is an abuse of notation, because Σ_{XX}^{-1} may not exist.

The following two propositions provide insights into the meaning of a conditional covariance operator. The former proposition relates the operator to the residual error of regression, and the latter proposition expresses the residual error in terms of the conditional variance.

Proposition 2. *For any $g \in \mathcal{H}_\mathcal{Y}$,*

$$\langle g, \Sigma_{YY|X} g \rangle_{\mathcal{H}_\mathcal{Y}} = \inf_{f \in \mathcal{H}_\mathcal{X}} E_{XY} \left| (g(Y) - E_Y[g(Y)]) - (f(X) - E_X[f(X)]) \right|^2.$$

Proof. Let $\Sigma_{YX} = \Sigma_{YY}^{1/2} V_{YX} \Sigma_{XX}^{1/2}$ be the decomposition in Eq. (5), and define $\mathcal{E}_g(f) = E_{YX} \left| (g(Y) - E_Y[g(Y)]) - (f(X) - E_X[f(X)]) \right|^2$. From the equality

$$\mathcal{E}_g(f) = \|\Sigma_{XX}^{1/2} f\|_{\mathcal{H}_\mathcal{X}}^2 - 2 \langle V_{XY} \Sigma_{YY}^{1/2} g, \Sigma_{XX}^{1/2} f \rangle_{\mathcal{H}_\mathcal{X}} + \|\Sigma_{YY}^{1/2} g\|_{\mathcal{H}_\mathcal{Y}}^2,$$

replacing $\Sigma_{XX}^{1/2} f$ with an arbitrary $\phi \in \mathcal{H}_\mathcal{X}$ yields

$$\begin{aligned}
\inf_{f \in \mathcal{H}_\mathcal{X}} \mathcal{E}_g(f) &\geq \inf_{\phi \in \mathcal{H}_\mathcal{X}} \left\{ \|\phi\|_{\mathcal{H}_\mathcal{X}}^2 - 2 \langle V_{XY} \Sigma_{YY}^{1/2} g, \phi \rangle_{\mathcal{H}_\mathcal{X}} + \|\Sigma_{YY}^{1/2} g\|_{\mathcal{H}_\mathcal{Y}}^2 \right\} \\
&= \inf_{\phi \in \mathcal{H}_\mathcal{X}} \|\phi - V_{XY} \Sigma_{YY}^{1/2} g\|_{\mathcal{H}_\mathcal{X}}^2 + \langle g, \Sigma_{YY|X} g \rangle_{\mathcal{H}_\mathcal{Y}} \\
&= \langle g, \Sigma_{YY|X} g \rangle_{\mathcal{H}_\mathcal{Y}}.
\end{aligned}$$

For the opposite inequality, take an arbitrary $\varepsilon > 0$. From the fact that $V_{XY} \Sigma_{YY}^{1/2} g \in \overline{\mathcal{R}(\Sigma_{XX})} = \mathcal{R}(\Sigma_{XX}^{1/2})$, there exists $f_* \in \mathcal{H}_\mathcal{X}$ such that $\|\Sigma_{XX}^{1/2} f_* - V_{XY} \Sigma_{YY}^{1/2} g\|_{\mathcal{H}_\mathcal{X}} \leq \varepsilon$. For such f_*,

$$\begin{aligned}
\mathcal{E}_g(f_*) &= \|\Sigma_{XX}^{1/2} f_*\|_{\mathcal{H}_\mathcal{X}}^2 - 2 \langle V_{XY} \Sigma_{YY}^{1/2} g, \Sigma_{XX}^{1/2} f_* \rangle_{\mathcal{H}_\mathcal{X}} + \|\Sigma_{YY}^{1/2} g\|_{\mathcal{H}_\mathcal{Y}}^2 \\
&= \|\Sigma_{XX}^{1/2} f_* - V_{YX} \Sigma_{YY}^{1/2} g\|_{\mathcal{H}_\mathcal{X}}^2 + \|\Sigma_{YY}^{1/2} g\|_{\mathcal{H}_\mathcal{Y}} - \|V_{XY} \Sigma_{YY}^{1/2} g\|_{\mathcal{H}_\mathcal{X}}^2 \\
&\leq \langle g, \Sigma_{YY|X} g \rangle_{\mathcal{H}_\mathcal{Y}} + \varepsilon^2.
\end{aligned}$$

Because ε is arbitrary, we have $\inf_{f \in \mathcal{H}_\mathcal{X}} \mathcal{E}_g(f) \leq \langle g, \Sigma_{YY|X} g \rangle_{\mathcal{H}_\mathcal{Y}}$. □

Proposition 2 is an analog for operators of a well-known result on covariance matrices and linear regression: the conditional covariance matrix $C_{YY|X} = C_{YY} - C_{YX} C_{XX}^{-1} C_{XY}$ expresses the residual error of the least square regression problem as $b^T C_{YY|X} b = \min_a E\|b^T Y - a^T X\|^2$.

To relate the residual error in Proposition 2 to the conditional variance of $g(Y)$ given X, we make the following mild assumption:

(AS) $\mathcal{H}_\mathcal{X} + \mathbb{R}$ is dense in $L^2(P_X)$, where $\mathcal{H}_\mathcal{X} + \mathbb{R}$ denotes the direct sum of the RKHS $\mathcal{H}_\mathcal{X}$ and the RKHS $\mathbb{R}$ (Aronszajn, 1950).

A positive definite kernel on a compact set is called *universal* if the corresponding RKHS is dense in the Banach space of continuous functions with sup norm (Steinwart, 2001). The assumption (AS) is satisfied if $\mathcal{X}$ is compact and $k_\mathcal{X}$ is universal. One example of a universal kernel is the Gaussian radial basis function (RBF) kernel $k(x, y) = \exp\left(-\sigma^2 \|x - y\|^2\right)$ on a compact subset of $\mathbb{R}^m$.

Proposition 3. *Under the assumption (AS),*

$$\langle g, \Sigma_{YY|X} g \rangle_{\mathcal{H}_\mathcal{Y}} = E_X\big[\mathrm{Var}_{Y|X}[g(Y)|X]\big] \tag{7}$$

for all $g \in \mathcal{H}_\mathcal{Y}$.

Proof. From Proposition 2, we have

$$\begin{aligned}
\langle g, \Sigma_{YY|X} g \rangle_{\mathcal{H}_\mathcal{Y}} &= \inf_{f \in \mathcal{H}_\mathcal{X}} \mathrm{Var}[g(Y) - f(X)] \\
&= \inf_{f \in \mathcal{H}_\mathcal{X}} \Big\{ \mathrm{Var}_X\big[E_{Y|X}[g(Y) - f(X)|X]\big] + E_X\big[\mathrm{Var}_{Y|X}[g(Y) - f(X)|X]\big] \Big\} \\
&= \inf_{f \in \mathcal{H}_\mathcal{X}} \mathrm{Var}_X\big[E_{Y|X}[g(Y)|X] - f(X)\big] + E_X\big[\mathrm{Var}_{Y|X}[g(Y)|X]\big].
\end{aligned}$$

Let $\varphi(x) = E_{Y|X}[g(Y)|X = x]$. Since $\varphi \in L^2(P_X)$ from $\mathrm{Var}[\varphi(X)] \le \mathrm{Var}[g(Y)] < \infty$, the assumption (AS) implies that for an arbitrary $\varepsilon > 0$ there exists $f \in \mathcal{H}_\mathcal{X}$ and $c \in \mathbb{R}$ such that $h = f + c$ satisfies $\|\varphi - h\|_{L^2(P_X)} < \varepsilon$. Because $\mathrm{Var}[\varphi(X) - f(X)] \le \|\varphi - h\|^2_{L^2(P_X)} \le \varepsilon^2$ and ε is arbitrary, we have $\inf_{f \in \mathcal{H}_\mathcal{X}} \mathrm{Var}_X\big[E_{Y|X}[g(Y)|X] - f(X)\big] = 0$, which completes the proof. □

Proposition 3 improves a result due to Fukumizu et al. (2004, Proposition 5), where the much stronger assumption $E[g(Y)|X = \cdot\,] \in \mathcal{H}_\mathcal{X}$ was imposed.

Propositions 2 and 3 imply that the operator $\Sigma_{YY|X}$ can be interpreted as capturing the predictive ability for Y of the explanatory variable X.

2.2 Criterion of kernel dimension reduction

Let $M(m \times n; \mathbb{R})$ be the set of real-valued $m \times n$ matrices. For a natural number $d \le m$, the Stiefel manifold $\mathbb{S}^m_d(\mathbb{R})$ is defined by

$$\mathbb{S}^m_d(\mathbb{R}) = \{B \in M(m \times d; \mathbb{R}) \mid B^T B = I_d\},$$

which consists of d orthonormal vectors in $\mathbb{R}^m$.[1] It is well known that $\mathbb{S}^m_d(\mathbb{R})$ is a compact smooth manifold. For $B \in \mathbb{S}^m_d(\mathbb{R})$, the matrix BB^T defines an orthogonal projection of $\mathbb{R}^m$ onto the d-dimensional subspace spanned by the column vectors of B.

Hereafter, $\mathcal{X}$ is assumed to be a Borel measurable subset of the m-dimensional Euclidean space such that $BB^T\mathcal{X} \subset \mathcal{X}$ for all $B \in \mathbb{S}^m_d(\mathbb{R})$.

[1] Although the Grassmann manifold is often used in the study of sets of subspaces in $\mathbb{R}^m$, we find the Stiefel manifold more convenient as it allows us to use matrix notation explicitly.

Let $\mathbb{B}_d^m \subseteq \mathbb{S}_d^m(\mathbb{R})$ denote the subset of matrices whose columns span a dimension reduction subspace; for each $B_0 \in \mathbb{B}_d^m$, we have

$$p_{Y|X}(y|x) = p_{Y|B_0^T X}(y|B_0^T x), \tag{8}$$

where $p_{Y|X}(y|x)$ and $p_{Y|B^T X}(y|u)$ are the conditional probability densities of Y given X, and Y given $B^T X$, respectively. The existence and positivity of these conditional probability densities are always assumed hereafter. As we have discussed in Introduction, under conditions given by Cook (1998, Section 6.4) this subset represents the central subspace (under the assumption that d is the minimum dimensionality of the dimension reduction subspaces).

We now turn to the key problem of characterizing the subset $\mathbb{B}_d^m$ using conditional covariance operators on reproducing kernel Hilbert spaces. In the following, we assume that $k_d(z, \tilde{z})$ is a positive definite kernel on $\mathcal{Z} = \cup_{B \in \mathbb{S}_d^m(\mathbb{R})} B^T \mathcal{X}$ such that $E_X[k_d(B^T X, B^T X)] < \infty$ for all $B \in \mathbb{S}_d^m(\mathbb{R})$, and we let $k_{\mathcal{X}}^B$ denote a positive definite kernel on $\mathcal{X}$ given by

$$k_{\mathcal{X}}^B(x, \tilde{x}) = k_d(B^T x, B^T \tilde{x}), \tag{9}$$

for each $B \in \mathbb{S}_d^m(\mathbb{R})$. The RKHS associated with $k_{\mathcal{X}}^B$ is denoted by $\mathcal{H}_{\mathcal{X}}^B$. As seen later in Theorem 4, if $\mathcal{X}$ and $\mathcal{Y}$ are subsets of Euclidean spaces and Gaussian RBF kernels are used for $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$, under some conditions the subset $\mathbb{B}_d^m$ is characterized by the set of solutions of an optimization problem:

$$\mathbb{B}_d^m = \arg \min_{B \in \mathbb{S}_d^m(\mathbb{R})} \Sigma_{YY|X}^B, \tag{10}$$

where Σ_{YX}^B and Σ_{XX}^B denote the (cross-) covariance operators with respect to the kernel k^B, and

$$\Sigma_{YY|X}^B = \Sigma_{YY} - \Sigma_{YX}^B {\Sigma_{XX}^B}^{-1} \Sigma_{XY}^B.$$

The minimization in Eq. (10) refers to the least operator in the partial order of the self-adjoint operators.

We use the trace to evaluate the partial order of the self-adjoint operators. While other possibilities exist (e.g., the determinant), the trace has the advantage of yielding a relatively simple theoretical analysis. The operator $\Sigma_{YY|X}^B$ is trace class for all $B \in \mathbb{S}_d^m(\mathbb{R})$ by $\Sigma_{YY|X}^B \leq \Sigma_{YY}$. Henceforth the minimization in Eq.(10) should thus be understood as that of minimizing $\mathrm{Tr}[\Sigma_{YY|X}^B]$.

From Propositions 2 and 3, minimization of $\mathrm{Tr}[\Sigma_{YY|X}^B]$ is equivalent to the minimization of the sum of the residual errors for the optimal prediction of functions of Y using $B^T X$, where the sum is taken over a complete

orthonormal system of $\mathcal{H}_\mathcal{Y}$. This is intuitively reasonable as a criterion of choosing B, and we will see that this is equivalent to finding the central subspace under some conditions.

Let $(\Omega, \mathcal{B})$ be a measurable space, let $(\mathcal{H}, k)$ be a RKHS over Ω with the kernel k measurable and bounded, and let $\mathcal{S}$ be the set of all probability measures on $(\Omega, \mathcal{B})$. The RKHS $\mathcal{H}$ is called *probability-determining* if the map

$$\mathcal{S} \ni P \quad \mapsto \quad (f \mapsto E_{X\sim P}[f(X)]) \in \mathcal{H}^* \tag{11}$$

is one-to-one, where $\mathcal{H}^*$ is the dual space of $\mathcal{H}$. It is easy to see that $\mathcal{H}$ is probability-determining if and only if the map $\mathcal{S} \ni P \mapsto E_{X\sim P}[k(\cdot, X)] \in \mathcal{H}$ is one-to-one.

Suppose Ω is a topological space equipped with the Borel σ-field. It is known that a finite Borel measure is necessarily a Radon measure for many "nice" spaces such as Polish spaces. From the Riesz representation theorem for Radon measures (see, for example, Berg et al., 1984, Chapter 2), on a locally compact space the linear functional $f \mapsto E_P[f(X)]$ on the space of functions of compact support uniquely determines a Radon probability measure P. Thus, if Ω is a compact Polish space, a universal kernel on Ω is probability-determining. In particular, any universal kernel on a compact subset of Euclidean space is probability-determining. It is also known that Gaussian RBF kernels on all of $\mathbb{R}^m$ are probability-determining (Fukumizu et al., 2004, Theorem 6). Note also that if $\mathcal{X}$ is a finite set of ℓ elements, any positive definite kernel that gives an ℓ-dimensional RKHS is probability-determining.

The following theorem improves Theorem 7 in Fukumizu et al. (2004), and is the theoretical basis of kernel dimension reduction. In the following, let P_B denote the probability on $\mathcal{X}$ induced from P_X by the projection $BB^T : \mathcal{X} \to \mathcal{X}$.

Theorem 4. *Suppose that the closure of the $\mathcal{H}_\mathcal{X}^B$ in $L^2(P_X)$ is included in the closure of $\mathcal{H}_\mathcal{X}$ in $L^2(P_X)$ for any $B \in \mathbb{S}_d^m(\mathbb{R})$. Then,*

$$\Sigma_{YY|X}^B \geq \Sigma_{YY|X}, \tag{12}$$

where the inequality refers to the order of self-adjoint operators. If further $(\mathcal{H}_\mathcal{X}, P_X)$ and $(\mathcal{H}_\mathcal{X}^B, P_B)$ for every $B \in \mathbb{S}_d^m(\mathbb{R})$ satisfy (AS) and $\mathcal{H}_\mathcal{Y}$ is probability-determining, the following equivalence holds

$$\Sigma_{YY|X} = \Sigma_{YY|X}^B \qquad \Longleftrightarrow \qquad Y \perp\!\!\!\perp X \mid B^T X. \tag{13}$$

Proof. The first assertion is obvious from Proposition 2. For the second assertion, let C be an $m \times (m-d)$ matrix whose columns span the orthogonal complement to the subspace spanned by the columns of B, and let $(U, V) = (B^T X, C^T X)$ for notational simplicity. By taking the expectation of the well-known relation

$$\mathrm{Var}_{Y|U}[g(Y)|U] = E_{V|U}\big[\mathrm{Var}_{Y|U,V}[g(Y)|U,V]\big] + \mathrm{Var}_{V|U}\big[E_{Y|U,V}[g(Y)|U,V]\big]$$

with respect to V, we have

$$E_U\big[\mathrm{Var}_{Y|U}[g(Y)|U]\big] = E_X[\mathrm{Var}_{Y|X}[g(Y)|X]] + E_U\big[\mathrm{Var}_{V|U}\big[E_{Y|U,V}[g(Y)|U,V]\big]\big],$$

from which Proposition 3 yields

$$\langle g, (\Sigma^B_{YY|X} - \Sigma_{YY|X})g\rangle_{\mathcal{H}_\mathcal{Y}} = E_U\big[\mathrm{Var}_{V|U}\big[E_{Y|U,V}[g(Y)|U,V]\big]\big].$$

It follows that the right hand side of the equivalence in Eq. (13) holds if and only if $E_{Y|U,V}[g(Y)|U,V]$ does not depend on V almost surely. This is equivalent to

$$E_{Y|X}[g(Y)|X] = E_{Y|U}[g(Y)|U]$$

almost surely. Since $\mathcal{H}_\mathcal{Y}$ is probability-determining, this means that the conditional probability of Y given X is reduced to that of Y given U. □

The assumptions implying Eq. (12) are satisfied if $\mathcal{X}$ is compact and $k_\mathcal{X}$ is universal. Thus, if $\mathcal{X}$ and $\mathcal{Y}$ are compact subsets of Euclidean spaces, universal kernels such as Gaussian RBF kernels are sufficient to guarantee the equivalence given by Eq. (13).

2.3 Kernel dimension reduction procedure

We now use the characterization given in Theorem 4 to develop an optimization procedure for estimating the central subspace from an empirical sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. We assume that $(X_1, Y_1), \ldots, (X_n, Y_n)$ is sampled i.i.d. from P_{XY} and we assume that there exists $B_0 \in \mathbb{S}^m_d(\mathbb{R})$ such that $p_{Y|X}(y|x) = p_{Y|B_0^T X}(y|B_0^T x)$.

We define the *empirical cross-covariance operator* $\widehat{\Sigma}^{(n)}_{YX}$ by evaluating the cross-covariance operator at the empirical distribution $\frac{1}{n}\sum_{i=1}^n \delta_{X_i}\delta_{Y_i}$. When acting on functions $f \in \mathcal{H}_\mathcal{X}$ and $g \in \mathcal{H}_\mathcal{Y}$, the operator $\widehat{\Sigma}^{(n)}_{YX}$ gives the empirical covariance:

$$\langle g, \widehat{\Sigma}^{(n)}_{YX} f\rangle_{\mathcal{H}_\mathcal{Y}} = \frac{1}{n}\sum_{i=1}^n g(Y_i)f(X_i) - \left(\frac{1}{n}\sum_{i=1}^n g(Y_i)\right)\left(\frac{1}{n}\sum_{i=1}^n f(X_i)\right).$$

Also, for $B \in \mathbb{S}_d^m(\mathbb{R})$, let $\widehat{\Sigma}_{YY|X}^{B(n)}$ denote the *empirical conditional covariance operator*:

$$\widehat{\Sigma}_{YY|X}^{B(n)} = \widehat{\Sigma}_{YY}^{(n)} - \widehat{\Sigma}_{YX}^{B(n)} \big(\widehat{\Sigma}_{XX}^{B(n)} + \varepsilon_n I\big)^{-1} \widehat{\Sigma}_{XY}^{B(n)}. \tag{14}$$

The regularization term $\varepsilon_n I$ ($\varepsilon_n > 0$) is required to enable operator inversion and is thus analogous to Tikhonov regularization (Groetsch, 1984). We will see that the regularization term is also needed for consistency.

We now define the KDR estimator $\widehat{B}^{(n)}$ as any minimizer of $\mathrm{Tr}[\widehat{\Sigma}_{YY|X}^{B(n)}]$ on the manifold $\mathbb{S}_d^m(\mathbb{R})$; that is, any matrix in $\mathbb{S}_d^m(\mathbb{R})$ that minimizes

$$\mathrm{Tr}\big[\widehat{\Sigma}_{YX}^{B(n)} \big(\widehat{\Sigma}_{XX}^{B(n)} + \varepsilon_n I\big)^{-1} \widehat{\Sigma}_{XY}^{B(n)}\big]. \tag{15}$$

The KDR objective function in Eq. (15) can also be expressed in terms of Gram matrices (given a kernel k, the *Gram matrix* is the $n \times n$ matrix whose entries are the evaluations of the kernel on all pairs of n data points). Let $\phi_i^B \in \mathcal{H}_\mathcal{X}$ and $\psi_i \in \mathcal{H}_\mathcal{Y}$ ($1 \le i \le n$) be functions defined by

$$\phi_i^B = k^B(\cdot, X_i) - \frac{1}{n}\sum_{j=1}^n k^B(\cdot, X_j), \qquad \psi_i = k_\mathcal{Y}(\cdot, Y_i) - \frac{1}{n}\sum_{j=1}^n k_\mathcal{Y}(\cdot, Y_j).$$

Because $\mathcal{R}(\widehat{\Sigma}_{XX}^{B(n)}) = \mathcal{N}(\widehat{\Sigma}_{XX}^{B(n)})^\perp$ and $\mathcal{R}(\widehat{\Sigma}_{YY}^{(n)}) = \mathcal{N}(\widehat{\Sigma}_{YY}^{(n)})^\perp$ are spanned by $(\phi_i^B)_{i=1}^n$ and $(\psi_i)_{i=1}^n$, respectively, the trace of $\widehat{\Sigma}_{YY|X}^{B(n)}$ is equal to that of the matrix representation of $\widehat{\Sigma}_{YY|X}^{B(n)}$ on the linear hull of $(\psi_i)_{i=1}^n$. Note that although the vectors $(\psi_i)_{i=1}^n$ are over-complete, the trace of the matrix representation with respect to these vectors is equal to the trace of the operator.

For $B \in \mathbb{S}_d^m(\mathbb{R})$, the centered Gram matrix G_X^B with respect to the kernel k^B is defined by

$$(G_X^B)_{ij} = \langle \phi_i^B, \phi_j^B \rangle_{\mathcal{H}_\mathcal{X}^B} = k_\mathcal{X}^B(X_i, X_j) - \frac{1}{n}\sum_{b=1}^n k_\mathcal{X}^B(X_i, X_b) - \frac{1}{n}\sum_{a=1}^n k_\mathcal{X}^B(X_a, X_j) + \frac{1}{n^2}\sum_{a=1}^n\sum_{b=1}^n k_\mathcal{X}^B(X_a, X_b),$$

and G_Y is defined similarly. By direct calculation, it is easy to obtain

$$\widehat{\Sigma}_{YY|X}^{B(n)} \psi_i = \frac{1}{n}\sum_{j=1}^n \psi_j \big(G_Y\big)_{ji} - \frac{1}{n}\sum_{j=1}^n \psi_j \big(G_X^B (G_X^B + n\varepsilon_n I_n)^{-1} G_Y\big)_{ji}.$$

It follows that the matrix representation of $\widehat{\Sigma}^{B(n)}_{YY|X}$ with respect to $(\psi_i)_{i=1}^n$ is $\frac{1}{n}\{G_Y - G_X^B(G_X^B + n\varepsilon_n I_n)^{-1}G_Y\}$ and its trace is

$$\begin{aligned}\mathrm{Tr}\big[\widehat{\Sigma}^{B(n)}_{YY|X}\big] &= \frac{1}{n}\mathrm{Tr}\big[G_Y - G_X^B(G_X^B + n\varepsilon_n I_n)^{-1}G_Y\big] \\ &= \varepsilon_n \mathrm{Tr}\big[G_Y(G_X^B + n\varepsilon_n I_n)^{-1}\big].\end{aligned}$$

Omitting the constant factor, the KDR objective function in Eq. (15) thus reduces to

$$\mathrm{Tr}\big[G_Y(G_X^B + n\varepsilon_n I_n)^{-1}\big]. \tag{16}$$

The KDR method is defined as the optimization of this function over the manifold $\mathbb{S}_d^m(\mathbb{R})$.

Theorem 4 is the population justification of the KDR method. Note that this derivation imposes no strong assumptions either on the conditional probability of Y given X, or on the marginal distributions of X and Y. In particular, it does not require ellipticity of the marginal distribution of X, nor does it require an additive noise model. The response variable Y may be either continuous or discrete. We confirm this general applicability of the KDR method by the numerical results presented in the next section.

Because the objective function Eq. (16) is nonconvex, the minimization requires a nonlinear optimization technique; in our experiments we use the steepest descent method with line search. To alleviate potential problems with local optima, we use a continuation method in which the scale parameter in Gaussian RBF kernel is gradually decreased during the iterative optimization process.

3 Numerical Results

3.1 Simulation studies

In this section we compare the performance of the KDR method with that of several well-known dimension reduction methods. Specifically, we compare to SIR, pHd, and SAVE on synthetic data sets generated by the regressions in Examples 6.2, 6.3, and 6.4 of Li et al. (2005). The results are evaluated by computing the Frobenius distance between the projection matrix of the estimated subspace and that of the true subspace; this evaluation measure is invariant under change of basis and is equal to

$$\|B_0 B_0^T - \widehat{B}\widehat{B}^T\|_F,$$

where B_0 and $\widehat{B}$ are the matrices in the Stiefel manifold $\mathbb{S}_d^m(\mathbb{R})$ representing the true subspace and the estimated subspace, respectively. For the KDR method, a Gaussian RBF kernel $\exp(-\|z_1 - z_2\|^2/c)$ was used, with $c = 2.0$ for regression (A) and regression (C) and $c = 0.5$ for regression (B). The parameter estimate $\widehat{B}$ was updated 100 times by the steepest descent method. The regularization parameter was fixed at $\varepsilon = 0.1$. For SIR and SAVE, we optimized the number of slices for each simulation so as to obtain the best average norm.

Regression (A) is given by

$$\text{(A)} \qquad Y = \frac{X_1}{0.5 + (X_2 + 1.5)^2} + (1 + X_2)^2 + \sigma E,$$

where $X \sim N(0, I_4)$ is a four-dimensional explanatory variable, and $E \sim N(0, 1)$ is independent of X. Thus, the central subspace is spanned by the vectors $(1, 0, 0, 0)$ and $(0, 1, 0, 0)$. For the noise level σ, three different values were used: $\sigma = 0.1, 0.4$ and 0.8. We used 100 random replications with 100 samples each. Note that the distribution of the explanatory variable X satisfies the ellipticity assumption, as required by the SIR, SAVE, and pHd methods.

Table 1 shows the mean and the standard deviation of the Frobenius norm over 100 samples. We see that the KDR method outperforms the other three methods in terms of estimation accuracy. It is also worth noting that in the results presented by Li et al. (2005) for their GCR method, the average norm was 0.28, 0.33, 0.45 for $\sigma = 0.1, 0.4, 0.8$, respectively; again, this is worse than the performance of KDR.

The second regression is given by

$$\text{(B)} \qquad Y = \sin^2(\pi X_2 + 1) + \sigma E,$$

where $X \in \mathbb{R}^4$ is distributed uniformly on the set

$$[0, 1]^4 \backslash \{x \in \mathbb{R}^4 \mid x_i \leq 0.7 \ (i = 1, 2, 3, 4)\},$$

and $E \sim N(0, 1)$ is independent noise. The standard deviation σ is fixed at $\sigma = 0.1, 0.2$ and 0.3. Note that in this example the distribution of X does not satisfy the ellipticity assumption.

Table 2 shows the results of the simulation experiments for this regression. We see that KDR again outperforms the other methods.

The third regression is given by

$$\text{(C)} \qquad Y = \frac{1}{2}(X_1 - a)^2 E,$$

	KDR		SIR		SAVE		pHd	
σ	NORM	SD	NORM	SD	NORM	SD	NORM	SD
0.1	0.11	0.07	0.55	0.28	0.77	0.35	1.04	0.34
0.4	0.17	0.09	0.60	0.27	0.82	0.34	1.03	0.33
0.8	0.34	0.22	0.69	0.25	0.94	0.35	1.06	0.33

Table 1: Comparison of KDR and other methods for regression (A).

	KDR		SIR		SAVE		pHd	
σ	NORM	SD	NORM	SD	NORM	SD	NORM	SD
0.1	0.05	0.02	0.24	0.10	0.23	0.13	0.43	0.19
0.2	0.11	0.06	0.32	0.15	0.29	0.16	0.51	0.23
0.3	0.13	0.07	0.41	0.19	0.41	0.21	0.63	0.29

Table 2: Comparison of KDR and other methods for regression (B).

where $X \sim N(0, I_{10})$ is a ten-dimensional variable and $E \sim N(0, 1)$ is independent noise. The parameter a is fixed at $a = 0, 0.5$ and 1. Note that in this example the conditional probability $p(y|x)$ does not obey an additive noise assumption. The mean of Y is zero and the variance is a quadratic function of X_1. We generated 100 samples of 500 data.

The results for KDR and the other methods are shown by Table 3, in which we again confirm that the KDR method yields significantly better performance than the other methods. In this case, pHd fails to find the true subspace; this is due to the fact that pHd is incapable of estimating a direction that only appears in the variance (Cook and Li, 2002). We note also that the results in Li et al. (2005) show that the contour regression methods SCR and GCR yield average norms larger than 1.3.

Although the estimation of variance structure is generally more difficult than that of estimating mean structure, the KDR method nonetheless is effective at finding the central subspace in this case.

	KDR		SIR		SAVE		pHd	
a	NORM	SD	NORM	SD	NORM	SD	NORM	SD
0.0	0.17	0.05	1.83	0.22	0.30	0.07	1.48	0.27
0.5	0.17	0.04	0.58	0.19	0.35	0.08	1.52	0.28
1.0	0.18	0.05	0.30	0.08	0.57	0.20	1.58	0.28

Table 3: Comparison of KDR and other methods for regression (C).

3.2 Applications

We apply the KDR method to two data sets; one is binary classification and the other is regression with a continuous response variable. These data sets have been used previously in studies of dimension reduction methods.

The first data set that we studied is *Swiss bank notes* which has been previously studied in the dimension reduction context by Cook and Lee (1999), with the data taken from Flury and Riedwyl (1988). The problem is that of classifying counterfeit and genuine Swiss bank notes. The data is a sample of 100 counterfeit and 100 genuine notes. There are 6 continuous explanatory variables that represent aspects of the size of a note: length, height on the left, height on the right, distance of inner frame to the lower border, distance of inner frame to the upper border, and length of the diagonal. We standardize each of explanatory variables so that their standard deviation is 5.0.

As we have discussed in the Introduction, many dimension reduction methods (including SIR) are not generally suitable for binary classification problems. Because among inverse regression methods the estimated subspace given by SAVE is necessarily larger than that given by pHd and SIR (Cook and Lee, 1999), we compared the KDR method only with SAVE for this data set.

Figure 1 shows two-dimensional plots of the data projected onto the subspaces estimated by the KDR method and by SAVE. The figure shows that the results for KDR appear to be robust with respect to the values of the scale parameter a in the Gaussian RBF kernel. (Note that if a goes to infinity, the result approaches that obtained by a linear kernel, since the linear term in the Taylor expansion of the exponential function is dominant.) In the KDR case, using a Gaussian RBF with scale parameter $a = 10$ and 100 we obtain clear separation of genuine and counterfeit notes. Slightly less separation is obtained for the Gaussian RBF kernel with $a = 10,000$, for the linear kernel, and for SAVE; in these cases there is an isolated genuine data point that lies close to the class boundary, which is similar to the results using linear discriminant analysis and specification analysis (Flury and Riedwyl, 1988). We see that KDR finds a more effective subspace to separate the two classes than SAVE and the existing analysis. Finally, note that there are two clusters of counterfeit notes in the result of SAVE, while KDR does not show multiple clusters in either class. Although clusters have also been reported in other analyses (Flury and Riedwyl, 1988, Section 12), the KDR results suggest that the cluster structure may not be relevant to the classification.

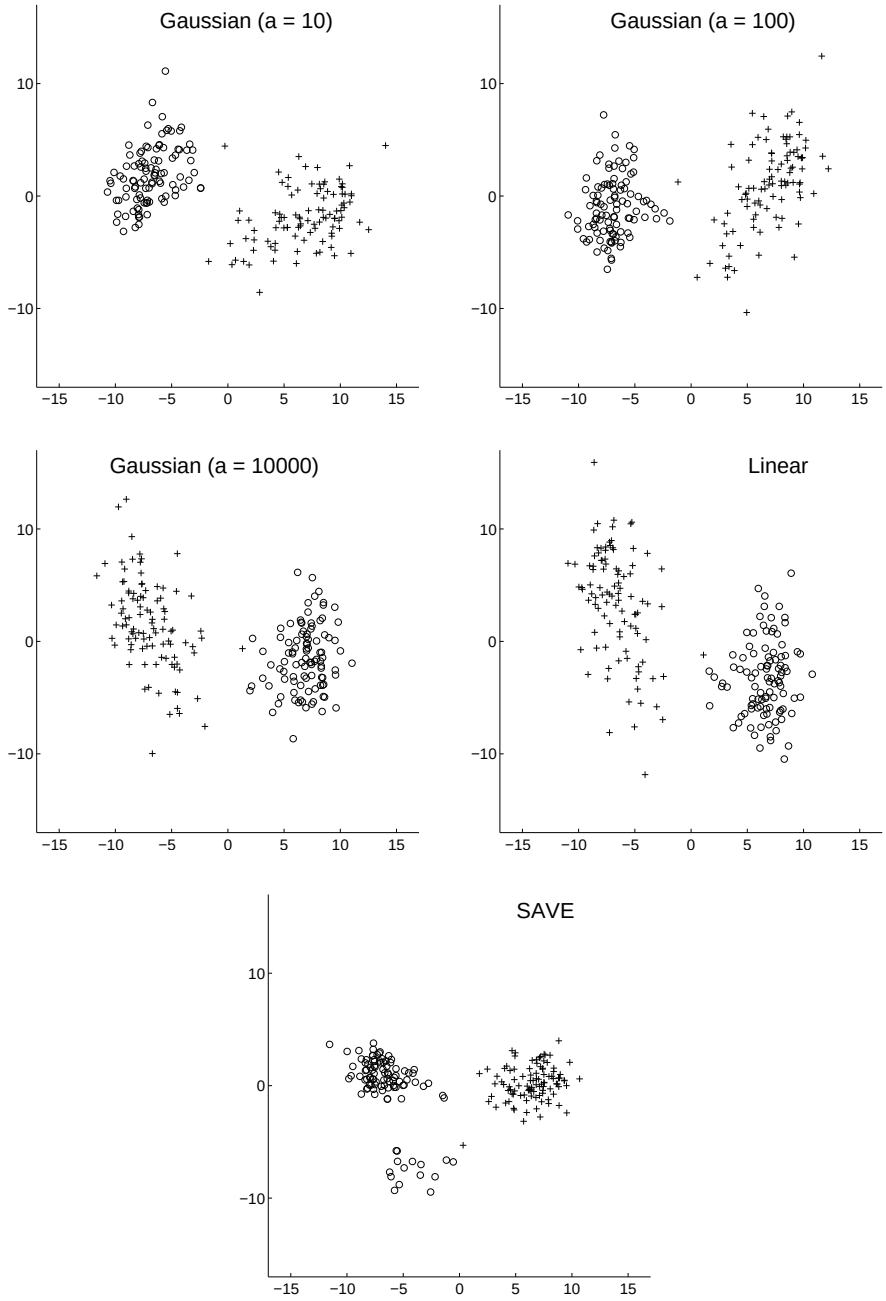

Figure 1: Two dimensional plots of *Swiss bank notes*. The crosses and circles show genuine and counterfeit notes, respectively. For the KDR methods, the Gaussian RBF kernel $\exp(-\|z_1 - z_2\|^2/a)$ is used with $a = 10, 100$ and 10000. For comparison, the plots given by KDR with a linear kernel and SAVE are shown.

We also analyzed the *Evaporation* data set, available in the *Arc* package (http://www.stat.umn.edu/arc/software.html). The data set is concerned with the effect on soil evaporation of various air and soil conditions. The number of explanatory variables is ten: maximum daily soil temperature (Maxst), minimum daily soil temperature (Minst), area under the daily soil temperature curve (Avst), maximum daily air temperature (Maxat), minimum daily air temperature (Minat), average daily air temperature (Avat), maximum daily humidity (Maxh), minimum daily humidity (Minh), area under the daily humidity curve (Avh), and total wind speed in miles/hour (Wind). The response variable is daily soil evaporation (Evap). The data were collected daily during 46 days; thus the number of data points is 46. This data set was studied in the context of contour regression methods for dimension reduction in Li et al. (2005). We standardize each variable so that the sample variance is equal to 5.0, and use the Gaussian RBF kernel $\exp\bigl(-\|z_1 - z_2\|^2/10\bigr)$.

Our analysis yielded an estimated two-dimensional subspace which is spanned by the vectors:

$$\begin{aligned} KDR_1 : & -0.25MAXST + 0.32MINST + 0.00AVST + (-0.28)MAXAT \\ & + (-0.23)MINAT + (-0.44)AVAT + 0.39MAXH + 0.25MINH \\ & + (-0.07)AVH + (-0.54)WIND. \\ KDR_2 : & 0.09MAXST + (-0.02)MINST + 0.00AVST + 0.10MAXAT \\ & + (-0.45)MINAT + 0.23AVAT + 0.21MAXH + (-0.41)MINH \\ & + (-0.71)AVH + (-0.05)WIND. \end{aligned}$$

In the first direction, Wind and Avat have a large factor with the same sign, while both have weak contributions on the second direction. In the second direction, Avh is dominant.

Figure 2 presents the scatter plots representing the response Y plotted with respect to each of the first two directions given by the KDR method. Both of these directions show a clear relation with Y. Figure 3 presents the scatter plot of Y versus the two-dimensional subspace found by KDR. The obtained two dimensional subspace is different from the one given by the existing analysis in Li et al. (2005); the contour regression method gives a subspace of which the first direction shows a clear monotonic trend, but the second direction suggests a U-shaped pattern. In the result of KDR, we do not see a clear folded pattern. Although without further analysis it is difficult to say which result expresses more clearly the statistical dependence, the plots suggest that the KDR method successfully captured the effective directions for regression.

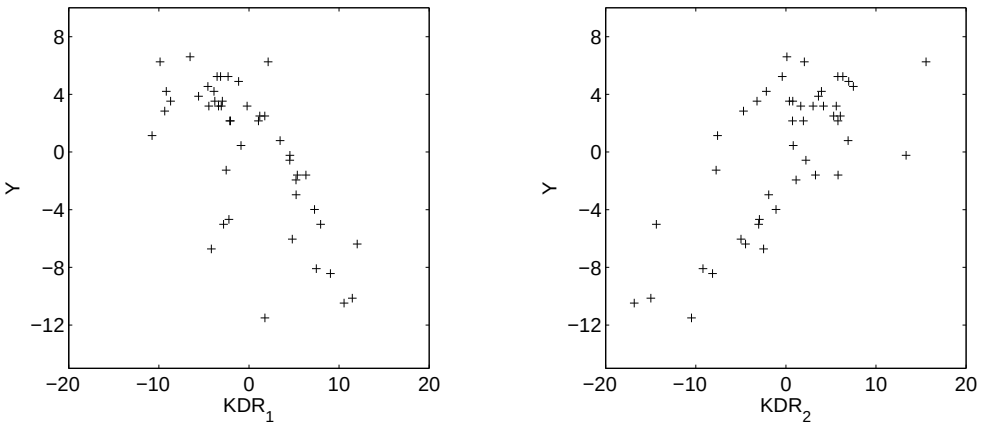

Figure 2: Two dimensional representation of *Evaporation* data for each of the first two directions

4 Consistency of kernel dimension reduction

In this section we prove that the KDR estimator is consistent. Our proof of consistency requires tools from empirical process theory, suitably elaborated to handle the RKHS setting. We establish convergence of the empirical objective function to the population objective function under a condition on the regularization coefficient ε_n, and from this result infer the consistency of $\widehat{B}^{(n)}$.

4.1 Main result

We assume hereafter that $\mathcal{Y}$ is a topological space. The Stiefel manifold $\mathbb{S}_d^m(\mathbb{R})$ is assumed to be equipped with a distance D which is compatible with the topology of $\mathbb{S}_d^m(\mathbb{R})$. It is known that geodesics define such a distance (see, for example, Kobayashi and Nomizu, 1963, Chapter IV).

The following technical assumptions are needed to guarantee the consistency of kernel dimension reduction:

(A-1) For any bounded continuous function g on $\mathcal{Y}$, the function

$$B \mapsto E_X\left[E_{Y|B^TX}[g(Y)|B^TX]^2\right]$$

is continuous on $\mathbb{S}_d^m(\mathbb{R})$.

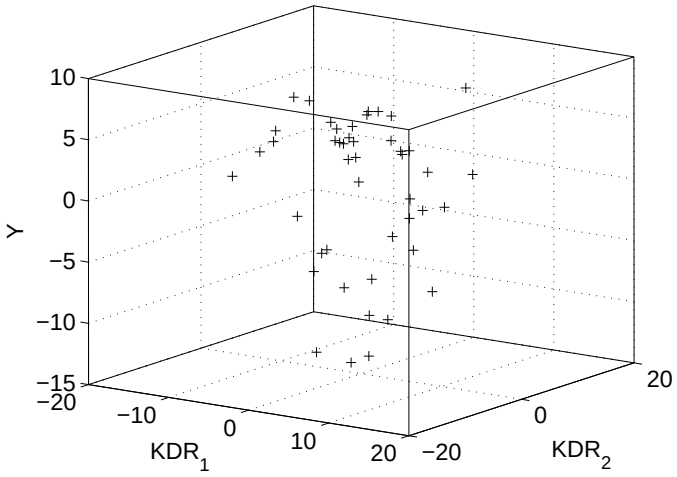

Figure 3: Three dimensional representation of *Evaporation* data.

(A-2) For $B \in \mathbb{S}_d^m(\mathbb{R})$, let P_B be the probability distribution of the random variable BB^TX on $\mathcal{X}$. The Hilbert space $\mathcal{H}_{\mathcal{X}}^B + \mathbb{R}$ is dense in $L^2(P_B)$ for any $B \in \mathbb{S}_d^m(\mathbb{R})$.

(A-3) There exists a measurable function $\phi : \mathcal{X} \to \mathbb{R}$ such that $E|\phi(X)|^2 < \infty$ and the Lipschitz condition

$$\|k_d(B^Tx, \cdot) - k_d(\tilde{B}^Tx, \cdot)\|_{\mathcal{H}_d} \leq \phi(x)D(B, \tilde{B})$$

holds for all $B, \tilde{B} \in \mathbb{S}_d^m(\mathbb{R})$ and $x \in \mathcal{X}$.

Theorem 5. *Suppose k_d in Eq. (9) is continuous and bounded, and suppose the regularization parameter ε_n in Eq. (14) satisfies*

$$\varepsilon_n \to 0, \qquad n^{1/2}\varepsilon_n \to \infty \qquad (n \to \infty). \tag{17}$$

Define the set of the optimum parameters $\mathbb{B}_d^m$ by

$$\mathbb{B}_d^m = \arg \min_{B \in \mathbb{S}_d^m(\mathbb{R})} \Sigma_{YY|X}^B.$$

Under the assumptions (A-1)*,* (A-2)*, and* (A-3)*, the set $\mathbb{B}_d^m$ is nonempty, and for an arbitrary open set U in $\mathbb{S}_d^m(\mathbb{R})$ with $\mathbb{B}_d^m \subset U$ we have*

$$\lim_{n\to\infty} \Pr\big(\widehat{B}^{(n)} \in U\big) = 1.$$

The assumptions (A-1) and (A-2) are used to establish the continuity of $\mathrm{Tr}[\Sigma^B_{YY|X}]$ in Lemma 12, and (A-3) is needed to derive the order of uniform convergence of $\widehat{\Sigma}^{B(n)}_{YY|X}$ in Lemma 8.

The assumption (A-1) is satisfied in various cases. Let $f(x) = E_{Y|X}[g(Y)|X = x]$, and assume $f(x)$ is continuous. This assumption holds, for example, if the conditional probability density $p_{Y|X}(y|x)$ is bounded and continuous on x. Let C be an element of $\mathbb{S}^m_{m-d}(\mathbb{R})$ such that the subspaces spanned by the column vectors of B and C are orthogonal; that is, the $m \times m$ matrix (B, C) is an orthogonal matrix. Define random variables U and V by $U = B^T X$ and $V = C^T X$. If X has the probability density function $p_X(x)$, the probability density function of (U, V) is given by $p_{U,V}(u, v) = p_X(Bu + Cv)$. Consider the situation in which u is given by $u = B^T\tilde{x}$ for $B \in \mathbb{S}^m_d(\mathbb{R})$ and $\tilde{x} \in \mathcal{X}$, and let $\mathcal{V}_{B,\tilde{x}} = \{v \in \mathbb{R}^{m-d} \mid BB^T\tilde{x} + Cv \in \mathcal{X}\}$. We have

$$E[g(Y)|B^T X = B^T\tilde{x}] = \frac{\int_{\mathcal{V}_{B,\tilde{x}}} f(BB^T\tilde{x} + Cv)p_X(BB^T\tilde{x} + Cv)dv}{\int_{\mathcal{V}_{B,\tilde{x}}} p_X(BB^T\tilde{x} + Cv)dv}.$$

If there exists an integrable function $r(v)$ such that $\chi_{\mathcal{V}_{B,\tilde{x}}}(v)p_X(BB^T\tilde{x} + Cv) \leq r(v)$ for all $B \in \mathbb{S}^m_d(\mathbb{R})$ and $\tilde{x} \in \mathcal{X}$, the dominated convergence theorem ensures (A-1). Thus, it is easy to see that a sufficient condition for (A-1) is that $\mathcal{X}$ is bounded, $p_X(x)$ is bounded, and $p_{Y|X}(y|x)$ is bounded and continuous on x, which is satisfied by a wide class of distributions.

The assumption (A-2) holds if $\mathcal{X}$ is compact and $k_d + 1$ is a universal kernel on $\mathcal{Z}$. The assumption (A-3) is satisfied by many useful kernels; for example, kernels with the property

$$\left|\frac{\partial^2}{\partial z_a \partial z_b} k_d(z_1, z_2)\right| \leq L\|z_1 - z_2\| \qquad (a, b = 1, 2),$$

for some $L > 0$. In particular Gaussian RBF kernels satisfy this property.

4.2 Proof of the consistency theorem

If the following proposition is shown, Theorem 5 follows straightforwardly by standard arguments establishing the consistency of M-estimators (see, for example, van der Vaart, 1998, Section 5.2).

Proposition 6. *Under the same assumptions as Theorem 5, the functions* $\mathrm{Tr}\big[\widehat{\Sigma}^{B(n)}_{YY|X}\big]$ *and* $\mathrm{Tr}\big[\Sigma^B_{YY|X}\big]$ *are continuous on* $\mathbb{S}^m_d(\mathbb{R})$*, and*

$$\sup_{B \in \mathbb{S}^m_d(\mathbb{R})} \left|\mathrm{Tr}\big[\widehat{\Sigma}^{B(n)}_{YY|X}\big] - \mathrm{Tr}\big[\Sigma^B_{YY|X}\big]\right| \quad \to \quad 0 \qquad (n \to \infty)$$

in probability.

The proof of Proposition 6 is divided into several lemmas. We decompose $\sup_B \left|\mathrm{Tr}[\Sigma^B_{YY|X}] - \mathrm{Tr}[\widehat{\Sigma}^{B(n)}_{YY|X}]\right|$ into two parts: $\sup_B \left|\mathrm{Tr}[\Sigma^B_{YY|X}] - \mathrm{Tr}[\Sigma^B_{YX}(\Sigma^B_{XX}+\varepsilon_n I)^{-1}\Sigma^B_{XY}]\right|$ and $\sup_B \left|\mathrm{Tr}[\Sigma^B_{YX}(\Sigma^B_{XX}+\varepsilon_n I)^{-1}\Sigma^B_{XY}] - \mathrm{Tr}[\widehat{\Sigma}^{B(n)}_{YY|X}]\right|$. Lemmas 7, 8, and 9 establish the convergence of the first part. The convergence of the second part is shown by Lemmas 10–13; in particular, Lemmas 11 and 12 establish the key result that the trace of the population conditional covariance operator is a continuous function of B.

The following lemmas make use of the trace norm and the Hilbert-Schmidt norm of operators. Recall that the *trace* of a positive operator A on a Hilbert space $\mathcal{H}$ is defined by

$$\mathrm{Tr}[A] = \sum_{i=1}^{\infty} \langle \varphi_i, A\varphi_i \rangle_{\mathcal{H}},$$

where $\{\varphi_i\}_{i=1}^{\infty}$ is a complete orthonormal system (CONS) of $\mathcal{H}$. A bounded operator T on a Hilbert space $\mathcal{H}$ is called *trace class* if $\mathrm{Tr}[(T^*T)^{1/2}]$ is finite. The set of all trace class operators on a Hilbert space is a Banach space with the trace norm $\|T\|_{tr} = \mathrm{Tr}[(T^*T)^{1/2}]$. A bounded operator $T : \mathcal{H}_1 \to \mathcal{H}_2$, where $\mathcal{H}_1$ and $\mathcal{H}_2$ are Hilbert spaces, is called *Hilbert-Schmidt* if $\mathrm{Tr}[T^*T] < \infty$, or equivalently, $\sum_{i=1}^{\infty} \|T\varphi_i\|^2_{\mathcal{H}_2} < \infty$ for a CONS $\{\varphi_i\}_{i=1}^{\infty}$ of $\mathcal{H}_1$. The set of all Hilbert-Schmidt operators from $\mathcal{H}_1$ to $\mathcal{H}_2$ is a Hilbert space with Hilbert-Schmidt inner product

$$\langle T_1, T_2 \rangle_{HS} = \sum_{i=1}^{\infty} \langle T_1\varphi_i, T_2\varphi_i \rangle_{\mathcal{H}_2},$$

where $\{\varphi_i\}_{i=1}^{\infty}$ is a CONS of $\mathcal{H}_1$. Thus, the Hilbert-Schmidt norm $\|T\|_{HS}$ satisfies $\|T\|^2_{HS} = \sum_{i=1}^{\infty} \|T\varphi_i\|^2_{\mathcal{H}_2}$.

Obviously, $\|T\| \leq \|T\|_{HS} \leq \|T\|_{tr}$ holds, if T is trace class or Hilbert-Schmidt. Recall also $\|AB\|_{tr} \leq \|A\| \, \|B\|_{tr}$ ($\|AB\|_{HS} \leq \|A\| \, \|B\|_{HS}$) for a bounded operator A and a trace class (Hilbert-Schmidt, resp.) operator B. If $A : \mathcal{H}_1 \to \mathcal{H}_2$ and $B : \mathcal{H}_2 \to \mathcal{H}_1$ are Hilbert-Schmidt, the product AB is trace-class with $\|AB\|_{tr} \leq \|A\|_{HS}\|B\|_{HS}$.

It is known that cross-covariance operators and covariance operators are Hilbert-Schmidt and trace class, respectively, under the assumption Eq. (2) (Gretton et al., 2005; Fukumizu et al., 2005). The Hilbert-Schmidt norm of Σ_{YX} is given by

$$\|\Sigma_{YX}\|^2_{HS} = \left\| E_{YX}[(k_{\mathcal{X}}(\cdot, X) - m_X)(k_{\mathcal{Y}}(\cdot, Y) - m_Y)] \right\|^2_{\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}}, \qquad (18)$$

where $\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}$ is the direct product of $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$, and the trace norm of Σ_{XX} is

$$\mathrm{Tr}[\Sigma_{XX}] = E_X\big[\|k_{\mathcal{X}}(\,\cdot\,, X) - m_X\|^2_{\mathcal{H}_{\mathcal{X}}}\big]. \tag{19}$$

Lemma 7.

$$\begin{aligned}
&\Big|\mathrm{Tr}\big[\widehat{\Sigma}^{(n)}_{YY|X}\big] - \mathrm{Tr}\big[\Sigma_{YY} - \Sigma_{YX}\big(\Sigma_{XX} + \varepsilon_n I\big)^{-1}\Sigma_{XY}\big]\Big| \\
&\le \frac{1}{\varepsilon_n}\Big\{\big(\big\|\widehat{\Sigma}^{(n)}_{YX}\big\|_{HS} + \big\|\Sigma_{YX}\big\|_{HS}\big)\big\|\widehat{\Sigma}^{(n)}_{YX} - \Sigma_{YX}\big\|_{HS} + \big\|\Sigma_{YY}\big\|_{tr}\big\|\widehat{\Sigma}^{(n)}_{XX} - \Sigma_{XX}\big\|\Big\} \\
&\quad + \big|\mathrm{Tr}\big[\widehat{\Sigma}^{(n)}_{YY} - \Sigma_{YY}\big]\big|.
\end{aligned}$$

Proof. Noting that the self-adjoint operator $\Sigma_{YX}(\Sigma_{XX}+\varepsilon_n I)^{-1}\Sigma_{XY}$ is trace class from $\Sigma_{YX}(\Sigma_{XX}+\varepsilon_n I)^{-1}\Sigma_{XY} \le \Sigma_{YY}$, the left hand side of the assertion is bounded from above by

$$\big|\mathrm{Tr}\big[\widehat{\Sigma}^{(n)}_{YY} - \Sigma_{YY}\big]\big| + \big|\mathrm{Tr}\big[\widehat{\Sigma}^{(n)}_{YX}\big(\widehat{\Sigma}^{(n)}_{XX} + \varepsilon_n I\big)^{-1}\widehat{\Sigma}^{(n)}_{XY} - \Sigma_{YX}\big(\Sigma_{XX} + \varepsilon_n I\big)^{-1}\Sigma_{XY}\big]\big|.$$

The second term is upper-bounded by

$$\begin{aligned}
&\big|\mathrm{Tr}\big[\big(\widehat{\Sigma}^{(n)}_{YX} - \Sigma_{YX}\big)\big(\widehat{\Sigma}^{(n)}_{XX} + \varepsilon_n I\big)^{-1}\widehat{\Sigma}^{(n)}_{XY}\big]\big| \\
&\quad + \big|\mathrm{Tr}\big[\Sigma_{YX}\big(\widehat{\Sigma}^{(n)}_{XX} + \varepsilon_n I\big)^{-1}\big(\widehat{\Sigma}^{(n)}_{XY} - \Sigma_{XY}\big)\big]\big| \\
&\quad + \big|\mathrm{Tr}\big[\Sigma_{YX}\big\{\big(\widehat{\Sigma}^{(n)}_{XX} + \varepsilon_n I\big)^{-1} - \big(\Sigma_{XX} + \varepsilon_n I\big)^{-1}\big\}\Sigma_{XY}\big]\big| \\
&\le \big\|\big(\widehat{\Sigma}^{(n)}_{YX} - \Sigma_{YX}\big)\big(\widehat{\Sigma}^{(n)}_{XX} + \varepsilon_n I\big)^{-1}\widehat{\Sigma}^{(n)}_{XY}\big\|_{tr} \\
&\quad + \big\|\Sigma_{YX}\big(\widehat{\Sigma}^{(n)}_{XX} + \varepsilon_n I\big)^{-1}\big(\widehat{\Sigma}^{(n)}_{XY} - \Sigma_{XY}\big)\big\|_{tr} \\
&\quad + \Big|\mathrm{Tr}\big[\big\{\big(\Sigma_{XX} + \varepsilon_n I\big)^{1/2}\big(\widehat{\Sigma}^{(n)}_{XX} + \varepsilon_n I\big)^{-1}\big(\Sigma_{XX} + \varepsilon_n I\big)^{1/2} - I\big\} \\
&\qquad\qquad \times \big(\Sigma_{XX} + \varepsilon_n I\big)^{-1/2}\Sigma_{XY}\Sigma_{YX}\big(\Sigma_{XX} + \varepsilon_n I\big)^{-1/2}\big]\Big| \\
&\le \frac{1}{\varepsilon_n}\big\|\widehat{\Sigma}^{(n)}_{YX} - \Sigma_{YX}\|_{HS}\big\|\widehat{\Sigma}^{(n)}_{XY}\big\|_{HS} + \frac{1}{\varepsilon_n}\big\|\Sigma_{YX}\big\|_{HS}\big\|\widehat{\Sigma}^{(n)}_{XY} - \Sigma_{XY}\big\|_{HS} \\
&\quad + \big\|\big(\Sigma_{XX} + \varepsilon_n I\big)^{1/2}\big(\widehat{\Sigma}^{(n)}_{XX} + \varepsilon_n I\big)^{-1}\big(\Sigma_{XX} + \varepsilon_n I\big)^{1/2} - I\big\| \\
&\qquad\qquad \times \big\|\big(\Sigma_{XX} + \varepsilon_n I\big)^{-1/2}\Sigma_{XY}\Sigma_{YX}\big(\Sigma_{XX} + \varepsilon_n I\big)^{-1/2}\big\|_{tr}.
\end{aligned}$$

Since the spectrum of A^*A and AA^* are identical, we have

$$\begin{aligned}
&\big\|\big(\Sigma_{XX} + \varepsilon_n I\big)^{1/2}\big(\widehat{\Sigma}^{(n)}_{XX} + \varepsilon_n I\big)^{-1}\big(\Sigma_{XX} + \varepsilon_n I\big)^{1/2} - I\big\| \\
&= \big\|\big(\widehat{\Sigma}^{(n)}_{XX} + \varepsilon_n I\big)^{-1/2}\big(\Sigma_{XX} + \varepsilon_n I\big)\big(\widehat{\Sigma}^{(n)}_{XX} + \varepsilon_n I\big)^{-1/2} - I\big\| \\
&\le \big\|\big(\widehat{\Sigma}^{(n)}_{XX} + \varepsilon_n I\big)^{-1/2}\big(\Sigma_{XX} - \widehat{\Sigma}^{(n)}_{XX}\big)\big(\widehat{\Sigma}^{(n)}_{XX} + \varepsilon_n I\big)^{-1/2}\big\| \\
&\le \frac{1}{\varepsilon_n}\big\|\widehat{\Sigma}^{(n)}_{XX} - \Sigma_{XX}\big\|.
\end{aligned}$$

The bound $\|(\Sigma_{XX}+\varepsilon_n I)^{-1/2}\Sigma_{XX}^{1/2}V_{XY}\| \leq 1$ yields

$$\left\|\left(\Sigma_{XX}+\varepsilon_n I\right)^{-1/2}\Sigma_{XY}\Sigma_{YX}\left(\Sigma_{XX}+\varepsilon_n I\right)^{-1/2}\right\|_{tr} \leq \left\|\Sigma_{YY}\right\|_{tr},$$

which concludes the proof. □

Lemma 8. *Under the assumption (A-3),*

$$\sup_{B\in\mathbb{S}_d^m(\mathbb{R})}\left\|\widehat{\Sigma}_{XX}^{B(n)}-\Sigma_{XX}^B\right\|_{HS}, \quad \sup_{B\in\mathbb{S}_d^m(\mathbb{R})}\left\|\widehat{\Sigma}_{XY}^{B(n)}-\Sigma_{XY}^B\right\|_{HS},$$
$$\textit{and} \quad \sup_{B\in\mathbb{S}_d^m(\mathbb{R})}\left|\mathrm{Tr}\left[\widehat{\Sigma}_{YY}^{B(n)}-\Sigma_{YY}^B\right]\right|$$

are of order $O_p(1/\sqrt{n})$ *as* $n\to\infty$.

The proof of Lemma 8 is deferred to the Appendix. From Lemmas 7 and 8, the following lemma is obvious.

Lemma 9. *If the regularization parameter* $(\varepsilon_n)_{n=1}^\infty$ *satisfies Eq. (17), under the assumption (A-3) we have*

$$\sup_{B\in\mathbb{S}_d^m(\mathbb{R})}\left|\mathrm{Tr}\left[\widehat{\Sigma}_{YY|X}^{B(n)}\right]-\mathrm{Tr}\left[\Sigma_{YY}-\Sigma_{YX}^B\left(\Sigma_{XX}^B+\varepsilon_n I\right)^{-1}\Sigma_{XY}^B\right]\right| = O_p\left(\varepsilon_n^{-1}n^{-1/2}\right),$$

as $n\to\infty$.

In the next four lemmas, we establish the uniform convergence of L_ε to L_0 ($\varepsilon\downarrow 0$), where $L_\varepsilon(B)$ is a function on $\mathbb{S}_d^m(\mathbb{R})$ defined by

$$L_\varepsilon(B)=\mathrm{Tr}\left[\Sigma_{YX}^B\left(\Sigma_{XX}^B+\varepsilon I\right)^{-1}\Sigma_{XY}^B\right],$$

for $\varepsilon>0$ and $L_0(B)=\mathrm{Tr}[\Sigma_{YY}^{1/2}V_{YX}^BV_{XY}^B\Sigma_{YY}^{1/2}]$. We begin by establishing pointwise convergence.

Lemma 10. *For arbitrary kernels with Eq. (2),*

$$\mathrm{Tr}\left[\Sigma_{YX}\left(\Sigma_{XX}+\varepsilon I\right)^{-1}\Sigma_{XY}\right]\to\mathrm{Tr}\left[\Sigma_{YY}^{1/2}V_{YX}V_{XY}\Sigma_{YY}^{1/2}\right] \qquad (\varepsilon\downarrow 0).$$

Proof. With a CONS $\{\psi_i\}_{i=1}^\infty$ for $\mathcal{H}_\mathcal{Y}$, the left hand side can be written as

$$\sum_{i=1}^\infty\langle\psi_i,\Sigma_{YY}^{1/2}V_{YX}\left\{I-\Sigma_{XX}^{1/2}(\Sigma_{XX}+\varepsilon I)^{-1}\Sigma_{XX}^{1/2}\right\}V_{XY}\Sigma_{YY}^{1/2}\psi_i\rangle_{\mathcal{H}_\mathcal{Y}}.$$

Since each summand is positive and upper bounded by $\langle \psi_i, \Sigma_{YY}^{1/2} V_{YX} V_{XY} \Sigma_{YY}^{1/2} \psi_i \rangle_{\mathcal{H}_\mathcal{Y}}$, and the sum over i is finite, by the dominated convergence theorem it suffices to show

$$\lim_{\varepsilon \downarrow 0} \langle \psi, \Sigma_{YY}^{1/2} V_{YX} \{ I - \Sigma_{XX}^{1/2} (\Sigma_{XX} + \varepsilon I)^{-1} \Sigma_{XX}^{1/2} \} V_{XY} \Sigma_{YY}^{1/2} \psi \rangle_{\mathcal{H}_\mathcal{Y}} = 0,$$

for each $\psi \in \mathcal{H}_\mathcal{Y}$.

Fix arbitrary $\psi \in \mathcal{H}_\mathcal{Y}$ and $\delta > 0$. From the fact $\mathcal{R}(V_{XY}) \subset \overline{\mathcal{R}(\Sigma_{XX})}$, there exists $h \in \mathcal{H}_\mathcal{X}$ such that $\| V_{XY} \Sigma_{YY}^{1/2} \psi - \Sigma_{XX} h \|_{\mathcal{H}_\mathcal{X}} < \delta$. Using the fact $I - \Sigma_{XX}^{1/2} (\Sigma_{XX} + \varepsilon_n I)^{-1} \Sigma_{XX}^{1/2} = \varepsilon_n (\Sigma_{XX} + \varepsilon_n I)^{-1}$, we have

$$\begin{aligned}
&\big\| \{ I - \Sigma_{XX}^{1/2} (\Sigma_{XX} + \varepsilon I)^{-1} \Sigma_{XX}^{1/2} \} V_{XY} \Sigma_{YY}^{1/2} \psi \big\|_{\mathcal{H}_\mathcal{X}} \\
&= \big\| \varepsilon \big(\Sigma_{XX} + \varepsilon I \big)^{-1} \Sigma_{XX} h \big\|_{\mathcal{H}_\mathcal{X}} + \big\| \varepsilon \big(\Sigma_{XX} + \varepsilon I \big)^{-1} \big(V_{XY} \Sigma_{YY}^{1/2} \psi - \Sigma_{XX} h \big) \big\|_{\mathcal{H}_\mathcal{X}} \\
&\leq \varepsilon \| h \|_{\mathcal{H}_\mathcal{X}} + \delta,
\end{aligned}$$

which is arbitrary small if ε is sufficiently small. This completes the proof. □

Lemma 11. *Suppose k_d is continuous and bounded. Then, for any $\varepsilon > 0$, the function $L_\varepsilon(B)$ is continuous on $\mathbb{S}_d^m(\mathbb{R})$.*

Proof. By an argument similar to that in the proof of Lemma 10, it suffices to show the continuity of $B \mapsto \langle \psi, \Sigma_{YX}^B (\Sigma_{XX}^B + \varepsilon I)^{-1} \Sigma_{XY}^B \psi \rangle_{\mathcal{H}_\mathcal{Y}}$ for each $\psi \in \mathcal{H}_\mathcal{Y}$.

Let $J_X^B : \mathcal{H}_X^B \to L^2(P_X)$ and $J_Y : \mathcal{H}_\mathcal{Y} \to L^2(P_Y)$ be the inclusions. As seen in Proposition 1, the operators Σ_{YX}^B and Σ_{XX}^B can be extended to the integral operators S_{YX}^B and S_{XX}^B on $L^2(P_X)$, respectively, so that $J_Y \Sigma_{YX}^B = S_{YX}^B J_X^B$ and $J_X^B \Sigma_{XX}^B = S_{XX}^B J_X^B$. It is not difficult to see also $J_X^B (\Sigma_{XX}^B + \varepsilon I)^{-1} = (S_{XX}^B + \varepsilon I)^{-1} J_X^B$ for $\varepsilon > 0$. These relations yield

$$\begin{aligned}
\langle \psi, \Sigma_{YX}^B \big(\Sigma_{XX}^B + \varepsilon I \big)^{-1} \Sigma_{XY}^B \psi \rangle_{\mathcal{H}_\mathcal{Y}} = \ & E_{XY} \big[\psi(Y) \big(\big(S_{XX}^B + \varepsilon I \big)^{-1} S_{XY}^B \psi \big)(X) \big] \\
& - E_Y[\psi(Y)] E_X \big[\big(\big(S_{XX}^B + \varepsilon I \big)^{-1} S_{XY}^B \psi \big)(X) \big],
\end{aligned}$$

where $J_Y \psi$ is identified with ψ. The assertion is obtained if we prove that the operators S_{XY}^B and $(S_{XX}^B + \varepsilon I)^{-1}$ are continuous with respect to B in operator norm. To see this, let $\tilde{X}$ be identically and independently distributed with X. We have

$$\begin{aligned}
\big\| (S_{XY}^B - S_{XY}^{B_0}) \psi \big\|_{L^2(P_X)}^2 &= E_{\tilde{X}} \Big[\mathrm{Cov}_{YX} \big[k_\mathcal{X}^B(X, \tilde{X}) - k_\mathcal{X}^{B_0}(X, \tilde{X}), \psi(Y) \big]^2 \Big] \\
&\leq E_{\tilde{X}} \Big[\mathrm{Var}_X \big[k_d(B^T X, B^T \tilde{X}) - k_d(B_0^T X, B_0^T \tilde{X}) \big] \mathrm{Var}_Y [\psi(Y)] \Big] \\
&\leq E_{\tilde{X}} E_X \big[\big(k_d(B^T X, B^T \tilde{X}) - k_d(B_0^T X, B_0^T \tilde{X}) \big)^2 \big] \| \psi \|_{L^2(P_Y)}^2,
\end{aligned}$$

from which the continuity of $B \mapsto S^B_{XY}$ is obtained by the continuity and boundedness of k_d. The continuity of $(S^B_{XX} + \varepsilon I)^{-1}$ is shown by $\|(S^B_{XX} + \varepsilon I)^{-1} - (S^{B_0}_{XX} + \varepsilon I)^{-1}\| = \|(S^B_{XX} + \varepsilon I)^{-1}(S^{B_0}_{XX} - S^B_{XX})(S^{B_0}_{XX} + \varepsilon I)^{-1}\| \leq \frac{1}{\varepsilon^2}\|S^{B_0}_{XX} - S^B_{XX}\|$. □

To establish the continuity of $L_0(B) = \mathrm{Tr}\big[\Sigma^B_{YX}{\Sigma^B_{XX}}^{-1}\Sigma^B_{XY}\big]$, the argument in the proof of Lemma 11 cannot be applied, because ${\Sigma^B_{XX}}^{-1}$ is not bounded in general. The assumptions (A-1) and (A-2) are used for the proof.

Lemma 12. *Suppose k_d is continuous and bounded. Under the assumptions* (A-1) *and* (A-2)*, the function $L_0(B)$ is continuous on $\mathbb{S}^m_d(\mathbb{R})$.*

Proof. By the same argument as in the proof of Lemma 10, it suffices to establish the continuity of $B \mapsto \langle \psi, \Sigma^B_{YY|X}\psi\rangle$ for $\psi \in \mathcal{H}_\mathcal{Y}$. From Proposition 2, the proof is completed if the continuity of the map

$$B \mapsto \inf_{f \in \mathcal{H}^B_\mathcal{X}} \mathrm{Var}_{XY}[g(Y) - f(X)]$$

is proved for any continuous and bounded function g.

Since $f(x)$ depends only on $B^T x$ for any $f \in \mathcal{H}^B_\mathcal{X}$, under the assumption (A-2), we use the same argument as in the proof of Proposition 3 to obtain

$$\begin{aligned}
&\inf_{f \in \mathcal{H}^B_\mathcal{X}} \mathrm{Var}_{XY}[g(Y) - f(X)] \\
&= \inf_{f \in \mathcal{H}^B_\mathcal{X}} \mathrm{Var}_X\big[E_{Y|BB^TX}[g(Y)|BB^TX] - f(X)\big] + E_X\big[\mathrm{Var}_{Y|BB^TX}[g(Y)|BB^TX]\big] \\
&= E_Y[g(Y)^2] - E_X\big[E_{Y|B^TX}[g(Y)|B^TX]^2\big],
\end{aligned}$$

which is a continuous function of $B \in \mathbb{S}^m_d(\mathbb{R})$ from Assumption (A-1). □

Lemma 13. *Suppose that k_d is continuous and bounded, and that ε_n converges to zero as n goes to infinity. Under the assumptions* (A-1) *and* (A-2)*, we have*

$$\sup_{B \in \mathbb{S}^m_d(\mathbb{R})} \mathrm{Tr}\big[\Sigma^B_{YY|X} - \big\{\Sigma_{YY} - \Sigma^B_{YX}(\Sigma^B_{XX} + \varepsilon_n I)^{-1}\Sigma^B_{XY}\big\}\big] \;\to\; 0 \quad (n \to \infty).$$

Proof. From Lemmas 10, 11 and 12, the continuous function $\mathrm{Tr}[\Sigma_{YY} - \Sigma_{YX}\big(\Sigma^B_{XX} + \varepsilon_n I\big)^{-1}\Sigma^B_{XY}]$ converges to the continuous function $\mathrm{Tr}[\Sigma^B_{YY|X}]$ for every $B \in \mathbb{S}^m_d(\mathbb{R})$. Because this convergence is monotone and $\mathbb{S}^m_d(\mathbb{R})$ is compact, it is necessarily uniform. □

The proof of Proposition 6 is now easily obtained.

Proof of Proposition 6. Lemmas 11 and 12 show the continuity of $\mathrm{Tr}\left[\widehat{\Sigma}^{B(n)}_{YY|X}\right]$ and $\mathrm{Tr}\left[\Sigma^{B}_{YY|X}\right]$. Lemmas 9 and 13 prove the uniform convergence. □

5 Conclusions

This paper has presented KDR, a new method for sufficient dimension reduction in regression. The method is based on a characterization of conditional independence using covariance operators on reproducing Hilbert spaces. This characterization is not restricted to first- or second-order conditional moments, but exploits high-order moments in the estimation of the central subspace. The KDR method is widely applicable; in distinction to most of the existing literature on SDR it does not impose strong assumptions on the probability distribution of the covariate vector X. It is also applicable to problems in which the response Y is discrete.

We have developed some asymptotic theory for the estimator, resulting in a proof of consistency of the estimator under weak conditions. The proof of consistency reposes on a result establishing the uniform convergence of the empirical process on a Hilbert space. In particular, we have established the rate $O_p(n^{-1/2})$ for uniform convergence, paralleling the results for ordinary real-valued empirical processes.

We have not yet developed distribution theory for the KDR method, and have left open the important problem of inferring the dimensionality of the central subspace. Our proof techniques do not straightforwardly extend to yield the asymptotic distribution of the KDR estimator, and new techniques may be required.

It should be noted, however, that inference of the dimensionality of the central subspace is not necessary for many of the applications of SDR. In particular, SDR is often used in the context of graphical exploration of data, where a data analyst may wish to explore views of varying dimensionality. Also, in high-dimensional prediction problems of the kind studied in statistical machine learning, dimension reduction may be carried out in the context of predictive modeling, in which case cross-validation and related techniques may be used to choose the dimensionality.

Finally, while we have focused our discussion on the central subspace as the object of inference, it is also worth noting that KDR applies even to situations in which a central subspace does not exist. As we have shown, the KDR estimate converges to the subset of projection matrices that sat-

isfy Eq. (1); this result holds regardless of the existence of a central subspace. That is, if the intersection of dimension-reduction subspaces is not a dimension-reduction subspace, but if the dimensionality chosen for KDR is chosen to be large enough such that subspaces satisfying Eq. (1) exist, then KDR will converge to one of those subspaces.

Acknowledgements

The authors thank Dr. Yoichi Nishiyama for his helpful comments on the uniform convergence of empirical processes. We would like to acknowledge support from JSPS KAKENHI 15700241, a research grant from the Inamori Foundation, a Scientific Grant from the Mitsubishi Foundation, and Grant 0412995 from the National Science Foundation.

A Uniform convergence of cross-covariance operators

In this appendix we present a proof of Lemma 8. The proof involves the use of random elements in a Hilbert space (Vakhania et al., 1987; Baker, 1973). Let $\mathcal{H}$ be a Hilbert space equipped with a Borel σ-field. A *random element* in the Hilbert space $\mathcal{H}$ is a measurable map $F : \Omega \to \mathcal{H}$ from a measurable space $(\Omega, \mathfrak{S})$. If $\mathcal{H}$ is an RKHS on a measurable set $\mathcal{X}$ with a measurable positive definite kernel k, a random variable X in $\mathcal{X}$ defines a random element in $\mathcal{H}$ by $k(\cdot, X)$.

A random element F in a Hilbert space $\mathcal{H}$ is said to have *strong order* p $(0 < p < \infty)$ if $E\|F\|^p$ is finite. For a random element F of strong order one, the expectation of F, which is defined as the element $m_F \in \mathcal{H}$ such that $\langle m_F, g\rangle_{\mathcal{H}} = E[\langle F, g\rangle_{\mathcal{H}}]$ for all $g \in \mathcal{H}$, is denoted by $E[F]$. With this notation, the the interchange of the expectation and the the inner product is justified: $\langle E[F], g\rangle_{\mathcal{H}} = E[\langle F, g\rangle_{\mathcal{H}}]$. Note also that for independent random elements F and G of strong order two, the relation

$$E[\langle F, G\rangle_{\mathcal{H}}] = \langle E[F], E[G]\rangle_{\mathcal{H}}$$

holds.

Let (X, Y) be a random vector on $\mathcal{X} \times \mathcal{Y}$ with law P_{XY}, and let $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$ be the RKHS with positive definite kernels $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$, respectively, which satisfy Eq. (2). The random element $k_{\mathcal{X}}(\cdot, X)$ has strong order two, and $E[k(\cdot, X)]$ equals m_X, where m_X is given by Eq. (4). The random

element $k_{\mathcal{X}}(\cdot, X)k_{\mathcal{Y}}(\cdot, Y)$ in the direct product $\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}$ has strong order one. Define the zero mean random elements $F = k_{\mathcal{X}}(\cdot, X) - E[k_{\mathcal{X}}(\cdot, X)]$ and $G = k_{\mathcal{Y}}(\cdot, Y) - E[k_{\mathcal{Y}}(\cdot, Y)]$.

For an i.i.d. sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ on $\mathcal{X} \times \mathcal{Y}$ with law P_{XY}, define random elements $F_i = k_{\mathcal{X}}(\cdot, X_i) - E[k_{\mathcal{X}}(\cdot, X)]$ and $G_i = k_{\mathcal{Y}}(\cdot, Y_i) - E[k_{\mathcal{Y}}(\cdot, Y)]$. Then, $F, F_1, \ldots, F_n$ and $G, G_1, \ldots, G_n$ are zero mean i.i.d. random elements in $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$, respectively. In the following, the notation $\mathcal{F} = \mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}$ is used for simplicity.

As shown in the proof of Lemma 4 in Fukumizu et al. (2005), we have

$$\big\|\widehat{\Sigma}_{YX}^{(n)} - \Sigma_{YX}\big\|_{HS} = \Big\|\frac{1}{n}\sum_{i=1}^{n}\Big(F_i - \frac{1}{n}\sum_{j=1}^{n} F_j\Big)\Big(G_i - \frac{1}{n}\sum_{j=1}^{n} G_j\Big) - E[FG]\Big\|_{\mathcal{F}},$$

which provides a bound

$$\begin{aligned}\sup_{B\in\mathbb{S}_d^m(\mathbb{R})}\big\|\widehat{\Sigma}_{YX}^{B(n)} - \Sigma_{YX}^{B}\big\|_{HS} \le{}& \sup_{B\in\mathbb{S}_d^m(\mathbb{R})}\Big\|\frac{1}{n}\sum_{i=1}^{n}\big(F_i^B G_i - E[FG]\big)\Big\|_{\mathcal{F}^B} \\ &+ \sup_{B\in\mathbb{S}_d^m(\mathbb{R})}\Big\|\frac{1}{n}\sum_{j=1}^{n} F_j^B\Big\|_{\mathcal{H}_{\mathcal{X}}^B}\Big\|\frac{1}{n}\sum_{j=1}^{n} G_j\Big\|_{\mathcal{H}_{\mathcal{Y}}}, \quad (20)\end{aligned}$$

where F_i^B are defined with the kernel k^B, and $\mathcal{F}^B = \mathcal{H}_{\mathcal{X}}^B \otimes \mathcal{H}_{\mathcal{Y}}$. Also, Eq. (19) implies

$$\begin{aligned}\mathrm{Tr}\big[\widehat{\Sigma}_{XX}^{(n)} - \Sigma_{XX}\big] &= \frac{1}{n}\sum_{i=1}^{n}\Big\|F_i - \frac{1}{n}\sum_{j=1}^{n} F_j\Big\|_{\mathcal{H}_{\mathcal{X}}}^2 - E\|F\|_{\mathcal{H}_{\mathcal{X}}}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\|F_i\|_{\mathcal{H}_{\mathcal{X}}}^2 - E\|F\|_{\mathcal{H}_{\mathcal{X}}}^2 - \Big\|\frac{1}{n}\sum_{i=1}^{n} F_i\Big\|_{\mathcal{H}_{\mathcal{X}}}^2,\end{aligned}$$

from which we have

$$\begin{aligned}\sup_{B\in\mathbb{S}_d^m(\mathbb{R})}\big|\mathrm{Tr}\big[\widehat{\Sigma}_{XX}^{B(n)} - \Sigma_{XX}^{B}\big]\big| \le{}& \sup_{B\in\mathbb{S}_d^m(\mathbb{R})}\Big|\frac{1}{n}\sum_{i=1}^{n}\|F_i^B\|_{\mathcal{H}_{\mathcal{X}}^B}^2 - E\|F^B\|_{\mathcal{H}_{\mathcal{X}}^B}^2\Big| \\ &+ \sup_{B\in\mathbb{S}_d^m(\mathbb{R})}\Big\|\frac{1}{n}\sum_{i=1}^{n} F_i^B\Big\|_{\mathcal{H}_{\mathcal{X}}^B}^2. \quad (21)\end{aligned}$$

It follows that Lemma 8 is proved if all the four terms on the right hand side of Eqs. (20) and (21) are of order $O_p(1/\sqrt{n})$.

Hereafter, the kernel k_d is assumed to be bounded. We begin by considering the first term on the right hand side of Eq. (21). This is the supremum

of a process which consists of real-valued random variables $\|F_i^B\|^2_{\mathcal{H}_\mathcal{X}^B}$. Let U^B be a random element in $\mathcal{H}_d$ defined by

$$U^B = k_d(\cdot, B^T X) - E[k_d(\cdot, B^T X)],$$

and let $C > 0$ be a constant such that $|k_d(z,z)| \leq C^2$ for all $z \in \mathcal{Z}$. From $\|U^B\|_{\mathcal{H}_d} \leq 2C$, we have for $B, \tilde{B} \in \mathbb{S}_d^m(\mathbb{R})$

$$\begin{aligned}\left|\|F^B\|^2_{\mathcal{H}_\mathcal{X}^B} - \|F^{\tilde{B}}\|^2_{\mathcal{H}_\mathcal{X}^{\tilde{B}}}\right| &= \left|\langle U^B - U^{\tilde{B}}, U^B + U^{\tilde{B}}\rangle_{\mathcal{H}_d}\right| \\ &\leq \|U^B - U^{\tilde{B}}\|_{\mathcal{H}_d}\|U^B + U^{\tilde{B}}\|_{\mathcal{H}_d} \\ &\leq 4C\|U^B - U^{\tilde{B}}\|_{\mathcal{H}_d}.\end{aligned}$$

The above inequality, combined with the bound

$$\|U^B - U^{\tilde{B}}\|_{\mathcal{H}_d} \leq 2\phi(x)D(B, \tilde{B}) \tag{22}$$

obtained from Assumption (A-3), provides a Lipschitz condition $\left|\|F^B\|^2_{\mathcal{H}_\mathcal{X}^B} - \|F^{\tilde{B}}\|^2_{\mathcal{H}_\mathcal{X}^{\tilde{B}}}\right| \leq 8C\phi(x)D(B, \tilde{B})$, which works as a sufficient condition for the uniform central limit theorem (van der Vaart, 1998, Example 19.7). This yields

$$\sup_{B \in \mathbb{S}_d^m(\mathbb{R})} \left|\frac{1}{n}\sum_{i=1}^n \|F_i^B\|^2_{\mathcal{H}_\mathcal{X}^B} - E\|F^B\|^2_{\mathcal{H}_\mathcal{X}^B}\right| = O_p(1/\sqrt{n}).$$

Our approach to the other three terms is based on a treatment of empirical processes in a Hilbert space. For $B \in \mathbb{S}_d^m(\mathbb{R})$, let $U_i^B = k_d(\cdot, B^T X_i) - E[k_d(\cdot, B^T X)]$ be a random element in $\mathcal{H}_d$. Then the relation $\langle k^B(\cdot, x), k^B(\cdot, \tilde{x})\rangle_{\mathcal{H}_\mathcal{X}^B} = k_d(B^T x, B^T \tilde{x}) = \langle k_d(\cdot, B^T x), k_d(\cdot, B^T \tilde{x})\rangle_{\mathcal{H}_d}$ implies

$$\left\|\frac{1}{n}\sum_{j=1}^n F_j^B\right\|_{\mathcal{H}_\mathcal{X}^B} = \left\|\frac{1}{n}\sum_{j=1}^n U_j^B\right\|_{\mathcal{H}_d}, \tag{23}$$

$$\left\|\frac{1}{n}\sum_{j=1}^n F_j^B G - E[FG]\right\|_{\mathcal{H}_\mathcal{X}^B \otimes \mathcal{H}_\mathcal{Y}} = \left\|\frac{1}{n}\sum_{j=1}^n U_j^B G - E[U^B G]\right\|_{\mathcal{H}_d \otimes \mathcal{H}_\mathcal{Y}}. \tag{24}$$

Note also that the assumption (A-3) gives

$$\|U^B G - U^{\tilde{B}} G\|_{\mathcal{H}_d \otimes \mathcal{H}_\mathcal{Y}} \leq 2\sqrt{k_\mathcal{Y}(y,y)}\phi(x)D(B, \tilde{B}). \tag{25}$$

From Eqs. (22), (23), (24), and (25), the proof of Lemma 8 is completed from the following proposition:

Proposition 14. *Let $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ be a measurable space, let Θ be a compact metric space with distance D, and let $\mathcal{H}$ be a Hilbert space. Suppose that $X, X_1, \ldots, X_n$ are i.i.d. random variables on $\mathcal{X}$, and suppose $F : \mathcal{X} \times \Theta \to \mathcal{H}$ is a Borel measurable map. If $\sup_{\theta \in \Theta} \|F(x;\theta)\|_{\mathcal{H}} < \infty$ for all $x \in \mathcal{X}$ and there exists a measurable function $\phi : \mathcal{X} \to \mathbb{R}$ such that $E[\phi(X)^2] < \infty$ and*

$$\|F(x;\theta_1) - F(x;\theta_2)\|_{\mathcal{H}} \leq \phi(x) D(\theta_1, \theta_2) \qquad (\forall\ \theta_1, \theta_2 \in \Theta), \tag{26}$$

then we have

$$\sup_{\theta \in \Theta} \Big\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \big(F(X_i;\theta) - E[F(X;\theta)] \big) \Big\|_{\mathcal{H}} = O_p(1) \qquad (n \to \infty).$$

The proof of Proposition 14 is similar to that for a real-valued random process, and is divided into several lemmas.

I.i.d. random variables $\sigma_1, \ldots, \sigma_n$ taking values in $\{+1, -1\}$ with equal probability are called *Rademacher* variables. The following concentration inequality is known for a Rademacher average in a Banach space:

Proposition 15. *Let $a_1, \ldots, a_n$ be elements in a Banach space, and let $\sigma_1, \ldots, \sigma_n$ be Rademacher variables. Then, for every $t > 0$*

$$\Pr\big(\big\| \textstyle\sum_{i=1}^{n} \sigma_i a_i \big\| > t \big) \leq 2 \exp\Big(-\frac{t^2}{32 \sum_{i=1}^{n} \|a_i\|^2} \Big).$$

Proof. See Ledoux and Talagrand (1991, Theorem 4.7 and the remark thereafter). □

With Proposition 15, the following exponential inequality is obtained with a slight modification of the standard symmetrization argument for empirical processes.

Lemma 16. *Let $X, X_1, \ldots, X_n$ and $\mathcal{H}$ be as in Proposition 14, and denote $(X_1, \ldots, X_n)$ by $\mathbf{X}_n$. Let $F : \mathcal{X} \to \mathcal{H}$ be a Borel measurable map with $E\|F(X)\|_{\mathcal{H}}^2 < \infty$. For a positive number M such that $E\|F(X)\|_{\mathcal{H}}^2 < M$, define an event A_n by $\frac{1}{n} \sum_{i=1}^{n} \|F(X_i)\|^2 \leq M$. Then, for every $t > 0$ and sufficiently large n,*

$$\Pr\Big(\Big\{ \mathbf{X}_n \,\Big|\, \Big\| \frac{1}{n} \sum_{i=1}^{n} \big(F(X_i) - E[F(X)] \big) \Big\|_{\mathcal{H}} > t \Big\} \cap A_n \Big) \leq 8 \exp\Big(-\frac{nt^2}{1024 M} \Big).$$

Proof. First, note that for any sufficiently large n we have $\Pr(A_n) \geq \frac{3}{4}$ and $\Pr\big(\frac{1}{n}\sum_{i=1}^n (F(X_i) - E[F(X)]) \leq \frac{t}{2}\big) \geq \frac{3}{4}$. We consider only such n in the following. Let $\tilde{\mathbf{X}}_n$ be an independent copy of $\mathbf{X}_n$, and let $\tilde{A}_n = \big\{\tilde{\mathbf{X}}_n \,\big|\, \frac{1}{n}\sum_{i=1}^n \|F(\tilde{X}_i)\|^2 \leq M\big\}$. The obvious inequality

$$\begin{aligned}
&\Pr\Big(\Big\{\mathbf{X}_n \,\Big|\, \Big\|\frac{1}{n}\sum_{i=1}^n (F(X_i) - E[F(X)])\Big\|_{\mathcal{H}} > t\Big\} \cap A_n\Big) \\
&\quad \times \Pr\Big(\Big\{\tilde{\mathbf{X}}_n \,\Big|\, \Big\|\frac{1}{n}\sum_{i=1}^n (F(\tilde{X}_i) - E[F(X)])\Big\|_{\mathcal{H}} \leq \frac{t}{2}\Big\} \cap \tilde{A}_n\Big) \\
&\qquad \leq \Pr\Big(\Big\{(\mathbf{X}_n, \tilde{\mathbf{X}}_n) \,\Big|\, \Big\|\frac{1}{n}\sum_{i=1}^n (F(X_i) - F(\tilde{X}_i))\Big\|_{\mathcal{H}} > \frac{t}{2}\Big\} \cap A_n \cap \tilde{A}_n\Big)
\end{aligned}$$

and the fact that $B_n := \big\{(\mathbf{X}_n, \tilde{\mathbf{X}}_n) \,\big|\, \frac{1}{2n}\sum_{i=1}^n \big(\|F(X_i)\|^2 + \|F(\tilde{X}_i)\|^2\big) \leq M\big\}$ includes $A_n \cap \tilde{A}_n$ gives a symmetrized bound

$$\begin{aligned}
&\Pr\Big(\Big\{\mathbf{X}_n \,\Big|\, \Big\|\frac{1}{n}\sum_{i=1}^n (F(X_i) - E[F(X)])\Big\|_{\mathcal{H}} > t\Big\} \cap A_n\Big) \\
&\qquad \leq 2\Pr\Big(\Big\{(\mathbf{X}_n, \tilde{\mathbf{X}}_n) \,\Big|\, \Big\|\frac{1}{n}\sum_{i=1}^n (F(X_i) - F(\tilde{X}_i))\Big\|_{\mathcal{H}} > \frac{t}{2}\Big\} \cap B_n\Big).
\end{aligned}$$

With Rademacher variables $\sigma_1, \ldots, \sigma_n$, the right hand side is equal to

$$2\Pr\Big(\Big\{(\mathbf{X}_n, \tilde{\mathbf{X}}_n, \{\sigma_i\}) \,\Big|\, \Big\|\frac{1}{n}\sum_{i=1}^n \sigma_i (F(X_i) - F(\tilde{X}_i))\Big\|_{\mathcal{H}} > \frac{t}{2}\Big\} \cap B_n\Big),$$

which is upper-bounded by

$$\begin{aligned}
&4\Pr\Big(\Big\|\frac{1}{n}\sum_{i=1}^n \sigma_i F(X_i)\Big\|_{\mathcal{H}} > \frac{t}{4} \text{ and } \frac{1}{2n}\sum_{i=1}^n \|F(X_i)\|_{\mathcal{H}}^2 \leq M\Big) \\
&= 4E_{\mathbf{X}_n}\Big[\Pr\Big(\Big\|\frac{1}{n}\sum_{i=1}^n \sigma_i F(X_i)\Big\|_{\mathcal{H}} > \frac{t}{4} \;\Big|\; \mathbf{X}_n\Big) 1_{\{\mathbf{X}_n \in C_n\}}\Big],
\end{aligned}$$

where $C_n = \big\{\mathbf{X}_n \,\big|\, \frac{1}{n}\sum_{i=1}^n \|F(X_i)\|_{\mathcal{H}}^2 \leq 2M\big\}$. From Proposition 15, the last line is upper-bounded by $4\exp\big(-\frac{(nt/4)^2}{32\sum_{i=1}^n \|F(X_i)\|^2}\big) \leq 4\exp\big(-\frac{nt^2}{1024M}\big)$. □

Let Θ be a set with semimetric d. For any $\delta > 0$, the *covering number* $N(\delta, d, \Theta)$ is the smallest $m \in \mathbb{N}$ for which there exist m points $\theta_1, \ldots, \theta_m$

in Θ such that $\min_{1\le i\le m} d(\theta,\theta_i) \le \delta$ holds for any $\theta \in \Theta$. We write $N(\delta)$ for $N(\delta, d, \Theta)$ if there is no confusion. For $\delta > 0$, the *covering integral* $J(\delta)$ for Θ is defined by

$$J(\delta) = \int_0^\delta \bigl(8\log(N(u)^2/u)\bigr)^{1/2} du.$$

The chaining lemma (Pollard, 1984), which plays a crucial role in the uniform central limit theorem, is readily extendable to a random process in a Banach space.

Lemma 17 (Chaining Lemma). *Let Θ be a set with semimetric d, and let $\{Z(\theta) \mid \theta \in \Theta\}$ be a family of random elements in a Banach space. Suppose Θ has a finite covering integral $J(\delta)$ for $0 < \delta < 1$ and suppose there exists a positive constant $R > 0$ such that for all $\theta, \eta \in \Theta$ and $t > 0$ the inequality*

$$\Pr\bigl(\|Z(\theta) - Z(\eta)\| > t d(\theta,\eta)\bigr) \le 8\exp\bigl(-\tfrac{1}{2R}t^2\bigr)$$

holds. Then, there exists a countable subset Θ^* of Θ such that for any $0 < \varepsilon < 1$

$$\Pr\Bigl(\sup_{\theta,\eta\in\Theta^*, d(\theta,\eta)\le\varepsilon} \|Z(\theta) - Z(\eta)\| > 26RJ(d(\theta,\eta))\Bigr) \le 2\varepsilon$$

holds. If $Z(\theta)$ has continuous sample paths, then Θ^* can be replaced by Θ.

Proof. By noting that the proof of the chaining lemma for a real-valued random process does not use any special properties of real numbers but the property of the norm (absolute value) for $Z(\theta)$, the proof applies directly to a process in a Banach space. See Pollard (1984, Section VII.2). □

Proof of Proposition 14. Note that Eq. (26) means

$$\Bigl\|\frac{1}{n}\sum_{i=1}^n (F(X_i;\theta_1) - F(X_i;\theta_2))\Bigr\|_{\mathcal{H}}^2 \le D(\theta_1,\theta_2)^2 \frac{1}{n}\sum_{i=1}^n \phi(X_i)^2.$$

Let $M > 0$ be a constant such that $E[\phi(X)^2] < M$, and let $A_n = \bigl\{\mathbf{X}_n \mid \bigl\|\frac{1}{n}\sum_{i=1}^n (F(X_i;\theta_1) - F(X_i;\theta_2))\bigr\|_{\mathcal{H}}^2 \le MD(\theta_1,\theta_2)^2\bigr\}$. Since the probability of A_n converges to zero as $n \to \infty$, it suffices to show that there exists $\delta > 0$ such that the probability

$$\mathbb{P}_n = \Pr\Bigl(\mathbf{X}_n \Bigm| A_n \cap \Bigl\{\sup_{\theta\in\Theta}\Bigl\|\frac{1}{\sqrt{n}}\sum_{i=1}^n \bigl(F(X_i;\theta) - E[F(X;\theta)]\bigr)\Bigr\|_{\mathcal{H}} > \delta\Bigr\}\Bigr)$$

satisfies $\limsup_{n\to\infty} \mathbb{P}_n = 0$.

With the notation $\tilde{F}_\theta(x) = F(x;\theta) - E[F(X;\theta)]$, from Lemma 16 we can derive

$$\begin{aligned}\Pr\Big(A_n \cap \Big\{\mathbf{X}_n \,\Big|\, \Big\|\frac{1}{\sqrt{n}}\sum_{i=1}^n \big(\tilde{F}_{\theta_1}(X_i) - \tilde{F}_{\theta_2}(X_i)\big)\Big\|_{\mathcal{H}} > t\Big\}\Big) \\ \le 8\exp\Big(-\frac{t^2}{512\cdot 2MD(\theta_1,\theta_2)^2}\Big),\end{aligned}$$

for any $t > 0$ and sufficiently large n. Because the covering integral $J(\delta)$ with respect to D is finite by the compactness of Θ, and the sample path $\Theta \ni \theta \mapsto \frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{F}_\theta(X_i) \in \mathcal{H}$ is continuous, the chaining lemma implies that for any $0 < \varepsilon < 1$

$$\begin{aligned}\Pr\Bigg(A_n \cap \Big\{\mathbf{X}_n \,\Big|\, \sup_{\theta_1,\theta_2\in\Theta, D(\theta_1,\theta_2)\le\varepsilon} \Big\|\frac{1}{\sqrt{n}}\sum_{i=1}^n \big(\tilde{F}_{\theta_1}(X_i) - \tilde{F}_{\theta_2}(X_i)\big)\Big\|_{\mathcal{H}} \\ > 26\cdot 512M\cdot J(\varepsilon)\Big\}\Bigg) \le 2\varepsilon.\end{aligned}$$

Take an arbitrary $\varepsilon \in (0,1)$. We can find a finite number of partitions $\Theta = \cup_{a=1}^{\nu(\varepsilon)}\Theta_a$ ($\nu(\varepsilon) \in \mathbb{N}$) so that any two points in each Θ_a are within the distance ε. Let θ_a be an arbitrary point in Θ_a. Then the probability $\mathbb{P}_n$ is bounded by

$$\begin{aligned}\mathbb{P}_n \le{}& \Pr\Big(\max_{1\le a\le\nu(\varepsilon)} \Big\|\frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{F}_{\theta_a}(X_i)\Big\|_{\mathcal{H}} > \frac{\delta}{2}\Big) \\ &+ \Pr\Bigg(A_n \cap \Big\{\mathbf{X}_n \,\Big|\, \sup_{\theta,\eta\in\Theta, D(\theta,\eta)\le\varepsilon} \Big\|\frac{1}{\sqrt{n}}\sum_{i=1}^n \big(\tilde{F}_{\theta}(X_i) - \tilde{F}_{\eta}(X_i)\big)\Big\|_{\mathcal{H}} > \frac{\delta}{2}\Big\}\Bigg).\end{aligned} \tag{27}$$

From Chebyshev's inequality the first term is upper-bounded by

$$\nu(\varepsilon)\Pr\Big(\Big\|\frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{F}_{\theta_a}(X_i)\Big\|_{\mathcal{H}} > \frac{\delta}{2}\Big) \le \frac{4\nu(\varepsilon)E\|\tilde{F}_{\theta_a}(X)\|_{\mathcal{H}}^2}{\delta^2}.$$

If we take sufficiently large δ so that $512MJ(\varepsilon) < \delta/2$ and $\frac{4\nu(\varepsilon)E\|\tilde{F}_{\theta_a}(X)\|_{\mathcal{H}}^2}{\varepsilon} < \delta^2$, the right hand side of Eq. (27) is bounded by 3ε, which competes the proof. □

References

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 69(3):337–404, 1950.

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

C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups*. Springer-Verlag, New York, 1984.

F. Chiaromonte and R. D. Cook. Sufficient dimension reduction and graphics in regression. *Annals of the Institute of Statistical Mathematics*, 54(4): 768–795, 2002.

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 B. Li. Dimension reduction for conditional mean in regression. *The Annals of Statistics*, 30(2):455–474, 2002.

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

R. D. Cook and X. Yin. Dimension reduction and visualization in discriminant analysis (with discussion). *Australian & New Zealand Journal of Statistics*, 43(2):147–199, 2001.

B. Flury and H. Riedwyl. *Multivariate Statistics: A Practical Approach*. Chapman and Hall, 1988.

K. Fukumizu, F. R. Bach, and A. Gretton. Consistency of kernel canonical correlation analysis. Research Memorandum 942, Institute of Statistical Mathematics, 2005.

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.

A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. Technical Report 140, Max-Planck-Institut für biologische Kybernetik, 2005.

C. W. Groetsch. *The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind*. Pitman, 1984.

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

S. Kobayashi and K. Nomizu. *Foundations of Differential Geometry, Vol. 1*. John Wiley & Sons, 1963.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces*. Springer-Verlag, 1991.

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

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.

D. Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, 1984.

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

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2001.

N. N. Vakhania, V. I. Tarieladze, and S. A. Chobanyan. *Probability Distributions on Banach Spaces*. D. Reidel Publishing Company, 1987.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

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

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