Dimensionality Reduction for Supervised Learning with Reproducing Kernel Hilbert Spaces

Kenji Fukumizu fukumizu@ism.ac.jp
Institute of Statistical Mathematics
Tokyo 106-8569, Japan

Francis R. Bach fbach@cs.berkeley.edu
Computer Science Division
University of California
Berkeley, CA 94720, USA

Michael I. Jordan jordan@cs.berkeley.edu
Computer Science Division and Department of Statistics
University of California
Berkeley, CA 94720, USA

Technical Report 641
May 25, 2003

Abstract

We propose a novel method of dimensionality reduction for supervised learning problems. Given a regression or classification problem in which we wish to predict a response variable Y from an explanatory variable X, we treat the problem of dimensionality reduction as that of finding a low-dimensional "effective subspace" of X which retains the statistical relationship between X and Y. We show that this problem can be formulated in terms of conditional independence. To turn this formulation into an optimization problem we establish a general nonparametric characterization of conditional independence using covariance operators on a reproducing kernel Hilbert space. This characterization allows us to derive a contrast function for estimation of the effective subspace. Unlike many conventional methods for dimensionality reduction in supervised learning, the proposed method requires neither assumptions on the marginal distribution of X, nor a parametric model of the conditional distribution of Y. We present experiments that compare the performance of the method with conventional methods.

1. Introduction

Many statistical learning problems involve some form of dimensionality reduction, either explicitly or implicitly. The goal may be one of *feature selection*, in which we aim to find linear or nonlinear combinations of the original set of variables, or one of *variable selection*, in which we wish to select a subset of variables from the original set. The setting may be unsupervised learning, in which a set of observations of a random vector X are available, or supervised learning, in which desired responses or labels Y are also available. Developing methods for dimensionality reduction requires being clear on the goal and the setting, as methods developed for one combination of goal and setting are not generally appropriate for another. There are additional motivations for dimensionality reduction that it is also helpful to specify, including: providing a simplified explanation of a phenomenon for a human (possibly as part of a visualization algorithm), suppressing noise so as to make a better prediction or decision, or reducing the computational burden. These various motivations are often complementary.

In this paper we study dimensionality reduction in the setting of supervised learning. Thus, we consider problems in which our data consist of observations of (X, Y) pairs, where X is an m-dimensional explanatory variable and where Y is an ℓ-dimensional response. The variable Y may be either continuous or discrete. We refer to these problems generically as "regression" problems, which indicates our focus on the conditional probability density function $p_{Y|X}(y \mid x)$. In particular, our framework includes discriminative approaches to classification problems, where Y is a discrete label.

We wish to solve a problem of feature selection in which the features are linear combinations of the components of X. In particular, we assume that there is an r-dimensional subspace $S \subset \mathbb{R}^m$ such that

$$p_{Y|X}(y \mid x) = p_{Y|\Pi_S X}(y \mid \Pi_S x), \tag{1}$$

for all x and y, where Π_S is the orthogonal projection of $\mathbb{R}^m$ onto S. The subspace S is called the *effective subspace for regression*. Based on a set of observations of (X, Y) pairs, we wish to recover a matrix whose columns span the effective subspace.

We approach the problem as a *semiparametric* statistical problem; in particular, we make no assumptions regarding the conditional distribution $p_{Y|\Pi_S X}(y \mid \Pi_S x)$, nor do we make any assumptions regarding the marginal distribution $p_X(x)$. That is, we wish to estimate a finite-dimensional parameter (a matrix whose columns span the effective subspace), while treating the distributions $p_{Y|\Pi_S X}(y \mid \Pi_S x)$ and $p_X(x)$ nonparametrically.

Having found an effective subspace, we may then proceed to build a parametric or nonparametric regression model on that subspace. Thus our approach is an explicit dimensionality reduction method for supervised learning that does not require any particular form of regression model, and can be used as a preprocessor for any supervised learner. This can be compared to the use of methods such as principal components analysis (PCA) in regression, which also make no assumption regarding the subsequent regression model, but fail to make use of the response variable Y.

There are a variety of related approaches in the literature, but most of them involve making specific assumptions regarding the conditional distribution $p_{Y|\Pi_S X}(y \mid \Pi_S x)$, the marginal distribution $p_X(x)$, or both. For example, classical two-layer neural networks involve a linear transformation in the first "layer," followed by a specific nonlinear function and a second layer (Bishop, 1995). Thus, neural networks can be seen as attempting to estimate an effective subspace based on specific assumptions about the regressor $p_{Y|\Pi_S X}(y \mid \Pi_S x)$. Similar comments apply to projection pursuit regression (Friedman and Stuetzle, 1981), ACE (Breiman and Friedman, 1985) and additive models (Hastie and Tibshirani, 1986), all of which provide a methodology for dimensionality reduction in which an additive model $E[Y \mid X] = g_1(\beta_1^T X) + \cdots + g_K(\beta_K^T X)$ is assumed for the regressor.

Canonical correlation analysis (CCA) and partial least squares (PLS, Höskuldsson, 1988, Helland, 1988) are classical multivariate statistical methods that can be used for dimensionality reduction in regression (Fung et al., 2002, Nguyen and Rocke, 2002). These methods are based on a linearity assumption for the regressor, however, and thus are quite strongly parametric.

The line of research that is closest to our work has its origin in a technique known as sliced inverse regression (SIR, Li, 1991). SIR is a semiparametric method for finding effective subspaces in regression. The basic idea is that the range of the response variable Y is partitioned into a set of "slices," and the sample means of the observations X are computed within each slice. This can be viewed as a rough approximation to the inverse regression of X on Y. For univariate Y the method is particularly easy to implement. Noting that the inverse regression must lie in the effective subspace if the forward regression lies in such a subspace, principal component analysis is then used on the sample means to find the effective subspace. Li (1991) has shown that this approach can find effective subspaces, but only under strong assumptions on the marginal distribution $p_X(x)$—in particular, the marginal distribution must be elliptically symmetric.

Further developments in the wake of SIR include principal Hessian directions (pHd, Li, 1992), and sliced average variance estimation (SAVE, Cook and Weisberg, 1991, Cook and Yin, 2001). These are all semiparametric methods in that they make no assumptions about the regressor (see also Cook, 1998). However, they again place strong restrictions on the probability distribution of the explanatory variables. If these assumptions do not hold, there is no guarantee of finding the effective subspace.

There are also related nonparametric approaches that estimate the derivative of the regressor to achieve dimensionality reduction, based on the fact that the derivative of the conditional expectation $E[y \mid B^T x]$ with respect to x belongs to the effective subspace (Samarov, 1993, Hristache et al., 2001). However, nonparametric estimation of derivatives is quite challenging in high-dimensional spaces.

There are also dimensionality reduction methods with a semiparametric flavor in the area of classification, notably the work of Torkkola (2003), who has proposed using nonparametric estimation of the mutual information between X and Y, and subsequent maximization of this estimate of mutual information with respect to a matrix representing the effective subspace.

In this paper we present a novel semiparametric method for dimensionality reduction that we refer to as *Kernel Dimensionality Reduction (KDR)*. KDR is based on the estimation and optimization of a particular class of operators on reproducing kernel Hilbert spaces (Aronszajn, 1950). Although our use of reproducing kernel Hilbert spaces is related to their role in algorithms such as the support vector machine and kernel PCA (Boser et al., 1992, Vapnik et al., 1997, Schölkopf et al., 1998), where the kernel function allows linear operations in function spaces to be performed in a computationally-efficient manner, our work differs in that it cannot be viewed as a "kernelization" of an underlying linear algorithm. Rather, we use reproducing kernel Hilbert spaces to provide characterizations of general notions of independence, and we use these characterizations to design objective functions to be optimized. We build on earlier work by Bach and Jordan (2002a), who showed how to use reproducing kernel Hilbert spaces to characterize *marginal independence* between pairs of variables, and thereby design an objective function for independent component analysis. In the current paper, we extend this line of work, showing how to characterize *conditional independence* using reproducing kernel Hilbert spaces. We achieve this by expressing conditional independence in terms of covariance operators on reproducing kernel Hilbert spaces.

How does conditional independence relate to our dimensionality reduction problem? Recall that our problem is to find a projection Π_S of X onto a subspace S such that the conditional probability of Y given X is equal to the conditional probability of Y given $\Pi_S X$. This is equivalent to finding a projection Π_S which makes Y and $(I - \Pi_S)X$ conditionally independent given $\Pi_S X$. Thus we can turn the dimensionality reduction problem into an optimization problem by expressing it in terms of covariance operators.

In a presence of a finite sample, we need to estimate the covariance operator so as to obtain a sampled-based objective function that we can optimize. We derive a natural plug-in estimate of the covariance operator, and find that the resulting estimate is identical to the *kernel generalized variance* that has been described earlier by Bach and Jordan (2002a) in the setting of independent component analysis. In that setting, the goal is to measure departures from independence, and the minimization of the kernel generalized variance can be viewed as a surrogate for minimizing a certain mutual information. In the dimensionality reduction setting, on the other hand, the goal is to measure *conditional* independence, and minimizing the kernel generalized variance can be viewed as a surrogate for *maximizing* a certain mutual information. Not surprisingly, the derivation that leads to the kernel generalized variance that we present here is quite different from the one presented in the earlier work on kernel ICA. Moreover, the argument that we present here can be viewed as providing a rigorous foundation for other, more heuristic, ways in which the kernel generalized variance has been used, including the model selection algorithms for graphical models presented by Bach and Jordan (2003).

The paper is organized as follows. In Section 2, we introduce the problem of dimensionality reduction for supervised learning, and describe its relation with conditional independence and mutual information. Section 3 derives the objective function for estimation of the effective subspace for regression, and describes the KDR method. All of the mathematical details needed for the results in Section 3 are presented in the Appendix, which also provides a general introduction to covariance operators in reproducing kernel Hilbert spaces. In Section 4, we present a series of experiments that test the effectiveness of our method, comparing it with several conventional methods. Section 5 describes an extension of KDR to the problem of variable selection. Section 6 presents our conclusions.

2. Dimensionality reduction for regression

We consider a regression problem, in which Y is an ℓ-dimensional random vector, and X is an m-dimensional explanatory variable. (Note again that we use "regression" in a generic sense that includes both continuous and discrete Y). The probability density function of Y given X is denoted by $p_{Y|X}(y \mid x)$. Assume that there is an r-dimensional subspace $S \subset \mathbb{R}^m$ such that

$$p_{Y|X}(y \mid x) = p_{Y|\Pi_S X}(y \mid \Pi_S x), \tag{2}$$

for all x and y, where Π_S is the orthogonal projection of $\mathbb{R}^m$ onto S. The subspace S is called the *effective subspace for regression*.

The problem that we treat here is that of finding the subspace S given an *i.i.d.* sample $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ from p_X and $p_{Y|X}$. The crux of the problem is that we assume no *a* *priori* knowledge of the regressor, and place no assumptions on the conditional probability $p_{Y|X}$.

As in the simpler setting of principal component analysis, we make the (generally unrealistic) assumption that the dimensionality r is known and fixed. We discuss various approaches to the estimation of the dimensionality in Section 6.

The notion of effective subspace can be formulated in terms of conditional independence. Let (B, C) be the m-dimensional orthogonal matrix such that the column vectors of B span the subspace S, and define $U = B^T X$ and $V = C^T X$. Because (B, C) is an orthogonal matrix, we have

$$p_X(x) = p_{U,V}(u, v), \qquad p_{X,Y}(x, y) = p_{U,V,Y}(u, v, y), \tag{3}$$

for the probability density functions. From Eq. (3), Eq. (2) is equivalent to

$$p_{Y|U,V}(y \mid u, v) = p_{Y|U}(y \mid u). \tag{4}$$

This shows that the effective subspace S is the one which makes Y and V conditionally independent given U (see Figure 1).

Mutual information provides another point of view on the equivalence between conditional independence and the existence of the effective subspace. From Eq. (3), it is straightforward to see that

$$I(Y, X) = I(Y, U) + E_U\big[I(Y|U, V|U)\big], \tag{5}$$

where $I(Z, W)$ denotes the mutual information defined by

$$I(Z, W) := \int\int p_{Z,W}(z, w) \log \frac{p_{Z,W}(z, w)}{p_Z(z)p_W(w)} dzdw. \tag{6}$$

Because Eq. (2) means $I(Y, X) = I(Y, U)$, the effective subspace S is characterized as the subspace which retains the mutual information of X and Y by the projection onto that subspace, or equivalently, which gives $I(Y|U, V|U) = 0$. This is again the conditional independence of Y and V given U.

The expression in Eq. (5) can be understood in terms of the decomposition of the mutual information according to a tree-structured graphical model—a quantity that has been termed the *T-mutual information* by Bach and Jordan (2002b). Considering the tree $Y - U - V$ in Figure 1(b), we have that the T-mutual information I^T is given by

$$I^T = I(Y, U, V) - I(Y, U) - I(U, V). \tag{7}$$

This is equal to the KL-divergence between a probability distribution on (Y, U, V) and its projection onto the family of distributions that factor according to the tree; that is, the set of distributions that verify $Y \perp\!\!\!\perp V \mid U$. Using Eq. (3), we can easily see that $I(Y, U, V) = I(Y, X) + I(U, V)$, and thus we obtain

$$I^T = I(Y, X) - I(Y, U) = E_U[I(Y|U, V|U)]. \tag{8}$$

Then, dimensionality reduction for regression can be viewed as the problem of minimizing the T-mutual information for the fixed tree structure in Figure 1(b).

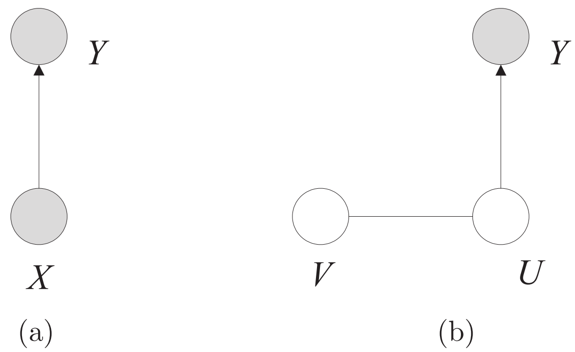


Figure 1: Graphical representation of dimensionality reduction for regression. The variables Y and V are conditionally independent given U, where $X = (U, V)$.

3. Kernel method for dimensionality reduction in regression

In this section we present our kernel-based method for dimensionality reduction. We discuss the basic definition and properties of cross-covariance operators on reproducing kernel Hilbert spaces, derive an objective function for characterizing conditional independence using cross-covariance operators, and finally present a sampled-based objective function based on this characterization.

3.1 Cross-covariance operators on reproducing kernel Hilbert spaces

We use cross-covariance operators on reproducing kernel Hilbert spaces to derive an objective function for dimensionality reduction. While cross-covariance operators are generally defined for random variables in Banach spaces (Vakhania et al., 1987, Baker, 1973), the theory is much simpler for reproducing kernel Hilbert spaces. We summarize only basic mathematical facts in this subsection, and defer the details to the Appendix. Let $(\mathcal{H}, k)$ be a reproducing kernel Hilbert space of functions on a set Ω with a positive definite kernel $k : \Omega \times \Omega \to \mathbb{R}$. The inner product of $\mathcal{H}$ is denoted by $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. We consider only real Hilbert spaces for simplicity. The most important aspect of reproducing kernel Hilbert spaces is the reproducing property:

$$\langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x) \quad \text{for all } x \in \Omega \text{ and } f \in \mathcal{H}. \tag{9}$$

Throughout this paper we use the Gaussian kernel

$$k(x_1, x_2) = \exp\bigl(-\|x_1 - x_2\|^2/\sigma^2\bigr), \tag{10}$$

which corresponds to a Hilbert space of smooth functions.

Let $(\mathcal{H}_1, k_1)$ and $(\mathcal{H}_2, k_2)$ be reproducing kernel Hilbert spaces over measurable spaces $(\Omega_1, \mathcal{B}_1)$ and $(\Omega_2, \mathcal{B}_2)$, respectively, with k_1 and k_2 measurable. For a random vector (X, Y) on $\Omega_1 \times \Omega_2$, the cross-covariance operator from $\mathcal{H}_1$ to $\mathcal{H}_2$ is defined by the relation

$$\langle g, \Sigma_{YX} f \rangle_{\mathcal{H}_2} = E_{XY}[f(X)g(Y)] - E_X[f(X)]E_Y[g(Y)] \tag{11}$$

for all $f \in \mathcal{H}_1$ and $g \in \mathcal{H}_2$. Eq. (11) implies that the covariance of $f(X)$ and $g(Y)$ is given by the action of the linear operator Σ_{YX} and the inner product. (See the Appendix for a basic exposition of cross-covariance operators.)

Covariance operators provide a useful framework for discussing conditional probability and conditional independence. As we show in Corollary 3 of the Appendix, the following relation holds between the conditional expectation and the cross-covariance operator, given that Σ_{XX} is invertible:[1]

$$E_{Y|X}[g(Y) \mid X] = \Sigma_{XX}^{-1}\Sigma_{XY}g \quad \text{for all } g \in \mathcal{H}_2, \tag{12}$$

Eq. (12) can be understood by analogy to the conditional expectation of Gaussian random variables. If X and Y are Gaussian random variables, it is well known that the conditional expectation is given by

$$E_{Y|X}[a^T Y \mid X = x] = x^T \Sigma_{XX}^{-1}\Sigma_{XY}a, \tag{13}$$

for an arbitrary vector a, where Σ_{XX} and Σ_{XY} are the variance-covariance matrices in the ordinary sense.

3.2 Conditional covariance operators and conditional independence

We derive an objective function for characterizing conditional independence using cross-covariance operators. Suppose we have random variables X and Y on $\mathbb{R}^m$ and $\mathbb{R}^\ell$, respectively. The variable X is decomposed into $U \in \mathbb{R}^r$ and $V \in \mathbb{R}^{m-r}$ so that $X = (U, V)$. For the function spaces corresponding to Y, U and V, we consider the reproducing kernel Hilbert spaces $(\mathcal{H}_1, k_1)$, $(\mathcal{H}_2, k_2)$, and $(\mathcal{H}_3, k_3)$ on $\mathbb{R}^\ell$, $\mathbb{R}^r$, and $\mathbb{R}^{m-r}$, respectively, each endowed with Gaussian kernels. We define the conditional covariance operator $\Sigma_{YY|U}$ on $\mathcal{H}_1$ by

$$\Sigma_{YY|U} := \Sigma_{YY} - \Sigma_{YU}\Sigma_{UU}^{-1}\Sigma_{UY}, \tag{14}$$

where Σ_{YY}, Σ_{UU}, Σ_{YU} are the corresponding covariance operators. As shown by Proposition 5 in the Appendix, the operator $\Sigma_{YY|U}$ captures the conditional variance of a random variable in the following way

$$\langle g, \Sigma_{YY|U}g\rangle_{\mathcal{H}_1} = E_U\big[\mathrm{Var}_{Y|U}[g(Y) \mid U]\big], \tag{15}$$

where g is an arbitrary function in $\mathcal{H}_1$. As in the case of Eq. (13), we can make an analogy to Gaussian variables. In particular, Eqs. (14) and (15) can be viewed as the analogs of the following well-known equality for the conditional variance of Gaussian variables:

$$\mathrm{Var}[a^T Y \mid U] = a^T(\Sigma_{YY} - \Sigma_{YU}\Sigma_{UU}^{-1}\Sigma_{UY})a. \tag{16}$$

It is natural to use minimization of $\Sigma_{YY|U}$ as a basis of a method for finding the most informative direction U. This intuition is justified theoretically by Theorem 7 in Appendix. That theorem shows that

$$\Sigma_{YY|U} \geq \Sigma_{YY|X} \qquad \text{for any} \quad U, \tag{17}$$

1. Even if Σ_{XX} is not invertible, a similar fact holds. See Corollary 3.

and

$$\Sigma_{YY|U} - \Sigma_{YY|X} = O \qquad \Longleftrightarrow \qquad Y \perp\!\!\!\perp V \mid U, \tag{18}$$

where, in Eq. (17), the inequality should be understood as the partial order of self-adjoint operators. From these relations, the effective subspace S can be characterized in terms of the solution to the following minimization problem:

$$\min_S \Sigma_{YY|U}, \quad \text{subject to } U = \Pi_S X. \tag{19}$$

In the following section we show how to turn this population-based criterion into a sampled-based criterion that can be optimized in the presence of a finite sample.

3.3 Kernel generalized variance for dimensionality reduction

To derive a sampled-based objective function from Eq. (19), we have to estimate the conditional covariance operator with given data, and choose a specific way to evaluate the size of self-adjoint operators.

For the estimation of the operator, we follow the procedure described by Bach and Jordan (2002a) in their derivation of kernel ICA. Let $\hat{K}_Y$ be the centralized Gram matrix (Bach and Jordan, 2002a, Schölkopf et al., 1998), defined by

$$\hat{K}_Y = \big(I_n - \tfrac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\big)G_Y\big(I_n - \tfrac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\big), \tag{20}$$

where $(G_Y)_{ij} = k_1(Y_i, Y_j)$ is the Gram matrix and $\mathbf{1}_n = (1, \ldots, 1)^T$ is the vector with all elements equal to 1. The matrices $\hat{K}_U$ and $\hat{K}_V$ are defined similarly, using $\{U_i\}_{i=1}^n$ and $\{V_i\}_{i=1}^n$, respectively. The empirical conditional covariance matrix $\hat{\Sigma}_{YY|U}$ is then defined by

$$\hat{\Sigma}_{YY|U} := \hat{\Sigma}_{YY} - \hat{\Sigma}_{YU}\hat{\Sigma}_{UU}^{-1}\hat{\Sigma}_{UY} = (\hat{K}_Y + \varepsilon I_n)^2 - \hat{K}_Y\hat{K}_U(\hat{K}_U + \varepsilon I_n)^{-2}\hat{K}_U\hat{K}_Y, \tag{21}$$

where $\varepsilon > 0$ is a regularization constant.

The size of $\hat{\Sigma}_{YY|U}$ in the ordered set of positive definite matrices can be evaluated by its determinant. Although there are other choices for measuring the size of $\hat{\Sigma}_{YY|U}$, such as the trace and the largest eigenvalue, we focus on the determinant in this paper. Using the Schur decomposition $\det(A - BC^{-1}B^T) = \det\begin{pmatrix} A & B \\ B^T & C \end{pmatrix}/\det C$, the determinant of $\hat{\Sigma}_{YY|U}$ can be written as follows:

$$\det \hat{\Sigma}_{YY|U} = \frac{\det \hat{\Sigma}_{[YU][YU]}}{\det \hat{\Sigma}_{UU}}, \tag{22}$$

where $\hat{\Sigma}_{[YU][YU]}$ is defined by

$$\hat{\Sigma}_{[YU][YU]} = \begin{pmatrix} \hat{\Sigma}_{YY} & \hat{\Sigma}_{YU} \\ \hat{\Sigma}_{UY} & \hat{\Sigma}_{UU} \end{pmatrix} = \begin{pmatrix} (\hat{K}_Y + \varepsilon I_n)^2 & \hat{K}_Y\hat{K}_U \\ \hat{K}_U\hat{K}_Y & (\hat{K}_U + \varepsilon I_n)^2 \end{pmatrix}. \tag{23}$$

We symmetrize the objective function by dividing by the constant $\det \hat{\Sigma}_{YY}$, which yields the following objective function

$$\frac{\det \hat{\Sigma}_{[YU][YU]}}{\det \hat{\Sigma}_{YY}\det \hat{\Sigma}_{UU}}. \tag{24}$$

We refer to the problem of minimizing this function with respect to the choice of subspace S as *Kernel Dimensionality Reduction (KDR)*.

Eq. (24) has been termed the "kernel generalized variance" by Bach and Jordan (2002a), who used it as a contrast function for independent component analysis. In that setting, the goal is to *minimize* a mutual information (among a set of recovered "source" variables), in the attempt to obtain independent components. Bach and Jordan (2002a) showed that the kernel generalized variance is in fact an approximation of the mutual information of the recovered sources, when this mutual information is expanded around the manifold of factorized distributions. In the current setting, on the other hand, our goal is to *maximize* the mutual information $I(Y,U)$, and we certainly do not expect to be near a manifold in which Y and U are independent. Thus the argument for the kernel generalized variance as an objective function in the ICA setting does not apply here. What we have provided in the previous section is an entirely distinct argument that shows that the kernel generalized variance is in fact an appropriate objective function for the dimensionality reduction problem, and that minimizing the kernel generalized variance in Eq. (24) can be viewed as a surrogate for maximizing the mutual information $I(Y,U)$.

Given that the numerical task that must be solved in KDR is the same as the numerical task that must be solved in kernel ICA, however, we can import all of the computational techniques developed by Bach and Jordan (2002a) for minimizing kernel generalized variance in the KDR setting. In particular, the optimization routine that we use in our experiments is gradient descent with a line search, where we exploit incomplete Cholesky decomposition to reduce the $n \times n$ matrices required in Eq. (24) to low-rank approximations. To cope with local optima, we make use of an annealing technique, in which the scale parameter σ for the Gaussian kernel is decreased gradually during the iterations of optimization. For a larger σ, the contrast function has fewer local optima, which makes optimization easier. The search becomes more accurate as σ is decreased.

4. Experimental results

We study the effectiveness of the new method through experiments, comparing it with several conventional methods: SIR, pHd, CCA, and PLS. For the experiments with SIR and pHd, we use an implementation for R due to Weisberg (2002).

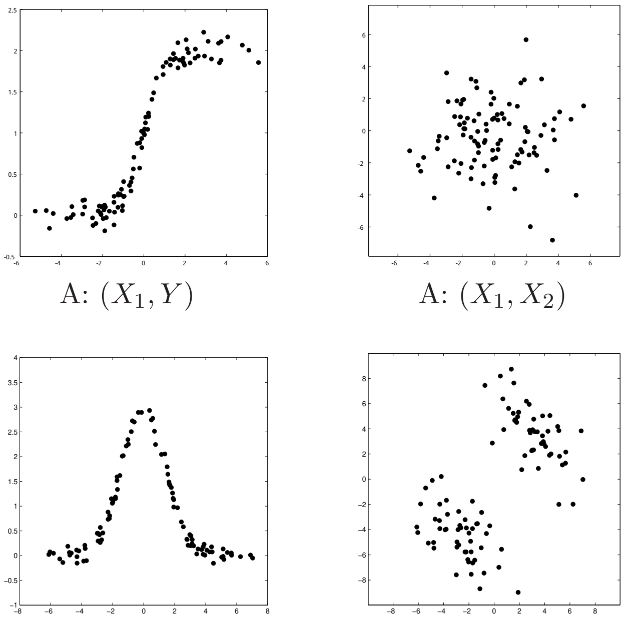
4.1 Synthetic data

The first data sets A and B comprise one-dimensional Y and two-dimensional $X = (X_1, X_2)$. One hundred *i.i.d.* data points are generated by

$$\begin{aligned} A: &\quad Y \sim 1/(1+\exp(-X_1)) + Z, \\ B: &\quad Y \sim 2\exp(-X_1^2) + Z, \end{aligned}$$

where $Z \sim N(0, 0.1^2)$, and $X = (X_1, X_2)$ follows a normal distribution and a normal mixture with two components for A and B, respectively. The effective subspace is spanned by $B_0 = (1,0)^T$ in both cases. The data sets are depicted in Figure 2.

Table 1 shows the angles between B_0 and the estimated direction. For Data A, all the methods except PLS yield a good estimate of B_0. Data B is surprisingly difficult for



A: (X_1, Y) A: (X_1, X_2)

B: (X_1, Y) B: (X_1, X_2)

Figure 2: Data A and B. One dimensional Y depends only on X_1 in $X = (X_1, X_2)$.

the conventional methods, presumably because the distribution of X is not spherical and the regressor has a strong nonlinearity. The KDR method succeeds in finding the correct direction for both data sets.

Data C has 300 samples of 17 dimensional X and one dimensional Y, which are generated by

$$C: \quad Y \sim 0.9X_1 + 0.2\frac{1}{1 + X_{17}} + Z, \tag{25}$$

where $Z \sim N(0, 0.01^2)$ and X follows a uniform distribution on $[0,1]^{17}$. The effective subspace is given by $\boldsymbol{b}_1 = (1, 0, \ldots, 0)$ and $\boldsymbol{b}_2 = (0, \ldots, 0, 1)$. We compare the KDR method with SIR and pHd only—CCA and PLS cannot find a 2-dimensional subspace, because Y is one-dimensional. To evaluate the accuracy of the results, we use the multiple correlation coefficient

$$R(\boldsymbol{b}) = \max_{\boldsymbol{\beta} \in B} \frac{\boldsymbol{\beta}^T \Sigma_{XX} \boldsymbol{b}}{\sqrt{\boldsymbol{\beta}^T \Sigma_{XX} \boldsymbol{\beta} \cdot \boldsymbol{b}^T \Sigma_{XX} \boldsymbol{b}}}, \qquad (\boldsymbol{b} \in B_0), \tag{26}$$

which is used in Li (1991). As shown in Table 2, the KDR method outperforms the others in finding the weak contribution of the second direction.

4.2 Real data: classification

In this section we apply the KDR method to classification problems. Many conventional methods of dimensionality reduction for regression are not suitable for classification. In particular, in the case of SIR, the dimensionality of the effective subspace must be less than the number of classes, because SIR uses the average of X in slices along the variable Y.

	SIR	pHd	CCA	PLS	Kernel
A: angle (rad.)	0.0087	-0.1971	0.0099	0.2736	-0.0014
B: angle (rad.)	-1.5101	-0.9951	-0.1818	0.4554	0.0052

Table 1: Angles between the true and the estimated spaces for Data A and B.

	SIR(10)	SIR(15)	SIR(20)	SIR(25)	pHd	Kernel
$R(\boldsymbol{b}_1)$	0.987	0.993	0.988	0.990	0.110	0.999
$R(\boldsymbol{b}_2)$	0.421	0.705	0.480	0.526	0.859	0.984

Table 2: Correlation coefficients for Data C. SIR(m) indicates the SIR with m slices.

Thus, in binary classification, only a one-dimensional subspace can be found, because at most two slices are available. The methods CCA and PLS have a similar limitation on the dimensionality of the effective subspace; they cannot find a subspace of larger dimensionality than that of Y. Thus our focus is the comparison between KDR and pHd, which is applicable to general binary classification problems. Note that Cook and Lee (1999) discuss dimensionality reduction methods for binary classification, and propose the *difference of covariance (DOC)* method. They compare pHd and DOC theoretically, and show that these methods are the same in binary classification if the population ratio of the classes is 1/2, which is almost the case in our experiments.

In the first experiment, we show the visualization capability of the dimensionality reduction methods. We use the *Wine* data set in the UCI machine learning repository (Murphy and Aha, 1994) to see how the projection onto a low-dimensional space realizes an effective description of data. The wine data consist of 178 samples with 13 variables and a label of three classes. We apply the KDR method, CCA, PLS, SIR, and pHd to these data. Figure 3 shows the projection onto the 2-dimensional subspace estimated by each method. The KDR method separates the data into three classes most completely, while CCA also shows perfect separation. We can see that the data are nonlinearly separable in the two-dimensional space. The other methods do not separate the classes completely.

Next we investigate how much information on Y is preserved in the estimated subspace. After reducing the dimensionality, we use the support vector machine (SVM) method to build a classifier in the reduced space, and compare its accuracy with an SVM trained using the full dimensional vector X[2]. We use the *Heart-disease* data set [3], *Ionosphere*, and *Wisconsin-breast-cancer* from the UCI repository. A description of these data is presented in Table 3.

Figure 4 shows the classification rates for the test set in subspaces of various dimensionality. We can see that KDR yields good separation even in low-dimensional subspaces,

2. In our experiments with the SVM, we used the Matlab Support Vector Toolbox by S. Gunn; see http://www.isis.ecs.soton.ac.uk/resources/svminfo.
3. We use the *Cleveland* data set, created by Dr. Robert Detrano of V.A. Medical Center, Long Beach and Cleveland Clinic Foundation. Although the original data set has five classes, we use only "no presence" (0) and "presence" (1-4) for the binary class labels. Samples with missing values are removed in our experiments.

Data set	dim. of X	training sample	test sample
Heart-disease	13	149	148
Ionosphere	34	151	200
Breast-cancer-Wisconsin	30	200	369

Table 3: Data description for the binary classification problem.

while pHd is much worse in low dimensions. It is noteworthy that in the Ionosphere data set the classifier in dimensions 5, 10, and 20 outperforms the classifier in the full dimensional space. This is presumably due to the suppression of noise irrelevant to the prediction of Y. These results show that the kernel method successfully finds an effective subspace which preserves the class information even when the dimensionality is reduced significantly.

5. Extension to variable selection

In this section, we describe an extension of the KDR method to the problem of variable selection. Variable selection is different from dimensionality reduction; the former involves selecting a subset of the explanatory variables $\{X_1, \ldots, X_m\}$ in order to obtain a simplified prediction of Y from X, while the latter involves finding linear combinations of the variables. However, the objective function that we have presented for dimensionality reduction can be extended straightforwardly to variable selection. In particular, given a fixed number of variables to be selected, we can compare the KGV for subspaces spanned by combinations of this number of selected variables. This gives a reasonable way to select variables, because for a subset $W = \{X_{j_1}, \ldots, X_{j_r}\} \subset \{X_1, \ldots, X_m\}$, the variables Y and W^C are conditionally independent given W if and only if Y and $\Pi_{W^c}X$ are conditionally independent given $\Pi_W X$, where Π_W and Π_{W^C} are the orthogonal projections onto the subspaces spanned by W and W^C, respectively. If we try to select r variables from among m explanatory variables, the total number of evaluations is $\binom{m}{r}$.

When $\binom{m}{r}$ is large, we must address the computational cost that arises in comparing large numbers of subsets. As in most other approaches to variable selection (see, e.g., Guyon and Elisseeff, 2003), we propose the use of a greedy algorithm and random search for this combinatorial aspect of the problem. (In the experiments presented in the current paper, however, we confine ourselves to small problems in which all combinations are tractably evaluated).

We apply this kernel-based method of variable selection to the *Boston Housing* data (Harrison and Rubinfeld, 1978) and the *Ozone* data (Breiman and Friedman, 1985), which have been often used as testbed examples for variable selection. Tables 4 and 5 give the detailed description of the data sets. There are 506 samples in the Boston Housing data, for which the variable MV, the median value of house prices in a tract, is estimated by using the 13 other variables. We use the corrected version of the data set given by Gilley and Pace (1996). In the Ozone data in which there are 330 samples, the variable UPO3 (the ozone concentration) is to be predicted by 9 other variables.

Table 6 shows the best three sets of four variables that attain the smallest values of the kernel generalized variance. For the Boston Housing data, RM and LSTAT are included in all the three of the result sets in Table 6, and PTRATIO and TAX are included in two

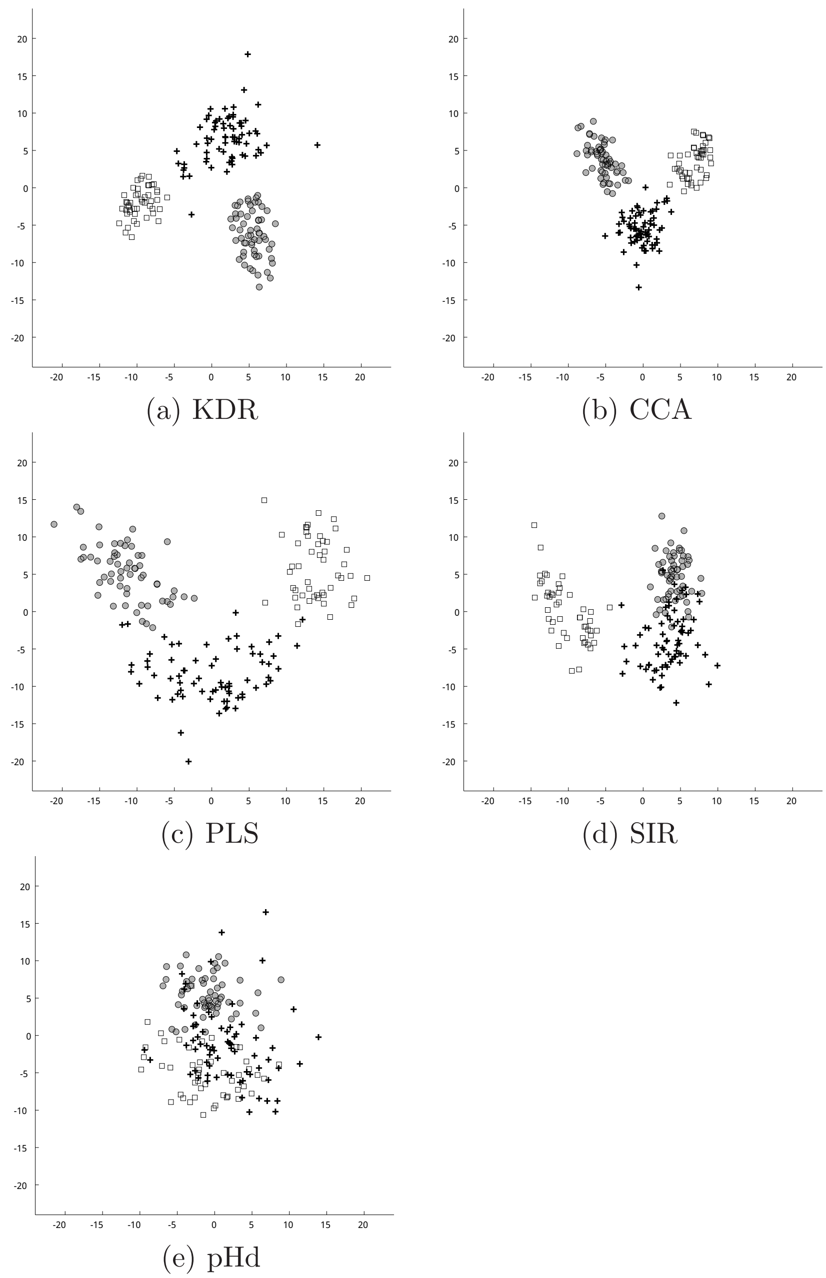


Figure 3: Wine data. Projections onto the estimated two-dimensional space. The symbols '+', '□', and gray '◯' represent the three classes.

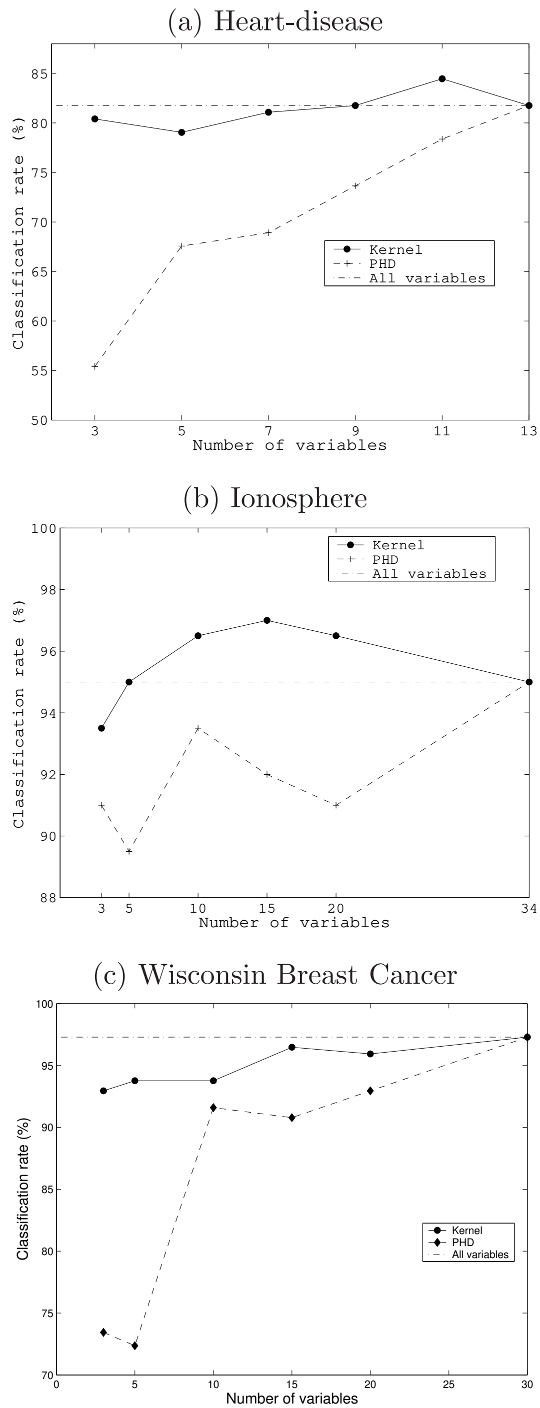


Figure 4: Classification accuracy of the SVM for test data after dimensionality reduction.

Variable		Description
MV	—	median value of owner-occupied home
CRIM	—	crime rate by town
ZN	—	proportion of town's residential land zoned for lots greater than 25,000 square feet
INDUS	—	proportion of nonretail business acres per town
CHAS	—	Charles River dummy (= 1 if tract bounds the Charles River, 0 otherwise)
NOX	—	nitrogen oxide concentration in pphm
RM	—	average number of rooms in owner units
AGE	—	proportion of owner units build prior to 1940
DIS	—	weighted distances to five employment centers in the Boston region
RAD	—	index of accessibility to radial highways
TAX	—	full property tax rate ($/$10,000)
PTRATIO	—	pupil-teacher ratio by town school district
B	—	black proportion of population
LSTAT	—	proportion of population that is lower status

Table 4: Boston Housing Data

of them. This observation agrees well with the analysis using alternating conditional expectation (ACE) by Breiman and Friedman (1985), which gives RM, LSTAT, PTRATIO, and TAX as the four major contributors. The original motivation in the study was to investigate the influence of nitrogen oxide concentration (NOX) on the house price (Harrison and Rubinfeld, 1978). In accordance with the previous studies, our analysis shows a relatively small contribution of NOX. For the Ozone data, all three of the result sets in the variable selection method include HMDT, SBTP, and IBHT. The variables IBTP, DGPG, and VDHT are chosen in one of the sets. This shows a fair accordance with earlier results by Breiman and Friedman (1985) and Li et al. (2000); the former concludes by ACE that SBTP, IBHT, DGPG, and VSTY are the most influential, and the latter selects HMDT, IBHT, and DGPG using a pHd-based method.

6. Conclusion

We have presented KDR, a new kernel-based approach to dimensionality reduction for regression and classification. One of the most notable aspects of this method is its generality—we do not impose any strong assumptions on either the conditional or the marginal distribution. This allows the method to be applicable to a wide range of problems, and gives it a significant practical advantage over existing methods such as CCA, PPR, SIR, pHd, and so on. These methods all impose significant restrictions on the conditional probability, the marginal distribution, or the dimensionality of the effective subspaces.

Our experiments have shown that the KDR method can provide many of the desired effects of dimensionality reduction: it provides data visualization capabilities, it can successfully select important explanatory variables in regression, and it can yield classification

Variable		Description
UPO3	—	upland ozone concentratin (ppm)
VDHT	—	Vandenburg 500 millibar height (m)
HMDT	—	himidity (percent)
IBHT	—	inversion base height (ft.)
DGPG	—	Daggett pressure gradient (mmhg)
IBTP	—	inversion base temperature (℉)
SBTP	—	Sandburg Air Force Base temperature (℃)
VSTY	—	visibility (miles)
WDSP	—	wind speed (mph)
DAY	—	day of the year

Table 5: Ozone data

Boston	1st	2nd	3rd
CRIM		X	
ZN			
INDUS			
CHAS			
NOX			
RM	X	X	X
AGE			
DIS			X
RAD			
TAX	X		X
PTRATIO	X	X	
B			
LSTAT	X	X	X
KGV	.1768	.1770	.1815

Ozone	1st	2nd	3rd
VDHT			X
HMDT	X	X	X
IBHT	X	X	X
DGPG		X	
IBTP	X		
SBTP	X	X	X
VSTY			
WDSP			
DAY			
KGV	.2727	.2736	.2758

Table 6: Variable selection using the proposed kernel method.

performance that is better than the performance achieved with the full-dimensional covariate space. We have also discussed the extension of the KDR method to variable selection. Experiments with classical data sets has shown an accordance with the previous results on these data sets and suggest that further study of this application of KDR is warranted.

The theoretical basis of KDR lies in the nonparametric characterization of conditional independence that we have presented in this paper. Extending earlier work on the kernel-based characterization of independence in ICA (Bach and Jordan, 2002a), we have shown that conditional independence can be characterized in terms of covariance operators on a reproducing kernel Hilbert space. While our focus has been on the problem of dimensionality reduction, it is also worth noting that there are many other possible applications of this characterization. In particular, conditional independence plays an important role in the structural definition of probabilistic graphical models, and our results may have applications to model selection and inference in graphical models.

There are several statistical problems which need to be addressed in further research on KDR. First, a basic analysis of the statistical consistency of the KDR-based estimator—the convergence of the estimator to the true subspace when such a space really exists—is needed. Second, and most significantly, we need rigorous methods for choosing the dimensionality of the effective subspace. If the goal is that of achieving high predictive performance after dimensionality reduction, we can use one of many existing methods (e.g., cross-validation, penalty-based methods) to assess the expected generalization as a function of dimensionality. Note in particular that by using KDR as a method to select an estimator given a fixed dimensionality, we have substantially reduced the number of hypotheses being considered, and expect to find ourselves in a regime in which methods such as cross-validation are likely to be effective. It is also worth noting, however, that the goals of dimensionality reduction are not always simply that of prediction; in particular, the search for small sets of explanatory variables will need to be guided by other principles. Finally, asymptotic analysis may provide useful guidance for selecting the dimensionality; an example of such an analysis that we believe can be adopted for KDR has been presented by Li (1991) for the SIR method.

Acknowledgments

This work was done while the first author was visiting the University of California, Berkeley. The authors thank Dr. Noboru Murata of Waseda University and Dr. Motoaki Kawanabe of Fraunhofer, FIRST for their helpful comments on the early version of this work. We wish to acknowledge support from JSPS KAKENHI 15700241, ONR MURI N00014-00-1-0637, NSF grant IIS-9988642, and a grant from Intel Corporation.

Appendix A. Cross-covariance operators on reproducing kernel Hilbert spaces and independence of random variables

A.1 Cross-covariance operators

While cross-covariance operators are generally defined for random variables on Banach spaces Vakhania et al. (1987), Baker (1973), they are more easily defined on reproducing kernel Hilbert spaces (RKHS). In this subsection, we summarize some of the basic mathematical facts used in Sections 3.1 and 3.3. While we discuss only real Hilbert spaces, extension to the complex case is straightforward.

Theorem 1 *Let $(\Omega_1, \mathcal{B}_1)$ and $(\Omega_2, \mathcal{B}_2)$ be measurable spaces, and let $(\mathcal{H}_1, k_1)$ and $(\mathcal{H}_2, k_2)$ be reproducing kernel Hilbert spaces on Ω_1 and Ω_2, respectively, with k_1 and k_2 measurable. Suppose we have a random vector (X, Y) on $\Omega_1 \times \Omega_2$ such that $E_X[k_1(X, X)]$ and $E_Y[k_2(Y, Y)]$ are finite. Then, there exists a unique operator Σ_{YX} from $\mathcal{H}_1$ to $\mathcal{H}_2$ such that*

$$\langle g, \Sigma_{YX} f \rangle_{\mathcal{H}_2} = E_{XY}[f(X)g(Y)] - E_X[f(X)]E_Y[g(Y)] \tag{27}$$

holds for all $f \in \mathcal{H}_1$ and $g \in \mathcal{H}_2$. This is called the cross-covariance operator*.*

Proof Obviously, the operator is unique, if it exists. From Riesz's representation theorem (see Reed and Simon, 1980, Theorem II.4, for example), the existence of $\Sigma_{YX} f \in \mathcal{H}_2$ for a fixed f can be proved by showing that the right hand side of Eq. (27) is a bounded linear functional on $\mathcal{H}_2$. The linearity is obvious, and the boundedness is shown by

$$\begin{aligned}
&\bigl| E_{XY}[f(X)g(Y)] - E_X[f(X)]E_Y[g(Y)] \bigr| \\
&/leq E_{XY} \bigl| \langle k_1(\cdot, X), f \rangle_{\mathcal{H}_1} \langle k_2(\cdot, Y), g \rangle_{\mathcal{H}_2} \bigr| + E_X \bigl| \langle k_1(\cdot, X), f \rangle_{\mathcal{H}_1} \bigr| \cdot E_Y \bigl| \langle k_2(\cdot, Y), g \rangle_{\mathcal{H}_2} \bigr| \\
&\leq E_{XY} \bigl[\|k_1(\cdot, X)\|_{\mathcal{H}_1} \|f\|_{\mathcal{H}_1} \|k_2(\cdot, Y)\|_{\mathcal{H}_2} \|g\|_{\mathcal{H}_2} \bigr] \\
&\qquad + E_X \bigl[\|k_1(\cdot, X)\|_{\mathcal{H}_1} \|f\|_{\mathcal{H}_1} \bigr] E_Y \bigl[\|k_2(\cdot, Y)\|_{\mathcal{H}_2} \|g\|_{\mathcal{H}_2} \bigr] \\
&\leq \bigl\{ E_X[k_1(X, X)]^{1/2} E_Y[k_2(Y, Y)]^{1/2} + E_X[k_1(X, X)^{1/2}] E_Y[k_2(Y, Y)^{1/2}] \bigr\} \|f\|_{\mathcal{H}_1} \|g\|_{\mathcal{H}_2}.
\end{aligned} \tag{28}$$

For the last inequality, $\|k(\cdot, x)\|_{\mathcal{H}}^2 = k(x, x)$ is used. The linearity of the map Σ_{YX} is given by the uniqueness part of Riesz's representation theorem. ■

From Eq. (28), Σ_{YX} is bounded, and by definition, we see $\Sigma_{YX}^* = \Sigma_{XY}$, where A^* denotes the adjoint of A. If the two RKHS are the same, the operator Σ_{XX} is called the *covariance operator*. A covariance operator Σ_{XX} is bounded, self-adjoint, and trace-class.

In an RKHS, conditional expectations can be expressed by cross-covariance operators, in a manner analogous to finite-dimensional Gaussian random variables.

Theorem 2 *Let $(\mathcal{H}_1, k_1)$ and $(\mathcal{H}_2, k_2)$ be RKHS on measurable spaces Ω_1 and Ω_2, respectively, with k_1 and k_2 measurable, and (X, Y) be a random vector on $\Omega_1 \times \Omega_2$. Assume that $E_X[k_1(X, X)]$ and $E_Y[k_2(Y, Y)]$ are finite, and for all $g \in \mathcal{H}_2$ the conditional expectation $E_{Y|X}[g(Y) \mid X = \cdot]$ is an element of $\mathcal{H}_1$. Then, we have for all $g \in \mathcal{H}_2$*

$$\Sigma_{XX} E_{Y|X}[g(Y) \mid X] = \Sigma_{XY} g, \tag{29}$$

where Σ_{XX} and Σ_{XY} are the covariance and cross-covariance operator.

Proof For any $f \in \mathcal{H}_1$, we have

$$\begin{aligned}
&\langle f, \Sigma_{XX} E_{Y|X}[g(Y) \mid X] \rangle_{\mathcal{H}_1} \\
&\qquad = E_X \bigl[f(X) E_{Y|X}[g(Y) \mid X] \bigr] - E_X[f(X)] E_X \bigl[E_{Y|X}[g(Y) \mid X] \bigr] \\
&\qquad = E_{XY}[f(X)g(Y)] - E_X[f(X)]E_Y[g(Y)] \quad = \quad \langle f, \Sigma_{XY} g \rangle_{\mathcal{H}_1}.
\end{aligned}$$

This completes the proof. ■

Corollary 3 *Let $\tilde{\Sigma}_{XX}^{-1}$ be the right inverse of Σ_{XX} on $(Ker\Sigma_{XX})^{\perp}$. Under the same assumptions as Theorem 2, we have*

$$\langle f, \tilde{\Sigma}_{XX}^{-1}\Sigma_{XY}g\rangle = \langle f, E_{Y|X}[g(Y) \mid X]\rangle \tag{30}$$

for all $f \in (\mathrm{Ker}\Sigma_{XX})^{\perp}$ and $g \in \mathcal{H}_2$. In particular, if $\mathrm{Ker}\Sigma_{XX} = 0$, we have

$$\Sigma_{XX}^{-1}\Sigma_{XY}g = E_{Y|X}[g(Y) \mid X]. \tag{31}$$

Proof Note that the product $\tilde{\Sigma}_{XX}^{-1}\Sigma_{XY}$ is well-defined, because $\overline{\mathrm{Range}\Sigma_{XY}} \subset \overline{\mathrm{Range}\Sigma_{XX}} = (\mathrm{Ker}\Sigma_{XX})^{\perp}$. The first inclusion is shown from the expression $\Sigma_{XY} = \Sigma_{XX}^{1/2}V\Sigma_{YY}^{1/2}$ with a bounded operator V (Baker, 1973, Theorem 1), and the second equation holds for any self-adjoint operator. Take $f = \Sigma_{XX}h \in \mathrm{Range}\Sigma_{XX}$. Then, Theorem 2 yields

$$\begin{aligned}\langle f, \tilde{\Sigma}_{XX}^{-1}\Sigma_{XY}g\rangle &= \langle h, \Sigma_{XX}\tilde{\Sigma}_{XX}^{-1}\Sigma_{XX}E_{Y|X}[g(Y) \mid X]\rangle \\ &= \langle h, \Sigma_{XX}E_{Y|X}[g(Y) \mid X]\rangle = \langle f, E_{Y|X}[g(Y) \mid X]\rangle.\end{aligned}$$

This completes the proof. ■

The assumption $E_{Y|X}[g(Y) \mid X = \cdot] \in \mathcal{H}_1$ in Theorem 2 can be simplified so that it can be checked without reference to a specific g.

Proposition 4 *Under the condition of Theorem 2, if there exists $C > 0$ such that*

$$E_{Y|X}[k_2(y_1, Y) \mid X = x_1]E_{Y|X}[k_2(y_2, Y) \mid X = x_2] \le Ck_1(x_1, x_2)k_2(y_1, y_2) \tag{32}$$

for all $x_1, x_2 \in \Omega_1$ and $y_1, y_2 \in \Omega_2$, then for all $g \in \mathcal{H}_2$ the conditional expectation $E_{Y|X}[g(Y) \mid X = \cdot]$ is an element of $\mathcal{H}_1$.

Proof See Theorem 2.3.13 in Alpay (2001). ■

For a function f in an RKHS, the expectation of $f(X)$ can be formulated as the inner product of f and a fixed element. Let $(\Omega, \mathcal{B})$ be a measurable space, and $(\mathcal{H}, k)$ be an RKHS on Ω with k measurable. Note that for a random variable X on Ω, the linear functional $f \mapsto E_X[f(X)]$ is bounded if $E_X[k(X, X)]$ exists. By Riesz's theorem, there is $u \in \mathcal{H}$ such that $\langle u, f\rangle_{\mathcal{H}} = E_X[f(X)]$ for all $f \in \mathcal{H}$. If we define $E_X[k(\cdot, X)] \in \mathcal{H}$ by this element u, we formally obtain the equality

$$\langle E_X[k(\cdot, X)], f\rangle_{\mathcal{H}} = E_X[\langle k(\cdot, X), f\rangle_{\mathcal{H}}], \tag{33}$$

which looks like the interchangeability of the expectation by X and the inner product. While the expectation $E_X[k(\cdot, X)]$ can be defined, in general, as an integral with respect to the distribution on $\mathcal{H}$ induced by $k(\cdot, X)$, the element $E_X[k(\cdot, X)]$ is formally obtained as above in a reproducing kernel Hilbert space.

A.2 Conditional covariance operator and conditional independence

We define the conditional (cross-)covariance operator, and derive its relation with the conditional covariance of random variables. Let $(\mathcal{H}_1, k_1)$, $(\mathcal{H}_2, k_2)$, let $(\mathcal{H}_3, k_3)$ be RKHS on measurable spaces Ω_1, Ω_2, and Ω_3, respectively, and let (X, Y, Z) be a random vector on $\Omega_1 \times \Omega_2 \times \Omega_3$. The *conditional cross-covariance operator of* (X, Y) *given* Z is defined by

$$\Sigma_{YX|Z} := \Sigma_{YX} - \Sigma_{YZ}\tilde{\Sigma}_{ZZ}^{-1}\Sigma_{ZX}. \tag{34}$$

Because $\mathrm{Ker}\Sigma_{ZZ} \subset \mathrm{Ker}\Sigma_{YZ}$ from the fact $\Sigma_{YZ} = \Sigma_{YY}^{1/2} V \Sigma_{ZZ}^{1/2}$ for some bounded operator V (Baker, 1973, Theorem 1), the operator $\Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{YX}$ can be uniquely defined, even if Σ_{ZZ}^{-1} is not unique. By abuse of notation, we write $\Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX}$, when cross-covariance operators are discussed.

The conditional cross-covariance operator is related to the conditional covariance of the random variables.

Proposition 5 *Let $(\mathcal{H}_1, k_1)$, $(\mathcal{H}_2, k_2)$, and $(\mathcal{H}_3, k_3)$ be reproducing kernel Hilbert spaces on measurable spaces Ω_1, Ω_2, and Ω_3, respectively, with k_i measurable, and let (X, Y, Z) be a measurable random vector on $\Omega_1 \times \Omega_2 \times \Omega_3$ such that $E_X[k_1(X, X)]$, $E_Y[k_2(Y, Y)]$, and $E_Z[k_3(Z, Z)]$ are finite. It is assumed that $E_{X|Z}[f(X) \mid Z]$ and $E_{Y|Z}[g(Y) \mid Z]$ are elements of $\mathcal{H}_3$ for all $f \in \mathcal{H}_1$ and $g \in \mathcal{H}_2$. Then, for all $f \in \mathcal{H}_1$ and $g \in \mathcal{H}_2$, we have*

$$\begin{aligned}\langle g, \Sigma_{YX|Z} f\rangle_{\mathcal{H}_2} &= E_{XY}[f(X)g(Y)] - E_Z\big[E_{X|Z}[f(X) \mid Z]E_{Y|Z}[g(Y) \mid Z]\big] \\ &= E_Z\big[\mathrm{Cov}_{XY|Z}\big(f(X), g(Y) \mid Z\big)\big]. \end{aligned} \tag{35}$$

Proof From the decomposition $\Sigma_{YZ} = \Sigma_{YY}^{1/2} V \Sigma_{ZZ}^{1/2}$, we have $\Sigma_{ZY} g \in (\mathrm{Ker}\Sigma_{ZZ})^{\perp}$. Then, by Corollary 3, we obtain

$$\begin{aligned}\langle g, \Sigma_{YZ}\tilde{\Sigma}_{ZZ}^{-1}\Sigma_{ZX} f\rangle = \langle \Sigma_{ZY} g, \tilde{\Sigma}_{ZZ}^{-1}\Sigma_{ZX} f\rangle &= \langle \Sigma_{ZY} g, E_{X|Z}[f(X) \mid Z]\rangle \\ &= E_{YZ}\big[g(Y)E_{X|Z}[f(X) \mid Z]\big] - E_X[f(X)]E_Y[g(Y)]. \end{aligned}$$

From this equation, the theorem is proved by

$$\begin{aligned}\langle g, \Sigma_{YX|Z} f\rangle &= E_{XY}[f(X)g(Y)] - E_X[f(X)]E_Y[g(Y)] \\ &\quad - E_{YZ}\big[g(Y)E_{X|Z}[f(X) \mid Z]\big] + E_X[f(X)]E_Y[g(Y)] \\ &= E_{XY}[f(X)g(Y)] - E_Z\big[E_{X|Z}[f(X) \mid Z]E_{Y|Z}[g(Y) \mid Z]\big]. \end{aligned} \tag{36}$$

∎

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

$$\mathcal{S} \ni P \quad \mapsto \quad (f \mapsto E_{X\sim P}[f(X)]) \in \mathcal{H}^* \tag{37}$$

is one-to-one, where $\mathcal{H}^*$ is the dual space of $\mathcal{H}$. From Riesz's theorem, $\mathcal{H}$ is probability-determining if and only if the map

$$\mathcal{S} \ni P \quad \mapsto \quad E_{X\sim P}[k(\cdot, X)] \in \mathcal{H}$$

is one-to-one. Theorem 2 in (Bach and Jordan, 2002a) shows the following fact:

Theorem 6 (Bach and Jordan 2002a) *For an arbitrary $\sigma > 0$, the reproducing kernel Hilbert space with Gaussian kernel $k(x, y) = \exp(-\|x - y\|^2/\sigma)$ on $\mathbb{R}^m$ is probability-determining.*

Recall that for two RKHS $\mathcal{H}_1$ and $\mathcal{H}_2$ on Ω_1 and Ω_2, respectively, the direct product $\mathcal{H}_1 \otimes \mathcal{H}_2$ is the RKHS on $\Omega_1 \times \Omega_2$ with the positive definite kernel k_1k_2 (see Aronszajn, 1950). The relation between conditional independence and the conditional covariance operator is given by the following theorem:

Theorem 7 *Let $(\mathcal{H}_{11}, k_{11})$, $(\mathcal{H}_{12}, k_{12})$, and $(\mathcal{H}_2, k_2)$ be reproducing kernel Hilbert spaces on measurable spaces Ω_{11}, Ω_{12}, and Ω_2, respectively, with continuous and bounded kernels. Let $(X, Y) = (Z, W, Y)$ be a random vector on $\Omega_{11} \times \Omega_{12} \times \Omega_2$, where $X = (Z, W)$, and let $\mathcal{H}_1 = \mathcal{H}_{11} \otimes \mathcal{H}_{12}$ be the direct product. It is assumed that $E_{Y|Z}[g(Y) \mid Z] \in \mathcal{H}_{11}$ and $E_{Y|X}[g(Y) \mid X] \in \mathcal{H}_1$ for all $g \in \mathcal{H}_2$. Then, we have*

$$\Sigma_{YY|Z} \geq \Sigma_{YY|X}, \tag{38}$$

where the inequality refers to the order of self-adjoint operators, and if further $\mathcal{H}_2$ is probability-determining, the following equivalence holds

$$\Sigma_{YY|X} = \Sigma_{YY|Z} \qquad \Longleftrightarrow \qquad Y \perp\!\!\!\perp W \mid Z. \tag{39}$$

Proof The right hand side of Eq. (39) is equivalent to $P_{Y|X} = P_{Y|Z}$, where $P_{Y|X}$ and $P_{Y|Z}$ are the conditional probability of Y given X and given Z, respectively. Taking the expectation of the well-known equality

$$V_{Y|Z}[g(Y) \mid Z] = E_{W|Z}\big[V_{Y|Z,W}[g(Y) \mid Z, W]\big] + V_{W|Z}\big[E_{Y|Z,W}[g(Y) \mid Z, W]\big] \tag{40}$$

with respect to Z, we derive

$$E_Z\big[V_{Y|Z}[g(Y) \mid Z]\big] = E_X\big[V_{Y|X}[g(Y) \mid X]\big] + E_Z\big[V_{W|Z}[E_{Y|X}[g(Y) \mid X]]\big]. \tag{41}$$

Since the last term of Eq. (41) is nonnegative, we obtain Eq. (38) from Proposition 5.

Equality holds if and only if $V_{W|Z}[E_{Y|X}[g(Y) \mid X]] = 0$ for almost every Z, which means $E_{Y|X}[g(Y) \mid X]$ does not depend on W almost surely. This is equivalent to

$$E_{Y|X}[g(Y) \mid X] = E_{Y|Z}[g(Y) \mid Z] \tag{42}$$

for almost every Z and W. Because $\mathcal{H}_2$ is probability-determining, this means $P_{Y|X} = P_{Y|Z}$. ■

A.3 Conditional cross-covariance operator and conditional independence

Theorem 7 characterizes conditional independence using the conditional covariance operator. Another formulation is possible with a conditional cross-covariance operator.

Let $(\Omega_1, \mathcal{B}_1)$, $(\Omega_2, \mathcal{B}_2)$, and $(\Omega_3, \mathcal{B}_3)$ be measurable spaces, and let (X, Y, Z) be a random vector on $\Omega_1 \times \Omega_2 \times \Omega_3$ with law P_{XYZ}. We define a probability measure $E_Z[P_{X|Z} \otimes P_{Y|Z}]$ on $\Omega_1 \times \Omega_2$ by

$$E_Z[P_{X|Z} \otimes P_{Y|Z}](A \times B) = E_Z\big[E_{X|Z}[\chi_A|Z]\, E_{Y|Z}[\chi_B \mid Z]\big], \tag{43}$$

where χ_A is the characteristic function of a measurable set A. It is canonically extended to any product-measurable sets in $\Omega_1 \times \Omega_2$.

Theorem 8 *Let $(\Omega_i, \mathcal{B}_i)$ $(i = 1, 2, 3)$ be a measurable space, let $(\mathcal{H}_i, k_i)$ be a RKHS on Ω_i with kernel measurable and bounded, and let (X, Y, Z) be a random vector on $\Omega_1 \times \Omega_2 \times \Omega_3$. It is assumed that $E_{X|Z}[f(X) \mid Z]$ and $E_{Y|Z}[g(Y) \mid Z]$ belong to $\mathcal{H}_3$ for all $f \in \mathcal{H}_1$ and $g \in \mathcal{H}_2$, and that $\mathcal{H}_1 \otimes \mathcal{H}_2$ is probability-determining. Then, we have*

$$\Sigma_{YX|Z} = O \qquad \Longleftrightarrow \qquad P_{XY} = E_Z[P_{X|Z} \otimes P_{Y|Z}]. \tag{44}$$

Proof The right-to-left direction is trivial from Theorem 5 and the definition of $E_Z[P_{X|Z} \otimes P_{Y|Z}]$. The left-hand side yields $E_Z[E_{X|Z}[f(X) \mid Z]E_{Y|Z}[g(Y) \mid Z]] = E_{XY}[f(X)g(Y)]$ for all $f \in \mathcal{H}_1$ and $g \in \mathcal{H}_2$. By the definition of $\mathcal{H}_1 \otimes \mathcal{H}_2$, we have $E_{(X',Y')\sim Q}[h(X', Y')] = E_{XY}[h(X, Y)]$ for all $h \in \mathcal{H}_1 \otimes \mathcal{H}_2$, where $Q = E_Z[P_{X|Z} \otimes P_{Y|Z}]$. This implies the right-hand side, because $\mathcal{H}_1 \otimes \mathcal{H}_2$ is probability-determining. ■

The right-hand side of Eq. (44) is weaker than the conditional independence of X and Y given Z. However, if Z is a part of X, we obtain conditional independence.

Corollary 9 *Let $(\mathcal{H}_{11}, k_{11})$, $(\mathcal{H}_{12}, k_{12})$, and $(\mathcal{H}_2, k_2)$ be reproducing kernel Hilbert spaces on measurable spaces Ω_{11}, Ω_{12}, and Ω_2, respectively, with kernels measurable and bounded. Let $(X, Y) = (Z, W, Y)$ be a random vector on $\Omega_{11} \times \Omega_{12} \times \Omega_2$, where $X = (Z, W)$, and let $\mathcal{H}_1 = \mathcal{H}_{11} \otimes \mathcal{H}_{12}$ be the direct product. It is assumed that $E_{X|Z}[f(X) \mid Z]$ and $E_{Y|Z}[g(Y) \mid Z]$ belong to $\mathcal{H}_{11}$ for all $f \in \mathcal{H}_1$ and $g \in \mathcal{H}_2$, and $\mathcal{H}_1 \otimes \mathcal{H}_2$ is probability-determining. Then, we have*

$$\Sigma_{YX|Z} = O \qquad \Longleftrightarrow \qquad Y \perp\!\!\!\perp W \mid Z. \tag{45}$$

Proof For any measurable sets $A \subset \Omega_{11}$, $B \subset \Omega_{12}$, and $C \subset \Omega_2$, we have, in general,

$$\begin{aligned}
&E_Z\big[E_{X|Z}[\chi_{A\times B}(Z,W) \mid Z]E_{Y|Z}[\chi_C(Y) \mid Z]\big] - E_{XY}[\chi_{A\times B}(Z,W)\chi_C(Y)] \\
&= E_Z\big[E_{W|Z}[\chi_B(W) \mid Z]\chi_A(Z)E_{Y|Z}[\chi_C(Y) \mid Z]\big] - E_Z\big[E_{WY|Z}[\chi_B(W)\chi_C(Y) \mid Z]\chi_A(Z)\big] \\
&= \int_A \big\{P_{W|Z}(B \mid z)P_{Y|Z}(C \mid z) - P_{WY|Z}(B \times C \mid z)\big\}dP_Z(z).
\end{aligned} \tag{46}$$

From Theorem 8, the left-hand side of Eq. (45) is equivalent to $E_Z[P_{X|Z} \otimes P_{Y|Z}] = P_{XY}$, which implies that the last integral in Eq. (46) is zero for all A. This means $P_{W|Z}(B \mid z)P_{Y|Z}(C \mid z) - P_{WY|Z}(B \times C \mid z) = 0$ for almost every z-P_Z. Thus, Y and W are conditional independent given Z. The converse is trivial. ■

Note that the left-hand side of Eq. (45) is not $\Sigma_{YW|Z}$ but $\Sigma_{YX|Z}$, which is defined on the direct product $\mathcal{H}_{11} \otimes \mathcal{H}_{12}$.

References

Daniel Alpay. *The Schur Algorithm, Reproducing Kernel Spaces and System Theory*. American Mathematical Society, 2001.

Nachman Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 69(3): 337–404, 1950.

Francis R. Bach and Michael I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002a.

Francis R. Bach and Michael I. Jordan. Tree-dependent component analysis. In D. Mozer and N. Friedman, editors, *Uncertainty in Artificial Intelligence: Proceedings of the Eighteenth Conference)*, San Mateo, CA, 2002b. Morgan Kaufmann.

Francis R. Bach and Michael I. Jordan. Learning graphical models with Mercer kernels. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*. MIT Press, Cambridge, MA, 2003.

Charles R. Baker. Joint measures and cross-covariance operators. *Trans. Amer. Math. Soc.*, 186:273–289, 1973.

Christopher M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, 1995.

Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In D. Haussler, editor, *Fifth Annual ACM Workshop on Computational Learning Theory*, pages 144–152, Pittsburgh, PA, 1992. ACM Press.

Leo Breiman and Jerome H. Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80:580–598, 1985.

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

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

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

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

Jerome H. Friedman and Werner Stuetzle. Projection pursuit regression. *Journal of the American Statistical Association*, 76:817–823, 1981.

Wing Kam Fung, Xuming He, Li Liu, and Peide Shi. Dimension reduction based on canonical correlation. *Statistica Sinica*, 12(4):1093–1114, 2002.

Otis W. Gilley and R. Kelly Pace. On the Harrison and Rubingeld data. *Journal of Environmental Economics Management*, 31:403–405, 1996.

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

David Harrison and Daniel L. Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics Management*, 5:81–102, 1978.

Trevor Hastie and Robert Tibshirani. Generalized additive models. *Statistical Science*, 1: 297–318, 1986.

Inge S. Helland. On the structure of partial least squares. *Communications in Statistics - Simulations and Computation*, 17(2):581–607, 1988.

Agnar Höskuldsson. PLS regression methods. *Journal of Chemometrics*, 2:211–228, 1988.

Marian Hristache, Anatoli Juditsky, Jörg Polzehl, and Vladimir Spokoiny. Structure adaptive approach for dimension reduction. *The Annals of Statistics*, 29(6):1537–1566, 2001.

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

Ker-Chau 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.

Ker-Chau Li, Heng-Hui Lue, and Chun-Houh Chen. Interactive tree-structured regression via principal Hessian directions. *Journal of the American Statistical Association*, 95(450): 547–560, 2000.

Patrick M. Murphy and David W. Aha. UCI repository of machine learning databases. Technical report, University of California, Irvine, Department of Information and Computer Science. http://www.ics.uci.edu/˜mlearn/MLRepository.html, 1994.

Danh V. Nguyen and David M. Rocke. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, 18(1):39–50, 2002.

Michael Reed and Barry Simon. *Functional Analysis*. Academic Press, 1980.

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

Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.

Kari Torkkola. Feature extraction by non-parametric mutual information maximization. *Journal of Machine Learning Research*, 3:1415–1438, 2003.

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

Vladimir N. Vapnik, Steven E. Golowich, and Alexander J. Smola. Support vector method for function approximation, regression estimation, and signal processing. In M. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems 9*, pages 281–287, Cambridge, MA, 1997. MIT Press.

Sanford Weisberg. Dimension reduction regression in R. *Journal of Statistical Software*, 7 (1), 2002.