Covariance Operator Based Dimensionality Reduction with Extension to Semi-Supervised Settings

Minyoung Kim and Vladimir Pavlovic
Department of Computer Science
Rutgers University, Piscataway, NJ 08854

Abstract

We consider the task of dimensionality reduction for regression (DRR) informed by real-valued multivariate labels. The problem is often treated as a regression task where the goal is to find a low dimensional representation of the input data that preserves the statistical correlation with the targets. Recently, Covariance Operator Inverse Regression (COIR) was proposed as an effective solution that exploits the covariance structures of both input and output. COIR addresses known limitations of recent DRR techniques and allows a closed-form solution without resorting to explicit output space slicing often required by existing IR-based methods. In this work we provide a unifying view of COIR and other DRR techniques and relate them to the popular supervised dimensionality reduction methods including the canonical correlation analysis (CCA) and the linear discriminant analysis (LDA). We then show that COIR can be effectively extended to a semi-supervised learning setting where many of the input points lack their corresponding multivariate targets. A study of benefits of proposed approaches is presented on several important regression problems in both fully-supervised and semi-supervised settings.

1 Introduction

Dimensionality reduction is a basic problem in modern machine learning, driven by applications such as the data visualization and compression. A large literature on dimension reduction is devoted to the *unsupervised* setting where one is given a set of data samples *alone*. Within this setting, discovering a low dimensional structure of the data can be accomplished by either extracting a global statistical information (e.g., PCA) or exploiting the geometric nature of data (e.g., LLE (Roweis and Saul, 2000), ISOMAP (Tenenbaum et al., 2000)). In the *supervised* setting, on the other hand, data is accompanied with additional label information that guides the formation of the low-dimensional embedded space. The labels often take discrete class values, indicating which data points have to be grouped together or far apart from one another in the embedded space. This class-labeled supervised dimension reduction framework, sometimes referred to as *metric learning*, has received considerable attention in the community. The well-known (kernel) Linear Discriminant Analysis and its generalizations (Globerson and Roweis, 2005; Weinberger et al., 2005) are some examples of this framework.

In certain situations, however, grouping data into a finite number of classes may be inappropriate. In many applications is reasonable to regard the label as a smoothly varying response (of the underlying phenomenon) in a *real-valued multivariate* domain. Existing class-labeled techniques may not be well suited for this setting. Instead, one can treat the problem in a regression framework where the targets (output responses) are regressed from data points (input covariates). In such a setting the task of dimension reduction, termed *dimension reduction for regression* (DRR), is formally defined as finding a low dimensional representation $\mathbf{z} \in \mathbb{R}^q$ of the input $\mathbf{x} \in \mathbb{R}^p$ ($q \ll p$) for regressing the output $\mathbf{y} \in \mathbb{R}^d$. DRR is well-suited for visualization of high-dim data, efficient regressor design with a reduced input dimension, and eliminating noise in data $\mathbf{x}$ through uncovering the essential information $\mathbf{z}$ for predicting $\mathbf{y}$.

A crucial notion related to DRR is the *sufficiency in*

Appearing in Proceedings of the 12^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2009, Clearwater Beach, Florida, USA. Volume 5 of JMLR: W&CP 5. Copyright 2009 by the authors.

dimension reduction (SDR) (Cook, 1998; Fukumizu et al., 2004; Li, 1991). Formally, SDR states that finding the low-dimensional representation is equivalent to recovering the subspace bases (or basis functions in a nonlinear case) $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_q]$ such that $\mathbf{y}$ and $\mathbf{x}$ are conditionally independent given $\mathbf{z} = \mathbf{B}^\top \mathbf{x}$, i.e., $\mathbf{y} \perp \mathbf{x} \mid \mathbf{z}$. This, in turn, implies that the dimension reduction is achieved with no loss of information for the purpose of predicting $\mathbf{y}$ from $\mathbf{x}$. A (minimal) subspace with this property is called the *central subspace*[1].

A number of methods originating in the statistics community have tackled the task of recovering the central subspace. The kernel dimension reduction (KDR) (Fukumizu et al., 2004) and the manifold KDR (mKDR) (Nilsson et al., 2007) directly reduces the task of imposing conditional independence to the optimization problem of minimizing the conditional covariance operator in RKHS (reproducing kernel Hilbert space). However, both methods introduce non-convex objectives, potentially suffering from existence of local minima. Alternatively, the inverse regression (IR) approach (Li, 1991) exploits the fact that the inverse regression $\mathbb{E}[\mathbf{x}|\mathbf{y}]$ can lie on the central subspace, leading to the possibility of estimating $\mathbf{B}$ (bases for the central subspace) from the slice-driven covariance estimates of the IR, the Sliced IR (SIR). While its kernel extension, KSIR (Wu, 2006), overcomes the linearity of SIR its performance may still suffer from the need to slice $\mathbf{y}$, which is suboptimal and can be unreliable for high dimensional outputs.

Covariance Operator Inverse Regression (COIR), a nonlinear method for DRR that jointly exploits the covariance structure of both input and output while preserving the input-output dependency, was recently proposed in (Kim and Pavlovic, 2008). COIR avoids explicit slicing of targets through an effective use of the covariance operators in RKHS. (Kim and Pavlovic, 2008) showed that COIR generalizes KSIR and allows a closed-form solution to the nonlinear central subspace estimation problem. In this paper we further study the properties of COIR and present a unifying view of COIR and other DRR techniques. We demonstrate that despite the apparent difference in their motivating tasks, the central subspace of COIR is identical to those recovered by KCCA (Hardoon et al., 2004), the kernelized canonical correlation analysis, and a generalization of the linear discriminant analysis (LDA), a method traditionally framed in classification settings.

We also extend the COIR framework to a semi-supervised setting, making it feasible to use large subsets of unlabeled data points in conjunction with a few labeled data. We follow the manifold regularization of (Belkin et al., 2005; Zhu et al., 2003) to affect the underlying geometry of the central subspace In addition, we introduce a nonlinear extension that admits arbitrary output kernel functions.

The paper is organized as follows. In Sec. 2 we briefly review related DRR approaches: KDR/mKDR and SIR/KSIR. We present COIR in Sec. 3, and establish a unifying relationship among different DRR methods in Sec. 4. We then discuss semi-supervised extensions of COIR in Sec. 5. In Sec. 6 the benefits of the proposed approaches are demonstrated on several regression problems in both fully-supervised and semi-supervised settings, followed by conclusion in Sec. 7.

2 Previous Approaches

Throughout the paper (except Sec. 5), we assume fully labeled data $\{(\mathbf{x}_i \in \mathbb{R}^p, \mathbf{y}_i \in \mathbb{R}^d)\}_{i=1}^n$, n i.i.d. samples from the (unknown) distribution $P(\mathbf{x}, \mathbf{y})$. All expectations and (co)variances that appear in the paper are w.r.t. $P(\mathbf{x}, \mathbf{y})$. We assume that the data points are centered at 0 without loss of generality.

2.1 Kernel Dimensionality Reduction

The kernel dimensionality reduction (KDR) (Fukumizu et al., 2004) finds a q-dimensional linear embedding matrix $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_q]$ ($\mathbf{b}_l \in \mathbb{R}^p$ for $l = 1, \ldots, q$) by directly reducing the SDR criterion (i.e., the task of imposing $\mathbf{y} \perp \mathbf{x} \mid \mathbf{z} = \mathbf{B}^\top \mathbf{x}$) to an optimization problem. The main idea is to quantify the notion of conditional dependency by a positive definite ordering of the expected covariance operators in what is called the probability-determining RKHS (e.g., the Hilbert space induced by the RBF kernel). More specifically, for two RKHS mappings, $\mathbf{y} \to \boldsymbol{\phi}(\mathbf{y}) \in \mathcal{H}_\mathbf{y}$ and $\mathbf{x} \to \boldsymbol{\phi}(\mathbf{x}) \in \mathcal{H}_\mathbf{x}$ induced by RBF kernels $k_\mathbf{y}(\cdot, \cdot)$ and $k_\mathbf{x}(\cdot, \cdot)$, respectively, we have the following theorem (Fukumizu et al., 2004):

Theorem 1. $\mathbb{E}[\mathbb{V}(\boldsymbol{\phi}(\mathbf{y})|\mathbf{x})] \preceq \mathbb{E}[\mathbb{V}(\boldsymbol{\phi}(\mathbf{y})|\mathbf{B}^\top \mathbf{x})]$, *where the equality holds if and only if* $\mathbf{y} \perp \mathbf{x} \mid \mathbf{z} = \mathbf{B}^\top \mathbf{x}$.

In KDR we minimize $\mathbb{E}[\mathbb{V}(\boldsymbol{\phi}(\mathbf{y})|\mathbf{z} = \mathbf{B}^\top \mathbf{x})]$, the uncertainty in predicting $\mathbf{y}$ from $\mathbf{z}$, which can be formulated as the following optimization problem:

$$\min_{\mathbf{B}} \ \mathrm{tr}\big[\mathbf{K}_\mathbf{y}(\mathbf{K}_\mathbf{z} + n\epsilon \mathbf{I}_n)^{-1}\big] \quad \text{s.t.} \ \ \mathbf{B}^\top \mathbf{B} = \mathbf{I}_q, \tag{1}$$

where $\mathbf{K}_\mathbf{y}$ and $\mathbf{K}_\mathbf{z}$ are $(n \times n)$ (centered) kernel Gram matrices computed over $\{\mathbf{y}_i\}_{i=1}^n$ and $\{\mathbf{z}_i = \mathbf{B}^\top \mathbf{x}_i\}_{i=1}^n$, respectively. $\mathbf{I}_a$ is the $(a \times a)$ identity matrix, and ϵ is a kernel regularizer. Although KDR does not assume any particular restriction on the underlying distribution $P(\mathbf{x}, \mathbf{y})$, the optimization of Eq.(1) is non-

[1] Although a *subspace* is usually meant for a linear case, we abuse it for referring to both linear and nonlinear cases.

convex, resorting to computationally demanding gradient search (every step requires inversion of Gram matrices). Moreover, despite its formulation in RKHS, KDR's final embedding is linear in the original space.

For a nonlinear extension of KDR, the manifold KDR (mKDR) has been proposed (Nilsson et al., 2007). It first learns m-dimensional nonlinear manifold maps $\{\mathbf{t}_i\}_{i=1}^n$ for the input data $\{\mathbf{x}_i\}_{i=1}^n$ (e.g., by the Laplacian Eigenmap (Belkin and Niyogi, 2003)). Then KDR is applied to the learned (nonlinear) manifold. Even though it generalizes KDR to the input space that lives in a nonlinear manifold, mKDR introduces a tight coupling between the central subspace and the separately learned input manifold, which restricts its applicability to transductive settings. Like KDR, mKDR still involves a non-convex optimization, potentially suffering from existence of local minima.

2.2 Inverse Regression

Inverse Regression (IR) is another interesting framework for DRR. The following theorem (Li, 1991) plays a crucial role in the IR framework.

Theorem 2. *If (i) there exists a q-dim central subspace with bases $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_q]$, i.e., $\mathbf{y} \perp \mathbf{x} | \mathbf{B}^\top \mathbf{x}$, and (ii) for any $\mathbf{a} \in \mathbb{R}^p$, $\mathbb{E}[\mathbf{a}^\top \mathbf{x} | \mathbf{B}^\top \mathbf{x}]$ is linear in $\{\mathbf{b}_l^\top \mathbf{x}\}_{l=1}^q$, then $\mathbb{E}[\mathbf{x}|\mathbf{y}]$ lie on the subspace spanned by $\{\mathbf{\Sigma_{xx}} \mathbf{b}_l\}_{l=1}^q$, where $\mathbf{\Sigma_{xx}}$ is the covariance of $\mathbf{x}$.*

From Thm.2, $\mathbf{B}$ can be obtained from q principal directions of $\mathbb{E}[\mathbf{x}|\mathbf{y}]$. That is, the column vectors of $\mathbf{B}$ coincide with the q largest eigenvectors of $\mathbb{V}(\mathbb{E}[\mathbf{x}|\mathbf{y}])$, pre-multiplied by $\mathbf{\Sigma}_{\mathbf{xx}}^{-1}$. Given the data, (Li, 1991) suggests to slice down (cluster) $\mathbf{y}$ so as to compute the sample estimate of $\mathbb{V}(\mathbb{E}[\mathbf{x}|\mathbf{y}])$, thus named *Sliced Inverse Regression* (SIR). More specifically, after clustering $\{\mathbf{y}_i\}_{i=1}^n$ into J slices, $S_1, \ldots, S_J$, and computing slicewise data means, $\mathbf{m}_j = \frac{1}{|S_j|} \sum_{i \in S_j} \mathbf{x}_i$, to approximate $\mathbb{E}[\mathbf{x}|\mathbf{y} \in S_j]$, the sample estimate is $\sum_j p_j \mathbf{m}_j \mathbf{m}_j^\top$, where $p_j = |S_j|/n$ is the j-th slice proportion.

It is known that the condition (ii) in Thm.2 equivalently imposes an elliptically-symmetric distribution (e.g., a Gaussian) of $\mathbf{x}$. In fact, SIR makes two assumptions: the linearity of the central subspace and the elliptical-symmetry of the marginal distribution of $\mathbf{x}$. These assumptions can be strong in certain situations, leading to SIR's failure if the conditions are not met. To relax the restrictions, (Wu, 2006) has suggested a fairly straightforward kernel extension of SIR (called KSIR) via the RKHS mapping $\mathbf{x} \to \boldsymbol{\phi}(\mathbf{x}) \in \mathcal{H}_\mathbf{x}$.

Letting $\mathbf{\Phi_x} = [\boldsymbol{\phi}(\mathbf{x}_1), \ldots, \boldsymbol{\phi}(\mathbf{x}_n)]$ and $\mathbf{C}$ be the $(n \times J)$ 0/1 cluster indicator matrix whose i-th row has all 0's but 1 at the j-th position for $i \in S_j$, KSIR estimates the central subspace comprised of the basis functions:

$$\mathbf{b} = \mathbf{\Phi_x} \boldsymbol{\beta}, \tag{2}$$

where $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_n]^\top$ is the solution[2] to:

$$\frac{1}{n^2} \mathbf{C} \mathbf{P}^{-1} \mathbf{C}^\top \mathbf{K}_\mathbf{x}^2 \boldsymbol{\beta} = \lambda \mathbf{K_x} \boldsymbol{\beta}. \tag{3}$$

Here, $\mathbf{K_x} = \mathbf{\Phi}_\mathbf{x}^\top \mathbf{\Phi_x}$ is the $(n \times n)$ Gram matrix, and $\mathbf{P} = \text{diag}(p_1, \ldots, p_J)$ is the $(J \times J)$ diagonal matrix whose entries are the column sums of $\mathbf{C}$, the proportions $p_j = n_j/n$ of points in S_j $(n_j = |S_j|)$.

KSIR allows a nonlinear central subspace with less restriction on the marginal distribution for $\mathbf{x}$. However, KSIR's slicing-based estimation of $\mathbb{V}(\mathbb{E}[\boldsymbol{\phi}(\mathbf{x})|\mathbf{y}])$ would be unreliable for high-dim $\mathbf{y}$, which restricts KSIR to single-output $(d = 1)$ regression or classification settings only (Wu, 2006). Below, we look into an alternative estimation method that avoids slicing by exploiting the kernel matrices of both input and output.

3 Covariance Operator IR (COIR)

COIR (Kim and Pavlovic, 2008) avoids explicit output space slicing by estimating $\mathbb{V}(\mathbb{E}[\boldsymbol{\phi}(\mathbf{x})|\mathbf{y}])$ using the covariance operator theorems (Baker, 1973; Fukumizu et al., 2004). Let $\mathbf{\Sigma_{xx}}$, $\mathbf{\Sigma_{yy}}$, $\mathbf{\Sigma_{xy}}$, and $\mathbf{\Sigma_{yx}}$ be the *covariance operators*[3] in and between the corresponding RKHSes of $\mathbf{x}$ and $\mathbf{y}$. (Kim and Pavlovic, 2008) showed that the covariance of the inverse regression $\mathbb{V}(\mathbb{E}[\boldsymbol{\phi}(\mathbf{x})|\mathbf{y}])$ can be expressed as:

$$\mathbb{V}(\mathbb{E}[\boldsymbol{\phi}(\mathbf{x})|\mathbf{y}]) = \mathbf{\Sigma_{xy}} \mathbf{\Sigma}_\mathbf{yy}^{-1} \mathbf{\Sigma_{yx}}. \tag{4}$$

Given the data $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$, the sample estimate of Eq.(4), can be written as $\widehat{\mathbb{V}}(\mathbb{E}[\boldsymbol{\phi}(\mathbf{x})|\mathbf{y}]) = \widehat{\mathbf{\Sigma}}_\mathbf{xy} \widehat{\mathbf{\Sigma}}_\mathbf{yy}^{-1} \widehat{\mathbf{\Sigma}}_\mathbf{yx}$, The sample covariance *operators* $(\widehat{\mathbf{\Sigma}})$ can be estimated in a similar manner as the sample covariance *matrices.* E.g., $\widehat{\mathbf{\Sigma}}_\mathbf{xy} = \frac{1}{n} \mathbf{\Phi_x} \mathbf{\Phi}_\mathbf{y}^\top$, where $\mathbf{\Phi_x} = [\boldsymbol{\phi}(\mathbf{x}_1), \ldots, \boldsymbol{\phi}(\mathbf{x}_n)]$ and $\mathbf{\Phi_y} = [\boldsymbol{\phi}(\mathbf{y}_1), \ldots, \boldsymbol{\phi}(\mathbf{y}_n)]$. Then $\widehat{\mathbb{V}}(\mathbb{E}[\boldsymbol{\phi}(\mathbf{x})|\mathbf{y}])$ is:

$$\begin{aligned}
&(\frac{1}{n} \mathbf{\Phi_x} \mathbf{\Phi}_\mathbf{y}^\top)(\frac{1}{n}(\mathbf{\Phi_y} \mathbf{\Phi}_\mathbf{y}^\top + n\epsilon \mathbf{I}))^{-1}(\frac{1}{n} \mathbf{\Phi_y} \mathbf{\Phi}_\mathbf{x}^\top) \\
&= \frac{1}{n} \mathbf{\Phi_x} \mathbf{\Phi}_\mathbf{y}^\top \mathbf{\Phi_y} (\mathbf{\Phi}_\mathbf{y}^\top \mathbf{\Phi_y} + n\epsilon \mathbf{I}_n)^{-1} \mathbf{\Phi}_\mathbf{x}^\top \\
&= \frac{1}{n} \mathbf{\Phi_x} \mathbf{K_y} (\mathbf{K_y} + n\epsilon \mathbf{I}_n)^{-1} \mathbf{\Phi}_\mathbf{x}^\top.
\end{aligned}$$

Here $\mathbf{I}$ is the $(dim(\mathcal{H}_\mathbf{y}) \times dim(\mathcal{H}_\mathbf{y}))$ identity operator, and $\mathbf{I}_n$ is the $(n \times n)$ identity matrix. A small positive

[2] We take q largest eigenvectors for q basis functions.

[3] The RKHS extension of the covariance matrix. For instance, $\mathbf{\Sigma_{yx}}$ is defined as follows: For $\forall \mathbf{g} \in \mathcal{H}_\mathbf{y}$ and $\forall \mathbf{f} \in \mathcal{H}_\mathbf{x}$, $\langle \mathbf{g}, \mathbf{\Sigma_{yx}} \mathbf{f} \rangle = \mathbb{E}[(\mathbf{g}(\mathbf{y}) - \mathbb{E}\mathbf{g}(\mathbf{y}))(\mathbf{f}(\mathbf{x}) - \mathbb{E}\mathbf{f}(\mathbf{x}))]$.

ϵ was added to the diagonal entries of $\mathbf{\Phi_y \Phi_y^\top}$ to circumvent potential rank deficiency in estimating $\mathbf{\Sigma_{yy}}$ and its inverse. ϵ plays an important role as a kernel regularizer in smoothing the affinity structure of $\mathbf{y}$. Finally, using the kernel trick similar to that of kernel PCA (Schölkopf et al., 1998), it is easy to show that finding $\boldsymbol{\beta}$ corresponds to solving the eigensystem,

$$\frac{1}{n}\mathbf{K_y}(\mathbf{K_y} + n\epsilon\mathbf{I}_n)^{-1}\mathbf{K_x^2}\boldsymbol{\beta} = \lambda \mathbf{K_x}\boldsymbol{\beta}. \tag{5}$$

Given $\boldsymbol{\beta}$'s, COIR's central subspace basis functions can be obtained from Eq.(2).

As a consequence, COIR has a closed-form solution (Eq.(5)) to the nonlinear central subspace, and makes few assumptions on the input distribution due to the nonlinear RKHS feature mapping. It also removes a potential risk of being caught at locally optimal solutions, the critical drawback of KDR/mKDR. In particular, COIR is a general case of KSIR (see (Kim and Pavlovic, 2008) and Sec. 4.1) where the explicit slicing is incorporated in a smooth output kernel. This makes COIR not only handle high dimensional output reliably, but also robust to potential noise in the output data. We next discuss important theoretical results that relate COIR to other supervised dimension reduction methods.

4 Relation to SIR, CCA, and LDA

In this section, we show that COIR generalizes the slice-driven IR techniques SIR/KSIR. Then we further investigate a relationship between COIR and other supervised dimensionality reduction techniques, namely CCA (canonical correlation analysis) and LDA (linear discriminant analysis). Despite apparent difference in motivating tasks, COIR's notion of SDR (sufficiency in dimension reduction) shares the main intuition with CCA which aims at dimensionality reduction based on input/output correlation. Indeed, we prove that COIR and the kernelized CCA (KCCA) are equivalent, yielding the same central subspace. Furthermore, we show that both COIR and KCCA can be identically derived from a generalization of kernelized LDA (KLDA) which extends traditional LDA's discrete (class) target space to a real multivariate domain.

4.1 KSIR as a special case of COIR

The equivalence between KSIR (Eq.(3)) and COIR (Eq.(5)) can be made, as in (Kim and Pavlovic, 2008), by setting:

$$\mathbf{K_y}(\mathbf{K_y} + n\epsilon\mathbf{I}_n)^{-1} = \frac{1}{n}\mathbf{C}\mathbf{P}^{-1}\mathbf{C}^\top. \tag{6}$$

Consider an ideal case where the output data $\{\mathbf{y}_i\}_{i=1}^n$ are collapsed to J distinct points that are infinitely far apart from one another[4]. We show that under this ideal case, Eq.(6) is indeed true when $\epsilon \to 0$.

Assuming an RBF kernel, $\mathbf{K_y}$ becomes a 0/1 block diagonal matrix, namely $\mathbf{K_y} = \text{diag}(\mathbf{E}_{|S_1|}, \ldots, \mathbf{E}_{|S_J|})$, where $\mathbf{E}_m$ denotes the $(m \times m)$ matrix with all 1's. Then it is easy to see that $\mathbf{K_y}(\mathbf{K_y} + n\epsilon\mathbf{I}_n)^{-1} = \text{diag}(c_1\mathbf{E}_{|S_1|}, \ldots, c_J\mathbf{E}_{|S_J|})$, where $c_j = \frac{1}{|S_j|+n\epsilon}$. Also, $\frac{1}{n}\mathbf{C}\mathbf{P}^{-1}\mathbf{C}^\top = \text{diag}(\frac{1}{np_1}\mathbf{E}_{|S_1|}, \ldots, \frac{1}{np_J}\mathbf{E}_{|S_J|})$, which reduces Eq.(6) to:

$$|S_j| + n\epsilon = np_j, \quad \text{for } j = 1, \ldots, J. \tag{7}$$

As $\epsilon \to 0$, Eq.(7) implies that $p_j = |S_j|/n$, which is exactly the maximum likelihood (ML) estimate of the cluster proportion employed by KSIR. That is, KSIR is a special case of COIR having 0/1 Gram matrix $\mathbf{K_y}$ (from the assumed J-collapsed perfect clustering) with a vanishing ϵ. For a non-negligible ϵ, the equivalence turns into $p_j = |S_j|/n + \epsilon$, where ϵ now serves as a regularizer (or a smoothing prior) in the ML estimation. For a general (non-0/1) kernel matrix $\mathbf{K_y}$, COIR can be naturally viewed as a smoothed extension of KSIR. Hence, COIR exploits the kernel structure of the output space through an effective use of covariance operators in RKHS, where ϵ acts as a kernel regularizer.

4.2 COIR and Kernel CCA

For two random vectors $\mathbf{x} \in \mathbb{R}^p$ and $\mathbf{y} \in \mathbb{R}^d$, CCA finds embeddings $\mathbf{w} \in \mathbb{R}^p$ and $\mathbf{u} \in \mathbb{R}^d$ such that $\text{Corr}(\mathbf{w}^\top\mathbf{x}, \mathbf{u}^\top\mathbf{y})$ is maximized. This reduces to solving the eigensystem:

$$\mathbf{\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}w} = \lambda\mathbf{\Sigma_{xx}w}. \tag{8}$$

CCA can be easily extended to a nonlinear setting, called the kernel CCA (KCCA) (Hardoon et al., 2004), using feature (Hilbert) space mappings on both input and output, $\mathbf{x} \to \boldsymbol{\phi}(\mathbf{x}) \in \mathcal{H}_\mathbf{x}$ and $\mathbf{y} \to \boldsymbol{\phi}(\mathbf{y}) \in \mathcal{H}_\mathbf{y}$.

From Eq.(4), it is not difficult to see that COIR's central subspace basis functions $\mathbf{b}$ can be obtained from:

$$\mathbf{\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}\Sigma_{xx}b} = \eta\mathbf{\Sigma_{xx}b}. \tag{9}$$

Although the equivalence between KCCA and COIR is not immediately obvious from Eq.(8) and Eq.(9), we prove that they indeed give rise to the same central subspace. We first introduce the following lemma on generalized eigensystems. The proof is rather straightforward using the spectral decomposition theorem, and skipped due to the space limit.

Lemma 3. *For two $(p \times p)$ symmetric PSD matrices $\mathbf{V}$ and $\mathbf{R}$, where $\mathbf{R}$ is invertible, let $\{(\lambda_j, \mathbf{w}_j)\}_{j=1}^p$ be the*

[4] Without loss of generality, we assume the data points are arranged according to their cluster indices.

eigenvalue/vector pairs of the generalized eigensystem $\mathbf{V}\mathbf{w} = \lambda\mathbf{R}\mathbf{w}$. *Then* $\mathbf{V} = \mathbf{R}\mathbf{W}\mathbf{\Lambda}\mathbf{W}^\top\mathbf{R}$, *where* $\mathbf{\Lambda} = diag(\lambda_1, \ldots, \lambda_p)$ *and* $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_p]$.

By denoting $\mathbf{\Sigma_{xy}}\mathbf{\Sigma_{yy}^{-1}}\mathbf{\Sigma_{yx}}$ in the KCCA eigensystem (Eq.(8)) as $\mathbf{V}$ in Lemma 3, we have:

$$\mathbf{V} = \mathbf{\Sigma_{xx}}\mathbf{W}\mathbf{\Lambda}\mathbf{W}^\top\mathbf{\Sigma_{xx}}, \qquad (10)$$

where $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_p]$ has the eigenvectors of Eq.(8) as its column vectors. Similarly, from the COIR eigensystem (Eq.(9)), letting $\mathbf{u} = \mathbf{\Sigma_{xx}}\mathbf{b}$ and $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_p]$ $(= \mathbf{\Sigma_{xx}}\mathbf{B})$ yields:

$$\mathbf{V} = \mathbf{U}\mathbf{H}\mathbf{U}^\top = \mathbf{\Sigma_{xx}}\mathbf{B}\mathbf{H}\mathbf{B}^\top\mathbf{\Sigma_{xx}}, \qquad (11)$$

where $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_p]$ and $\mathbf{H} = \text{diag}(\eta_1, \ldots, \eta_p)$ are the eigenvectors and eigenvalues of Eq.(9), respectively. Hence, the following relationship holds:

$$\mathbf{W}\mathbf{\Lambda}\mathbf{W}^\top = \mathbf{B}\mathbf{H}\mathbf{B}^\top. \qquad (12)$$

Let q_1 and q_2 be the numbers of non-zero eigenvalues for Eq.(8) and Eq.(9), respectively. We denote the non-zero eigenvalue/vector pairs by $\{(\lambda_j, \mathbf{w}_j)\}_{j=1}^{q_1}$ for Eq.(8), and $\{(\eta_j, \mathbf{b}_j)\}_{j=1}^{q_2}$ for Eq.(9). Then we will show that $\{\mathbf{w}_1, \ldots, \mathbf{w}_{q_1}\}$ and $\{\mathbf{b}_1, \ldots, \mathbf{b}_{q_2}\}$ span the same (central) subspace (so, $q_1 = q_2$ automatically follows), which would complete the proof.

For any $\mathbf{a} \in \mathbb{R}^p$ orthogonal to $\mathbf{w}_j$ for all $j = 1, \ldots, q_1$, $0 = \mathbf{a}^\top\mathbf{W}\mathbf{\Lambda}\mathbf{W}^\top\mathbf{a} = \mathbf{a}^\top\mathbf{B}\mathbf{H}\mathbf{B}^\top\mathbf{a} = \sum_{j=1}^{q_2}\eta_j(\mathbf{b}_j^\top\mathbf{a})^2$, which implies that $\mathbf{a}$ is also orthogonal to $\mathbf{b}_j$ for all $j = 1, \ldots, q_2$. Since the other direction also trivially holds, they share the same orthogonal subspace and, hence, the same subspace.

4.3 COIR and Generalized Kernel LDA

The Linear Discriminant Analysis (LDA) finds a linear embedding direction $\mathbf{w} \in \mathbb{R}^p$ for the input $\mathbf{x} \in \mathbb{R}^p$ that maximizes the between-class scatter ($\mathbf{S}_B$), and at the same time, minimizes the within-class scatter ($\mathbf{S}_W$), that is, $\max_{\mathbf{w}} \frac{\mathbf{w}^\top\mathbf{S}_B\mathbf{w}}{\mathbf{w}^\top\mathbf{S}_W\mathbf{w}}$. This reduces to solving the generalized eigensystem: $\mathbf{S}_B\mathbf{w} = \lambda\mathbf{S}_W\mathbf{w}$.

Although LDA assumes a discrete class label $\mathbf{y}$, it is possible to extend it to a real-multivariate label (Barker and Rayens, 2003). The extension generalizes the notion of between/within-class scatter, namely $\mathbf{S}_B = \mathbb{V}(\mathbb{E}[\mathbf{x}|\mathbf{y}])$ and $\mathbf{S}_W = \mathbb{E}[\mathbb{V}(\mathbf{x}|\mathbf{y})]$. We denote this the *generalized LDA*. It is easy to see that the standard LDA (with discrete labels) is a special case since $\mathbb{V}(\mathbb{E}[\mathbf{x}|\mathbf{y}]) = \sum_c \frac{n_c}{n}(\overline{\mathbf{x}}_c - \overline{\mathbf{x}})(\overline{\mathbf{x}}_c - \overline{\mathbf{x}})^\top$ and $\mathbb{E}[\mathbb{V}(\mathbf{x}|\mathbf{y})] = \frac{1}{n}\sum_c\sum_{i\in c}(\mathbf{x}_i - \overline{\mathbf{x}}_c)(\mathbf{x}_i - \overline{\mathbf{x}}_c)^\top$, where n (n_c) and $\overline{\mathbf{x}}$ ($\overline{\mathbf{x}}_c$) are the (class-wise) cardinality and data mean, respectively. The kernelization of the generalized LDA (KLDA) is also fairly straightforward. Using the *E-V-V-E* identity, the generalized KLDA entails the eigensystem:

$$\mathbb{V}(\mathbb{E}[\boldsymbol{\phi}(\mathbf{x})|\mathbf{y}])\mathbf{w} = \lambda\mathbb{V}(\boldsymbol{\phi}(\mathbf{x}))\mathbf{w}, \qquad (13)$$

which simply turns into Eq.(8), meaning that the generalized KLDA is equivalent to KCCA (and COIR).

5 Semi-Supervised Extension of COIR

In the semi-supervised setting, we are given the labeled data $L = \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_l, \mathbf{y}_l)\}$ and the unlabeled data $U = \{\mathbf{x}_{l+1}, \ldots, \mathbf{x}_n\}$. The unlabeled data can be exploited to estimate the unknown entries of the output kernel matrix, $\mathbf{K_y}(i,j)$ for $i \in \{l+1, \ldots, n\}$ and/or $j \in \{l+1, \ldots, n\}$. Once we have $\mathbf{K_y}$, we can readily find the COIR central subspace from Eq.(5).

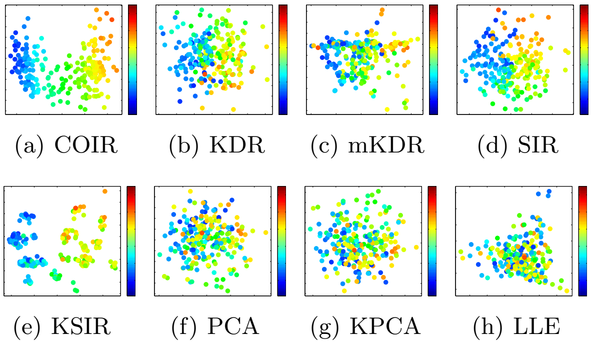
We extend the manifold regularization (Zhu et al., 2003; Belkin et al., 2005), a semi-supervised regression framework that propagates labels along the manifold whose structure is discovered from both labeled and unlabeled data points. Admitting an RKHS mapping in the $\mathbf{y}$ space (i.e., $\mathbf{y} \to \boldsymbol{\phi}(\mathbf{y}) \in \mathcal{H}_\mathbf{y}$), we minimize the kernel-weighted $L2$ difference in the output feature space, namely $\text{tr}(\mathbf{K_y}\boldsymbol{\mathcal{L}}_\mathbf{x}) = \frac{1}{2}\sum_{i,j}k_\mathbf{x}(\mathbf{x}_i, \mathbf{x}_j)||\boldsymbol{\phi}(\mathbf{y}_i) - \boldsymbol{\phi}(\mathbf{y}_j)||^2_{\mathcal{H}_\mathbf{y}}$, an objective similar to the one used in manifold regularization. Here, $\boldsymbol{\mathcal{L}}_\mathbf{x}$ is the graph Laplacian of the input data (e.g., $\boldsymbol{\mathcal{L}}_\mathbf{x} = \mathbf{D_x} - \mathbf{K_x}$, where $\mathbf{D_x}$ is the diagonal matrix having row sums of $\mathbf{K_x}$ in its entries). We estimate $\mathbf{K_y}$ by solving:

$$\begin{aligned} \min_{\mathbf{K_y}\succeq 0} \quad & \text{tr}(\mathbf{K_y}\boldsymbol{\mathcal{L}}_\mathbf{x}), \\ \text{s.t.} \quad & \mathbf{K_y}(i,j) = k_\mathbf{y}(\mathbf{y}_i, \mathbf{y}_j), \ 1 \le i,j \le l. \end{aligned} \qquad (14)$$

Eq.(14) is an instance of semidefinite programs that can be solved by general SDP solvers. Note that the manifold regularization is a special case of Eq.(14) with a linear output kernel $\mathbf{K_y} = \mathbf{Y}\mathbf{Y}^\top$, where we generalize it to an arbitrary nonlinear output kernel.

6 Evaluation

We evaluate the performance of different DRR methods on synthetic and real data for both fully- and semi-supervised settings. We highlight COIR's reliability in estimating a central subspace compared to the previous IR techniques based on output slicing (i.e., SIR/KSIR). We also contrast it with the non-convex KDR/mKDR. We finally show how the semi-supervised COIR (denoted by SS-COIR) discussed in Sec. 5 can benefit from unlabeled data when data are partially labeled. For baseline comparison, we often demonstrate the results of unsupervised dimension reduction techniques such as PCA, kernel PCA (KPCA), and LLE (Roweis and Saul, 2000).



(a) COIR (b) KDR (c) mKDR (d) SIR

(e) KSIR (f) PCA (g) KPCA (h) LLE

Figure 1: 2D subspace embeddings for Noisy Curves.

Unless stated otherwise, the kernel-based methods (i.e., KDR, mKDR, KSIR, and COIR) employ the RBF kernel. In mKDR, the (Laplacian Eigenmap) manifold map of a test point is estimated by a nearest neighbor search in train data. We use k-means for output slicing in SIR/KSIR. Model parameters such as the RBF kernel width, #slices in (K)SIR, and the manifold dimension in mKDR, are estimated by cross validation.

6.1 Synthetic Noisy Curves

The dataset called *curves* (Wu, 2006) is generated from the equation, $y = \mathrm{sign}(\mathbf{b}_1^\top \mathbf{x} + \epsilon_1) \cdot \log(|\mathbf{b}_2^\top \mathbf{x} + a_0 + \epsilon_2|)$ for some $\mathbf{b}_1, \mathbf{b}_2 \in \mathbb{R}^{15}$, $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_{15})$, $\epsilon_1, \epsilon_2 \sim \mathcal{N}(0, 1)$, and a constant a_0. The input is 15-dim, but the central subspace is at most 2-dim as y is decided by $\{\mathbf{b}_l^\top \mathbf{x}\}_{l=1}^2$. To simulate the noisy nature of real-world data, we additionally introduce 4 Gaussian white noise dimensions to the label y (5-dim in total). The task is to reduce the 15-dim $\mathbf{x}$ to 2-dim.

Fig. 1 shows the dimension-reduced input spaces estimated by the competing methods. To visualize the goodness of data layout, each point is colored by its true (noise-removed) y value: higher as warmer (reddish) and lower as cooler (bluish). Note that the adjoining points in highly different colors would significantly increase uncertainty in estimating predictors based on the dimension-reduced input data. Although two linear embeddings, KDR and SIR, yield similar data layouts that look well separated, some points with highly different output values (blue and red) adjoin one another too closely. In mKDR, data points are overall intermingled, probably due to the manifold learning on the isotropic Gaussian input data that can cause severe information loss for regression. In KSIR, we see several separated clusters, each of which contains data points mixed with different output values. This can be attributed to the slicing (clustering) error due to the noise in the output. On the other hand, COIR lays out data along the output values smoothly and discriminatively from blue/left to red/right. The unsupervised PCA, KPCA, and LLE produce random clutters since they ignore labels and simply project the isotropic Gaussian data onto a 2D plane.

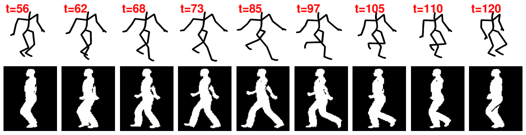


Figure 2: Example 3D human body poses (depicted as skeletons) and silhouette images for walking motion.

Table 1: Test errors for human body pose estimation.

Input Space	COIR	mKDR (KDR)	KSIR (SIR)	KPCA	$\mathbf{x}$ itself
NN Regr.	6.178	8.068 (8.277)	8.351 (8.496)	8.659	6.515
GP Regr.	5.863	7.204 (7.456)	7.311 (7.554)	8.083	5.954

6.2 Human Body Pose Estimation

We consider a regression problem to estimate the human body pose from a silhouette image. The task is particularly interesting for the supervised dimension reduction techniques as we may expect to find a more accurate intrinsic low-dim subspace of the human figure, guided by the pose information. We use the sequence of walking motion (about 3 walking cycles) obtained from the CMU motion capture database[5]. The output $\mathbf{y}$ is composed of 59 3D joint angles at 31 articulation points of the body. The input $\mathbf{x}$ is the silhouette image of size (160×100), i.e., a 16000-dim vector, taken at a side view (Fig. 2).

The first 80% frames of the sequence are used for training, and the rest for testing. The central subspace dimension is set to 2, the widely believed manifold dimension for the walking motion. Once the subspaces are learned, we conduct regression estimation using the dimension-reduced data as input. We employ two most popular regression methods: the nearest neighbor (NN) and the Gaussian Process (GP) regression (Williams and Rasmussen, 1996). Table 1 shows the test errors. Unlike other approaches, COIR rarely entails information loss in terms of pose prediction (the performance is even slightly better than that based on the silhouette image $\mathbf{x}$ itself as input), while achieving significant reduction of the data dimensionality.

6.3 Scratched USPS Digit Image Denoising

To test the behavior of COIR on high-dim output data we devise an image denoising experiment with the USPS hand-written digit images (LeCun et al.,

[5] http://mocap.cs.cmu.edu.

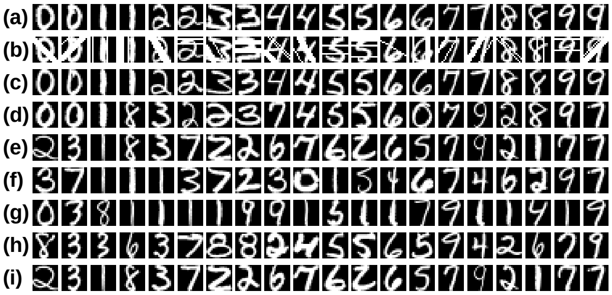


Figure 3: NN prediction examples for USPS test images. (a) Noise-free target images ($\mathbf{y}$), (b) Scratched input images ($\mathbf{x}$), (c) NN prediction on COIR subspace, (d) KDR, (e) mKDR, (f) SIR, (g) KSIR, (h) LLE, and (i) NN on $\mathbf{x}$ itself.

Table 2: Test errors for USPS image denoising.

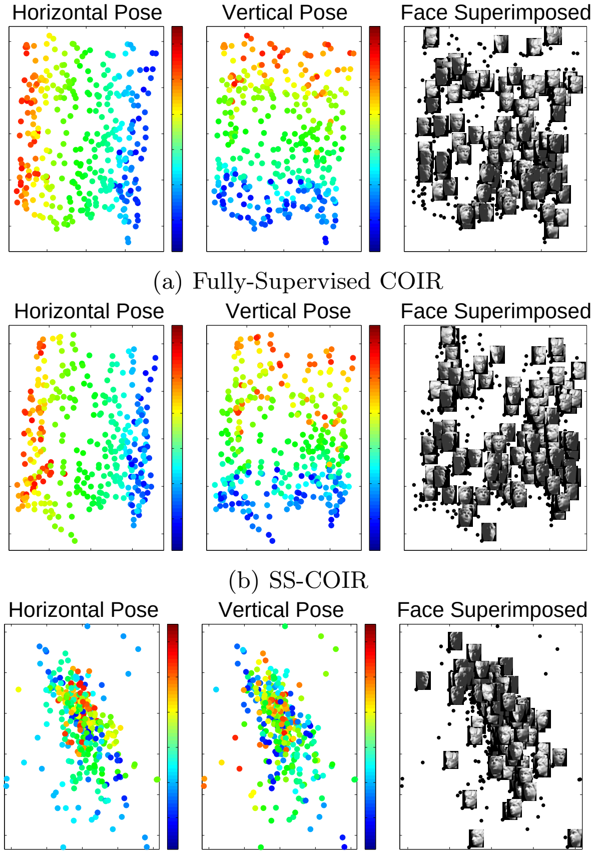
Input Space	COIR	mKDR (KDR)	KSIR (SIR)	LLE	$\mathbf{x}$ itself
NN Regr.	8.533	9.375 (9.133)	11.491 (10.962)	11.275	9.361
GP Regr.	8.145	9.132 (9.031)	10.726 (10.614)	10.792	9.104

1989). By adding random scratch lines with varying thickness and orientation on the normalized (16 × 16) digit images, the task is to denoise the corrupted images. The regression problem is to predict the original unscratched image (output $\mathbf{y}$) from the scratched input ($\mathbf{x}$). Both $\mathbf{y}$ and $\mathbf{x}$ are of 256-dim.

From the USPS database, we use 2000 images for training and other 2000 images for testing. The central subspace dimension is chosen as 30. The test reconstruction (denoising) RMS errors are shown in Table 2, while some of the denoised test images by the NN regression are depicted in Fig. 3. We see that COIR outperforms non-convex KDR/mKDR which can easily get caught at local optima. Moreover, COIR is robust to noise with improved prediction accuracy compared to the regression based on the image input itself. SIR/KSIR again suffer from unreliable slicing-based estimation in the high-dim output space.

6.4 Head Pose Estimation

The dataset[6] consists of about 700 face images of size (64 × 64), rendered from different views with varying lighting direction. The relevant regression task is to predict a 2D pose (horizontal and vertical rotation angles) from a 4096-dim image. We randomly partition the data into train/test sets with equal sizes. Fig. 4(a) shows the projection of test images onto a 2D central subspace estimated by COIR on the *fully-labeled* train data. For visualization, each point is colored by the true label (i.e., horizontal and vertical poses). We see that COIR lays out the data points along the head pose quite obviously, where X and Y axes roughly correspond to horizontal and vertical angles, respectively.



(a) Fully-Supervised COIR

(b) SS-COIR

(c) COIR on 10% Labeled Data Only

Figure 4: COIR subspaces for head pose estimation.

Table 3: Test errors for head pose estimation.

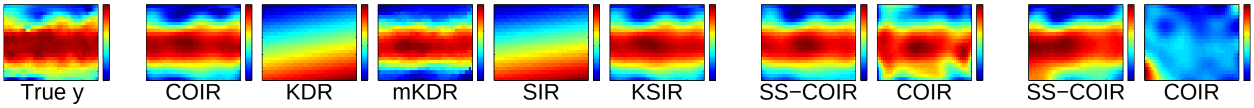
Fully-Supervised			10% Labeled	
COIR	mKDR	KSIR	SS-COIR	COIR
0.263	0.380	0.575	0.418	1.294

Then we form a *semi-supervised* setting by revealing the labels of only 10% of the train data which are randomly chosen. We compare the central subspace estimated by our semi-supervised COIR (SS-COIR) (Fig. 4(b)) with that trained on the labeled data only (Fig. 4(c)). As shown, the semi-supervised COIR discovers a well-discriminated central subspace similar to the fully-labeled case, while using only a small number of labeled data makes it severely mixed. We also estimate a linear regressor using the dimension-reduced input data. The test errors are shown in Table 3.

6.5 Temperature Map

We test on the globe temperature data obtained from the satellite measurement of temperatures in the mid-

[6] http://isomap.stanford.edu/datasets.html.



(a) Data (b) Fully-Supervised (c) 20% Labeled (d) 10% Labeled

Figure 5: (Temperature Map) Prediction by linear regression trained on different embedded spaces.

Table 4: (Temperature Map) Test errors.

Fully-Supervised			Semi-Supervised	
COIR	mKDR (KDR)	KSIR (SIR)	SS-COIR (20%/10%)	COIR (20%/10%)
1.080	2.578 (8.0733)	1.287 (8.0732)	1.145 / 2.417	3.008 / 8.858

dle troposphere (http://www.remss.com/). We take the map of Dec. 2004 (Fig. 5(a)), which was also used in (Nilsson et al., 2007). The map is a (72 × 144) matrix, where each element has a temperature (in K) at its position (latitude, longitude). We consider the regression problem with y = temp. and $\mathbf{x}$ = (latitude,longitude). We randomly split the data into 60%/40% train/test sets. In the fully-supervised setting, as shown in Fig. 5(b), the prediction by linear regression using nonlinear embeddings (mKDR, KSIR, and COIR) is good. On the other hand, the linear embeddings (KDR and SIR) fail due to the nonlinear ellipsoidal manifold structure of the input space. See also Table 4 for the test errors by linear regression.

In the semi-supervised setting, we randomly choose subsets of the train data with two different sizes (20% and 10% of the train data), where only their labels are used. As shown in Fig. 5(c) and 5(d), SS-COIR exhibits prediction results almost similar to the fully-supervised case (distorted a bit for 10% labeled case). However, when we use only the labeled data, the prediction results are much worse. Table 4 quantitatively demonstrates this as well.

7 Conclusion

In this work we presented a unifying view of COIR and other DRR techniques while relating them to CCA and LDA, traditionally formulated in discrete label output spaces. We also derived an extension of COIR to a semi-supervised setting, which allows the family of central subspace regression methods to effectively handle large datasets with partially known targets. Experiments on several synthetic and real datasets compared different related methods, and showed that COIR and its semi-supervised extension outperform competing DRR methods.

Acknowledgements

This work was supported in part by NSF IIS grant #0413105.

References

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

Barker, M. and Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics*, 17(3):166–173.

Belkin, M. and Niyogi, P. (2003). Laplacian Eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396.

Belkin, M., Niyogi, P., and Sindhwani, V. (2005). On manifold regularization. AISTATS.

Cook, R. D. (1998). *Regression graphics*. Wiley Inter-Science.

Fukumizu, K., Bach, F., and Jordan, M. (2004). Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *JMLR*.

Globerson, A. and Roweis, S. (2005). Metric learning by collapsing classes. NIPS.

Hardoon, D., Szedmak, S., and Shawe-Taylor, J. (2004). Canonical correlation analysis; An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664.

Kim, M. and Pavlovic, V. (2008). Dimensionality reduction using covariance operator inverse regression. CVPR.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Handwritten digit recognition with a back-propagation network. NIPS.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *JASA*.

Nilsson, J., Sha, F., and Jordan, M. (2007). Regression on manifolds using kernel dimension reduction. ICML.

Roweis, S. and Saul, L. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326.

Schölkopf, B., Smola, A. J., and Muller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319.

Tenenbaum, J. B., Silva, V. D., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323.

Weinberger, K., Blitzer, J., and Saul, L. (2005). Distance metric learning for large margin nearest neighbor classification. NIPS.

Williams, C. K. I. and Rasmussen, C. E. (1996). Gaussian processes for regression. NIPS.

Wu, H. M. (2006). Kernel sliced inverse regression with applications on classification. ICSA Applied Statistics Symposium.

Zhu, X., Ghahramani, Z., and Lafferty, J. (2003). Semi-supervised learning using Gaussian fields and harmonic functions. ICML.