Regression on Manifolds Using Kernel Dimension Reduction

Jens Nilsson JENSN@MATHS.LTH.SE
Centre for Mathematical Sciences, Lund University, Box 118, SE-221 00 Lund, Sweden

Fei Sha FEISHA@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

Abstract

We study the problem of discovering a manifold that best preserves information relevant to a nonlinear regression. Solving this problem involves extending and uniting two threads of research. On the one hand, the literature on sufficient dimension reduction has focused on methods for finding the best linear subspace for nonlinear regression; we extend this to manifolds. On the other hand, the literature on manifold learning has focused on unsupervised dimensionality reduction; we extend this to the supervised setting. Our approach to solving the problem involves combining the machinery of kernel dimension reduction with Laplacian eigenmaps. Specifically, we optimize cross-covariance operators in kernel feature spaces that are induced by the normalized graph Laplacian. The result is a highly flexible method in which no strong assumptions are made on the regression function or on the distribution of the covariates. We illustrate our methodology on the analysis of global temperature data and image manifolds.

1. Introduction

Dimension reduction is an important theme in machine learning. Dimension reduction problems can be approached from the point of view of either unsupervised learning or supervised learning. A classical example of the former is principal component analysis (PCA), a method that projects data onto a linear manifold. More recent research has focused on nonlinear manifolds, and the long list of "manifold learning" algorithms—including LLE, Isomap, and Laplacian eigenmaps—provide sophisticated examples of unsupervised dimension reduction (Tenenbaum et al., 2000; Roweis & Saul, 2000; Belkin & Niyogi, 2003; Donoho & Grimes, 2003). The supervised learning setting is somewhat more involved; one must make a choice of the family of manifolds to represent the covariate vectors, and one must also choose a family of functions to represent the regression surface (or classification boundary). Due in part to this additional complexity, most of the focus in the supervised setting has been on reduction to linear manifolds. This is true of classical linear discriminant analysis, and also of the large family of methods known as sufficient dimension reduction (SDR) (Li, 1991; Cook, 1998; Fukumizu et al., 2006). SDR aims to find a linear subspace $\mathcal{S}$ such that the response Y is conditionally independent of the covariate vector X, given the projection of X on $\mathcal{S}$. This formulation in terms of conditional independence means that essentially no assumptions are made on the form of the regression from X to Y, but strong assumptions are made on the manifold representation of X (it is a linear manifold). Finally, note that the large literature on feature selection for supervised learning can also be conceived of as a projection onto a family of linear manifolds.

It is obviously of interest to consider methods that combine manifold learning and sufficient dimension reduction. From the point of view of manifold learning, we can readily imagine situations in which some form of side information is available to help guide the choice of manifold. Such side information might come from a human user in an exploratory data analysis setting. We can also envisage regression and classification problems in which nonlinear representations of the covariate vectors are natural on subject matter grounds. For example, we will consider a problem involving atmosphere temperature close to the Earth's surface in which a manifold representation of the covariate vectors is quite natural. We will also consider an exam-

Appearing in *Proceedings of the 24th International Conference on Machine Learning*, Corvallis, OR, 2007.

ple involving image manifolds. Other examples involving dynamical systems are readily envisaged; for example, a torus is often a natural representation of robot kinematics and robot dynamics can be viewed as a regression on this manifold.

It is obviously an ambitious undertaking to attempt to identify a nonlinear manifold from a large family of nonlinear manifolds while making few assumptions regarding the regression surface. Nonetheless, it is an important undertaking, because the limitation to linear manifolds in SDR can be quite restrictive in practice, and the lack of a role for supervised data in manifold learning is limiting. As in much of the unsupervised manifold learning literature, we aim to make progress on this problem by focusing on visualization. Without attempting to define the problem formally, we attempt to study situations in which supervised manifold learning is natural and investigate the ability of algorithms to find useful visualizations.

The methodology that we develop combines techniques from SDR and unsupervised manifold learning. Specifically, we make use of ideas from kernel dimension reduction (KDR), a recently-developed approach to SDR that uses cross-covariance operators on reproducing kernel Hilbert spaces to measure quantities related to conditional independence. We will show that this approach combines naturally with representations of manifolds based on Laplacian eigenmaps.

The paper is organized as follows. In Section 2, we provide basic background on SDR, KDR and unsupervised manifold learning. Section 3 presents our new manifold kernel dimension reduction (mKDR) method. In Section 4, we present experimental results evaluating mKDR on both synthetic and real data sets. In Section 5, we comment briefly on related work. Finally, we conclude and discuss future directions in Section 6.

2. Background

We begin by outlining the SDR problem. We then describe KDR, a specific methodology for SDR in which the linear subspace is characterized by cross-covariance operators on reproducing kernel Hilbert spaces. Finally, we also provide a brief overview of unsupervised manifold learning.

2.1. Sufficient Dimension Reduction

Let $(\mathcal{X}, \mathcal{B}_\mathcal{X})$ and $(\mathcal{Y}, \mathcal{B}_\mathcal{Y})$ be measurable spaces of covariates X and response variables Y respectively. SDR aims at finding a linear subspace $\mathcal{S} \subset \mathcal{X}$ such that $\mathcal{S}$ contains as much predictive information regarding the response Y as the original covariate space. This desideratum is captured formally as a conditional independence assertion:

$$Y \perp\!\!\!\perp X \mid \boldsymbol{B}^{\mathrm{T}} X \tag{1}$$

where $\boldsymbol{B}$ denotes the orthogonal projection of $\mathcal{X}$ onto $\mathcal{S}$. The subspace $\mathcal{S}$ is referred to as a *dimension reduction subspace*. Dimension reduction subspaces are not unique. We can derive a unique "minimal" subspace, defined as the intersections of all reduction subspaces $\mathcal{S}$. This minimal subspace does not necessarily satisfy the conditional independence assertion; when it does, the subspace is referred to as the *central subspace*.

Many approaches have been developed to identify central subspaces (Li, 1991; Li, 1992; Cook & Li, 1991; Cook & Yin, 2001; Chiaromonte & Cook, 2002; Li et al., 2005). Many of these approaches are based on inverse regression; that is, the problem of estimating $\mathbf{E}[\![X|Y]\!]$. The intuition is that, if the forward regression model $\mathbf{P}(Y|X)$ is concentrated in a subspace of $\mathcal{X}$ then $\mathbf{E}[\![X|Y]\!]$ should lie in the same subspace. Moreover, the responses Y are typically of much lower dimension than the covariates X, and thus the subspace may be more readily identified via inverse regression. A difficulty with this approach, however, is that rather strong assumptions generally have to be imposed on the distribution of X (e.g., that the distribution be elliptical), and the methods can fail when these assumptions are not met. This issue is of particular importance in our setting, in which the focus is on capturing the structure underlying the distribution of X and in which such strong assumptions would be a significant drawback. We thus turn to a description of KDR, an approach to SDR that does not make such strong assumptions.

2.2. Kernel Dimension Reduction

The framework of kernel dimension reduction was first described in Fukumizu et al. (2004) and later refined in Fukumizu et al. (2006). The key idea of KDR is to map random variables X and Y to reproducing kernel Hilbert spaces (RKHS) and to characterize conditional independence using cross-covariance operators.

Let $\mathcal{H}_\mathcal{X}$ be an RKHS of functions on $\mathcal{X}$ induced by the kernel function $\mathbf{K}_\mathcal{X}(\cdot, X)$ for $X \in \mathcal{X}$. We define the space $\mathcal{H}_\mathcal{Y}$ and the kernel function $\mathbf{K}_\mathcal{Y}$ similarly. Define the cross-covariance between a pair of functions $f \in \mathcal{H}_\mathcal{X}$ and $g \in \mathcal{H}_\mathcal{Y}$ as follows:

$$\mathbf{C}_{fg} = \mathbf{E}_{XY}\left[\Big(f(X) - \mathbf{E}_X[\![f(X)]\!]\Big)\Big(g(Y) - \mathbf{E}_Y[\![g(Y)]\!]\Big)\right]. \tag{2}$$

It turns out that there exists a bilinear operator $\boldsymbol{\Sigma}_{YX}$ from $\mathcal{H}_\mathcal{X}$ to $\mathcal{H}_\mathcal{Y}$ such that $\mathbf{C}_{fg} = \langle g, \boldsymbol{\Sigma}_{YX} f\rangle_{\mathcal{H}_\mathcal{Y}}$ for *all* functions f and g. Similarly we can define covariance operators $\boldsymbol{\Sigma}_{XX}$ and $\boldsymbol{\Sigma}_{YY}$. Finally, we can use these operators to define a class of *conditional* cross-covariance operators in the fol-

lowing way:

$$\boldsymbol{\Sigma}_{YY|X} = \boldsymbol{\Sigma}_{YY} - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}. \quad (3)$$

This definition assumes that $\boldsymbol{\Sigma}_{XX}$ is invertible; more general cases are discussed in Fukumizu et al. (2006).

Note that the conditional covariance operator $\boldsymbol{\Sigma}_{YY|X}$ of eq. (3) is "less" than the covariance operator $\boldsymbol{\Sigma}_{YY}$, as the difference $\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}$ is positive semidefinite. This agrees with the intuition that conditioning reduces uncertainty. We gain further insight by noting the similarity between eq. (3) and the covariance matrix of the conditional distribution $\mathbf{P}(Y|X)$ when X and Y are jointly Gaussian.

Finally, we are ready to link these cross-covariance operators to the central subspace. Consider any subspace $\mathcal{S}$ in $\mathcal{X}$. Let us map this subspace to an RKHS $\mathcal{H}_{\mathcal{S}}$ with a kernel function $\mathbf{K}_{\mathcal{S}}$. Furthermore, let us define the conditional covariance operator $\boldsymbol{\Sigma}_{YY|S}$ as if we would regress Y on $\mathcal{S}$. What is the relation between $\boldsymbol{\Sigma}_{YY|S}$ and $\boldsymbol{\Sigma}_{YY|X}$?

Intuitively, $\boldsymbol{\Sigma}_{YY|S}$ would indicate greater residual error of predicting Y unless $\mathcal{S}$ contains the central subspace. This intuition was formalized in Fukumizu et al. (2006):

Theorem 1 *Suppose $Z = \boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}X \in \mathcal{S}$ where $\boldsymbol{B} \in \mathbb{R}^{D\times d}$ is a projection matrix such that $\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}$ is an identity matrix. Further, assume Gaussian RBF kernels for $\mathbf{K}_X$, $\mathbf{K}_{\mathcal{Y}}$ and $\mathbf{K}_{\mathcal{S}}$. Then*

- *$\boldsymbol{\Sigma}_{YY|X} \prec \boldsymbol{\Sigma}_{YY|Z}$, where $\prec$ stands for "less than or equal to" in some operator partial ordering.*
- *$\boldsymbol{\Sigma}_{YY|X} = \boldsymbol{\Sigma}_{YY|Z}$ if and only if $Y \perp\!\!\!\perp X|\boldsymbol{B}^{\mathrm{T}}X$, that is, $\mathcal{S}$ is the central subspace.*

Note that this theorem does not impose assumptions on either the marginal distributions of X and Y or the conditional distribution $\mathbf{P}(Y|X)$. Note also that although we have stated the theorem using Gaussian kernels, other kernels are possible and general conditions on these are discussed in Fukumizu et al. (2006).

Theorem 1 leads to an algorithm for estimating the central subspace, characterized by $\boldsymbol{B}$, for empirical samples. Using the trace to order operators, $\boldsymbol{B}$ is the matrix that minimizes $\mathrm{Tr}\,[\![\hat{\boldsymbol{\Sigma}}_{YY|Z}]\!]$, where $\hat{\boldsymbol{\Sigma}}_{YY|Z}$ is the empirical version of the conditional covariance operator (3). Let $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^N$ denote N samples from the joint distribution $\mathbf{P}(X, Y)$, and let $\boldsymbol{K}_Y \in \mathbb{R}^{N\times N}$ and $\boldsymbol{K}_Z \in \mathbb{R}^{N\times N}$ denote the Gram matrices computed over $\{\boldsymbol{y}_i\}$ and $\{\boldsymbol{z}_i = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{x}_i\}$. Fukumizu et al. (2006) show that this minimization problem can be formulated in terms of $\boldsymbol{K}_Y$ and $\boldsymbol{K}_Z$, so that $\boldsymbol{B}$ is the solution to:

$$\begin{array}{ll} \min & \mathrm{Tr}\,[\![\boldsymbol{K}_Y^c(\boldsymbol{K}_Z^c + N\epsilon\boldsymbol{I})^{-1}]\!] \\ \text{such that} & \boldsymbol{B}^{\mathrm{T}}\boldsymbol{B} = \boldsymbol{I} \end{array} \quad (4)$$

where $\boldsymbol{I}$ is the identity matrix of appropriate dimensionality, and ϵ a regularization coefficient. The matrix $\boldsymbol{K}^c$ denotes the centered kernel matrices

$$\boldsymbol{K}^c = \left(\boldsymbol{I} - \frac{1}{N}\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}}\right)\boldsymbol{K}\left(\boldsymbol{I} - \frac{1}{N}\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}}\right) \quad (5)$$

where $\boldsymbol{e}$ is a vector of all ones.

2.3. Manifold Learning

Many real-world data sets are generated with very few degrees of freedom; an example is pictures of the same object under different imaging conditions, such as rotation angle or translation. The extrinsic dimensionality of these images as data points in the Euclidean space spanned by the pixel intensities far exceeds the intrinsic dimensionality determined by those underlying factors. When these factors vary smoothly, the data points can be seen as lying on a low-dimensional submanifold embedded in the high dimensional ambient space. Discovering such submanifolds and finding low-dimensional representations of them has been a focus of much recent work on unsupervised learning (Roweis & Saul, 2000; Tenenbaum et al., 2000; Belkin & Niyogi, 2003; Donoho & Grimes, 2003; Sha & Saul, 2005). A key theme of these learning algorithms is to preserve (local) topological and geometrical properties (for example, geodesics, proximity, symmetry, angle) while projecting data points to low dimensional representations.

In this section, we briefly review the method of Laplacian eigenmaps, which is the manifold learning method on which we base our extension to supervised manifold learning. This method is based on the (normalized) graph Laplacian which can be seen as a discrete approximation to the Laplace-Beltrami operator on continuous manifolds.

Let $\{\boldsymbol{x}_i \in \mathbb{R}^D\}_{i=1}^N$ denote N data points sampled from a submanifold. We start by constructing a graph which has N vertices, one for each $\boldsymbol{x}_i$. Vertex i and vertex j are linked by an edge if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are nearest neighbors. Let $\boldsymbol{W}$ be the matrix whose element $W_{ij} = \exp(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2/\sigma^2)$ if there is an edge between vertex i and j, and zero otherwise. Furthermore, let $\boldsymbol{D}$ be the diagonal matrix whose diagonal elements are the row sums of $\boldsymbol{W}$; i.e., $D_{ii} = \sum_j W_{ij}$. The aim of the Laplacian eigenmap procedure is to find an $m < D$ dimensional embedding $\{\boldsymbol{u}_i \in \mathbb{R}^m\}$ such that $\boldsymbol{u}_i$ and $\boldsymbol{u}_j$ are close if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are close. Formally, let $\boldsymbol{v}_m \in \mathbb{R}^N$ be column vectors such that $[\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots \boldsymbol{u}_N] = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots \boldsymbol{v}_m]^{\mathrm{T}}$. Then $\boldsymbol{v}_m$ is chosen such that

$$\sum_{ij} \frac{W_{ij}(v_{mi} - v_{mj})^2}{\sqrt{D_{ii}}\sqrt{D_{jj}}} \quad (6)$$

is minimized, subject to the constraint that $\boldsymbol{v}_m$ is orthogonal to $\boldsymbol{v}_{m'}$ if $m \neq m'$. By the Rayleigh quotient theorem, the vector $\boldsymbol{v}_m$ must be the m-th bottom eigenvector of the

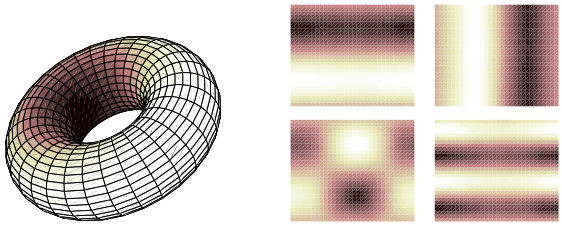


Figure 1. 3-*D* torus and bottom eigenvectors of graph Laplacian for data sampled from the torus. See text for details.

matrix $\boldsymbol{L} = \boldsymbol{D}^{-1/2}(\boldsymbol{D} - \boldsymbol{W})\boldsymbol{D}^{-1/2}$ excluding the very bottom constant eigenvector. The matrix $\boldsymbol{L}$ is referred to as the normalized (and symmetrized) graph Laplacian or graph Laplacian for brevity in the rest of the paper.

The eigenvectors of the graph Laplacian can be seen as discretized versions of harmonic functions (Lafon, 2004; Coifman et al., 2005); i.e., they are eigenfunctions of the continuous Laplace-Beltrami operator sampled at the locations of the $\boldsymbol{x}_i$. As an example, Fig. 1 shows some of the first non-constant eigenvectors (mapped onto 2-*D*) for data points sampled from a 3-*D* torus. The image intensities correspond to the high and low values of the eigenvectors. The variation in the intensities can be interpreted as the high and low frequency components of the harmonic functions. Intuitively, these eigenvectors can be used to approximate smooth functions on the manifold (Bengio et al., 2003; Ham et al., 2004). We will explore this intuition in the next section to parameterize the central subspace with these eigenfunctions.

3. Manifold KDR

Consider regression problems where the covariates have an intrinsic manifold structure. Can we incorporate such geometrical information into the process of identifying a predictive representation for the covariates? To this end, we present an algorithm that we refer to as *manifold kernel dimension reduction* (mKDR). The algorithm combines ideas from unsupervised manifold learning and KDR.

At a high level, the algorithm has three ingredients: (1) the computation of a low-dimensional embedding of covariates X; (2) the parametrization of the central subspace as a linear transformation of the low-dimensional embedding; (3) the computation of the coefficients of the optimal linear map using the KDR framework. The linear map yields directions in the low-dimensional embedding that contribute most significantly to the central subspace. Such directions can be used for data visualization.

Let us choose M eigenvectors $\{\boldsymbol{v}_m\}_{m=1}^{M}$, or equivalently, an M-dimensional embedding $\boldsymbol{U} \in \mathcal{U} \subset \mathbb{R}^{M \times N}$. As in the framework of KDR, we consider a kernel function that maps a point $\boldsymbol{B}^{\mathrm{T}}\boldsymbol{x}_i$ in the central subspace to the RKHS. We construct the mapping explicitly:

$$\mathbf{K}(\cdot, \boldsymbol{B}^{\mathrm{T}}\boldsymbol{x}_i) \approx \boldsymbol{\Phi}\boldsymbol{u}_i \tag{7}$$

by approximating it with a linear expansion $\boldsymbol{\Phi} \in \mathbb{R}^{M \times M}$ in the eigenfunctions. Note that the linear map $\boldsymbol{\Phi}$ is independent of $\boldsymbol{x}_i$, enforcing a globally smooth transformation on the embedding $\boldsymbol{U}$.

We identify the linear map $\boldsymbol{\Phi}$ in the framework of KDR. Specifically, we aim to minimize the contrast function of eq. (4) for statistical conditional independence between the response $\boldsymbol{y}_i$ and $\boldsymbol{x}_i$. The Gram matrix is then approximated and parameterized by the linear map $\boldsymbol{\Phi}$:

$$\langle\, \mathbf{K}(\cdot, \boldsymbol{B}^{\mathrm{T}}\boldsymbol{x}_i), \mathbf{K}(\cdot, \boldsymbol{B}^{\mathrm{T}}\boldsymbol{x}_j) \,\rangle \approx \boldsymbol{u}_i{}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{u}_j \tag{8}$$

Finally, we formulate the following optimization problem:

$$\begin{array}{ll} \min & \mathrm{Tr}\left[\!\left[\boldsymbol{K}_Y^c(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{U} + N\epsilon\boldsymbol{I})^{-1}\right]\!\right] \\ \text{such that} & \boldsymbol{\Omega} \geq 0 \\ & \mathrm{Tr}(\boldsymbol{\Omega}) = 1 \end{array} \tag{9}$$

where $\boldsymbol{\Omega} = \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}$. Note that if $\boldsymbol{\Omega}$ is allowed to grow arbitrarily large then the objective function attains an arbitrarily small value with infimum of zero. Therefore, we constrain the matrix $\boldsymbol{\Omega}$ to have unit trace: $\mathrm{Tr}(\boldsymbol{\Omega}) = 1$. Furthermore, the matrix $\boldsymbol{\Omega}$ needs to be constrained in the cone of positive semidefinitive matrices, i.e., $\boldsymbol{\Omega} \geq 0$, so that a linear map $\boldsymbol{\Phi}$ can be computed as the square root of the $\boldsymbol{\Omega}$.

Note that recovering $\boldsymbol{B}$ from the linear map $\boldsymbol{\Phi}$ is possible via inverting the map in eq. (7). However, we would like to point out that for the purpose of regression using reduced dimensionality, it is sufficient to build regression models for $\boldsymbol{Y}$ from the central subspace $\boldsymbol{\Phi}\boldsymbol{U}$ in the embedding space $\mathcal{U}$ generated by the graph Laplacian.

The optimization of eq. (9) is nonlinear and nonconvex. We have applied the projected gradient method to find local optimal minimizers. This method worked well in all of our experiments. Despite the nonconvexity of the problem, initialization with the identity matrix gave fast convergence in our experiments. Algorithm 1 gives the pseudocode for the mKDR algorithm. In our experiments we choose a regularized linear response kernel $\boldsymbol{K}_Y = \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y} + N\epsilon\boldsymbol{I}$, but other choices could also be considered.

The problem of choosing the dimensionality of the central subspace is still open in the case of KDR and is open in our case as well. A heuristic strategy is to choose M to be larger than a conservative prior estimate of the the dimensionality d of the central subspace and to rely on the fact that the solution to the optimization eq. (9) often leads to a low rank linear map $\boldsymbol{\Phi}$. The rank $r = \textsc{rank}(\boldsymbol{\Phi})$ then provides an empirical estimate of d. Let $\boldsymbol{\Phi}^r$ stand for the matrix formed

Algorithm 1 Manifold Kernel Dimension Reduction

Input:
Covariates $\{\boldsymbol{x}_i \in \mathbb{R}^D\}_{i=1}^N$ and responses $\{\boldsymbol{y}_i \in \mathbb{R}^{D'}\}_{i=1}^N$
$M \leq D$: number of eigenvectors
$d \leq M$: the dimensionality of the central subspace
Output: linear map $\boldsymbol{\Phi}$
Compute eigenvectors $\{\boldsymbol{v}_m\}_{m=1}^M$ via graph Laplacian
Initialize $\boldsymbol{\Omega} \leftarrow \boldsymbol{I}$
$t \leftarrow 1$
repeat
Set step size $\eta \leftarrow 1/t$
Update $\boldsymbol{\Omega}$ with gradient descent

$$\boldsymbol{\Omega} \leftarrow \boldsymbol{\Omega} - \eta \frac{\partial \operatorname{Tr}\left[\!\left[\boldsymbol{K}_Y^c (\boldsymbol{U}^{\mathrm{T}} \boldsymbol{\Omega} \boldsymbol{U} + N\epsilon \boldsymbol{I})^{-1}\right]\!\right]}{\partial \boldsymbol{\Omega}} \tag{10}$$

where the gradient is given by

$$-\boldsymbol{U}(\boldsymbol{U}^{\mathrm{T}} \boldsymbol{\Omega} \boldsymbol{U} + N\epsilon \boldsymbol{I})^{-1} \boldsymbol{K}_Y^c (\boldsymbol{U}^{\mathrm{T}} \boldsymbol{\Omega} \boldsymbol{U} + N\epsilon \boldsymbol{I})^{-1} \boldsymbol{U}^{\mathrm{T}}$$

Project $\boldsymbol{\Omega}$ to the positive semidefinite cone

$$\boldsymbol{\Omega} \leftarrow \sum_m \max(\lambda_m, 0) \boldsymbol{a}_m \boldsymbol{a}_m^{\mathrm{T}} \tag{11}$$

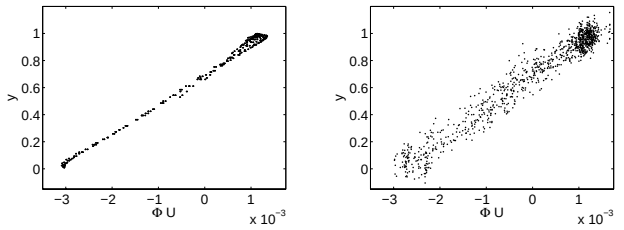
where $(\lambda_m, \boldsymbol{a}_m)$ are $\boldsymbol{\Omega}$'s eigenvalues and eigenvectors.
Scale $\boldsymbol{\Omega}$ to have unit trace: $\boldsymbol{\Omega} \leftarrow \boldsymbol{\Omega} / \operatorname{Tr}(\boldsymbol{\Omega})$
Increase t: $t \leftarrow t + 1$
until $|V(t) - V(t-1)| \,/\, |V(t)| < \texttt{tol}$, where $V(t)$ is the value of the objective function of eq. (9) at step t.
Compute $\boldsymbol{\Phi}$ as the square root of $\boldsymbol{\Omega}$

by the top r eigenvectors of $\boldsymbol{\Phi}$. We can approximate the subspace of $\boldsymbol{\Phi}\boldsymbol{U}$ with $\boldsymbol{\Phi}^r \boldsymbol{U}$.

Additionally, the row vectors of the matrix $\boldsymbol{\Phi}^r$ are combinations of eigenfunctions. We can select the eigenfunctions with the largest coefficients and use them to visualize the original data $\{\boldsymbol{x}_i\}$. Note that these eigenfunctions are not necessarily chosen *consecutively* from the spectral bottom of the graph Laplacian (in contradistinction to the unsupervised manifold learning setting). Indeed, the seemingly out-of-order selection of eigenfunctions implemented by our procedure can be interpreted as the "most predictive" functions that respect the intrinsic geometry of $\{\boldsymbol{x}_i\}$. In our experiments, we have used these strategies to reveal the central subspace as well as to visualize data.

4. Experimental Results

We demonstrate the effectiveness of the mKDR algorithm with one synthetic and two real-world data sets. For two of the three data sets, the true manifolds underlying the data can be directly visualized with 3-D graphics. The true manifold for the third data set is high-dimensional and cannot be visualized directly. Applying the algorithm of mKDR to these data sets yielded interesting and informative low-dimensional representations, confirming the potential of the algorithm for exploratory data analysis and visualization of high-dimensional data.



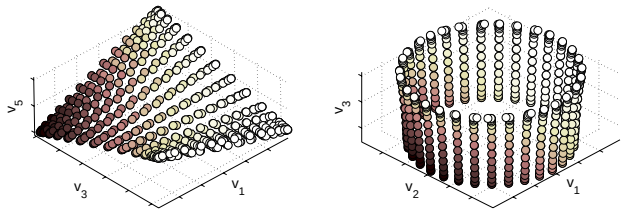
(a) (b)

Figure 2. Central subspaces of data on torus, sampled uniformly and randomly. See text for details.

4.1. Regression on a Torus

We begin by analyzing data points lying on the surface of a torus, illustrated in Fig. 1. A torus can be constructed by rotating a 2-D cycle in $\mathbb{R}^3$ with respect to an axis. Therefore, a data point on the surface has two degrees of freedom: the rotated angle θ_r with respect to the axis and the polar angle θ_p on the cycle. We synthesized our data set by sampling these two angles from the Cartesian product $[0 \quad 2\pi] \times [0 \quad 2\pi]$. The 3-D coordinates of our torus are thus given by $x_1 = (2 + \cos\theta_r)\cos\theta_p$, $x_2 = (2 + \cos\theta_r)\sin\theta_p$, and $x_3 = \sin\theta_r$. We then embed the torus in $\boldsymbol{x} \in \mathbb{R}^{10}$ by augmenting the coordinates with 7-dimensional all-zero or random vectors. To set up the regression problem, we define the response by $y = \sigma\left[-17\left(\sqrt{(\theta_r - \pi)^2 + (\theta_p - \pi)^2} - 0.6\pi\right)\right]$ where $\sigma[\cdot]$ is the sigmoid function. Note that y is radial symmetric, depending only on the distance between (θ_r, θ_p) and (π, π). The colors on the surface of the torus in Fig. 1 correspond to the value of the response.

We applied mKDR to the torus data set generated from 961 uniformly sampled angles θ_p and θ_r. We used $M = 50$ bottom eigenvectors from the graph Laplacian. The mKDR algorithm then computed the matrix $\boldsymbol{\Phi} \in \mathbb{R}^{50\times 50}$ that minimizes the empirical conditional covariance operator; cf. eq. (9). We found that this matrix is nearly rank 1 and can be approximated by $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}$ where $\boldsymbol{a}$ is the eigenvector corresponding to the largest eigenvalue. Hence, we projected the 50-D embedding of the graph Laplacian onto this principal direction $\boldsymbol{a}$. Fig. 2(a) shows the scatter plot of the projections and the responses of all samples. The scatter plot reveals a clear linear relation. Thus, if we want to regress the response on the eigenvectors returned by the graph Laplacian, a linear regression function is appropriate and very likely to be sufficient.



(a) (b)

Figure 3. Visualizing data using the most predictive eigenvectors, contributing the most to the central subspace, as well as the bottom eigenvectors ignoring responses. See text for details.

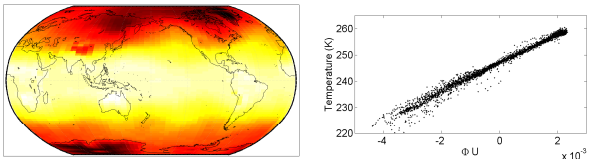
The linear relation revealed in Fig. 2(a) does not mean that it is appropriate to choose linear functions to regress the response on the *original* coordinates of the torus. Specifically, the coordinates are mapped to the kernel space via a highly nonlinear mapping induced by the graph Laplacian. For this particular example, the kernel mapping is powerful enough to make linear regression in the RKHS sufficient.

In practice it is often not possible to achieve a uniform sampling of the underlying manifold. Furthermore it is reasonable to assume that both covariates and response are noisy. To examine the robustness of the mKDR algorithm, we create a torus data set where θ_r and θ_p are randomly sampled and the covariate $\boldsymbol{x}$ and response y are disturbed by additive Gaussian noise. Fig. 2(b) illustrates the central subspace of the noisy data set. Comparing to the noiseless central subspace in Fig. 2(a), the central subspace is more diffuse although the general trend is still linear.

The principal direction $\boldsymbol{a}$ allows us to view central subspaces in the RKHS induced by the manifold learning kernel. It also gives rise to the possibility of visualizing data in the coordinate space induced by the graph Laplacian. Specifically, $\boldsymbol{a}$ encodes the combining coefficients of eigenvectors. To understand which eigenvectors are the most useful in forming the principal direction, we chose the three eigenvectors with largest coefficients in magnitude. They corresponded to the first, the third and the fifth bottom eigenvectors of graph Laplacian. We call them *predictive eigenvectors*. Fig. 3(a) shows the 3-*D* embedding of samples using the predictive eigenvectors as coordinates, where the color encodes the responses. As a contrast, Fig. 3(b) shows the 3-*D* embedding of the torus using the bottom three eigenvectors. The difference in how data is visualized is clear: mKDR arranges samples under the guidance of the responses while unsupervised graph Laplacian does so solely based on the intrinsic geometry of the covariates.

4.2. Predicting Global Temperature

To investigate the effectiveness of mKDR on complex nonlinear regression problems, we have applied it to visualize



(a) (b)

Figure 4. A map of the global temperature in Dec. 2004 and its central subspace. See text for details

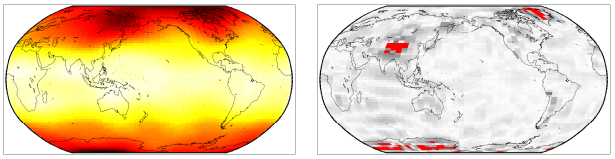
and analyze a set of 3168 satellite measurements of temperatures in the middle troposphere (Remote Sensing Systems, 2004). Fig. 4(a) shows these temperature encoded by colors on a world map, where yellow or white colors mean hotter temperature and red colors mean lower temperature. Our regression problem is to predict the temperature using the coordinates of latitude and longitude. Note that there are only two covariates in the problem. However, it is not obvious what is the most suitable regression function by reading the temperature map. Moreover, the domain of the covariates is not Euclidean, but rather ellipsoidal, a fact that a regression method should take into consideration when estimating the global temperature distribution.

We applied the algorithm of mKDR to the temperature data set. We choose $M = 100$ eigenvectors from the graph Laplacian. We projected the M-dimensional manifold embedding onto the principal direction of the linear map $\boldsymbol{\Phi}$. Fig. 4(b) displays the scatter plot of the projection against the temperatures. The relationship between the two is largely linear.

We tested the linearity by regressing the temperatures on the projections, using a linear regression function. The predicted temperatures and the prediction errors are shown in Fig. 5(a) and Fig. 5(b), with color encoding temperatures and errors respectively. Note that the central space predicts the overall temperature pattern well. In areas of inner Greenland, inner Antarctica and the Himalayas, the prediction error are relatively large, shown in red color in Fig. 5(b). Climates in these areas are typically extreme and vary significantly even across small local regions. Such variations might not be well represented by the graph Laplacian eigenfunctions which are smooth.

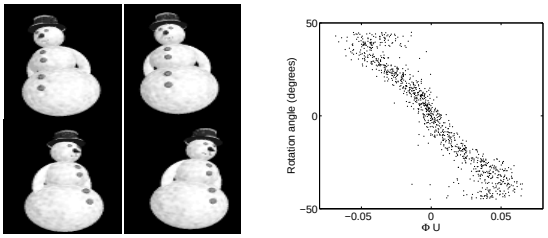
4.3. Regression on Image Manifolds

In our two previous experiments, the underlying manifolds are low dimensional and can be directly visualized. In this section, we experiment with a real-world data set whose underlying manifold is high-dimensional and unknown. Our data set contains one thousand 110×80 images of a snowman in a three-dimensional scene. Each snowman is rotated around its axis with an angle chosen uniformly random from the interval $[-45°, 45°]$, and tilted with an an-



(a) (b)

Figure 5. Prediction and prediction errors of the global temperature. See text for details.



(a) (b)

Figure 6. Images of rotating, tilting and translating snowman and its central subspace when the rotation angle is used as regression response. See text for details.

gle chosen uniformly random from $[-10°, 10°]$. Moreover, the objects are subject to random vertical plane translations spanning 5 pixels in each direction. A few representative images of such variations are shown in Fig. 6(a). Note that all variations are chosen independently. Therefore, the data set resides on a four-dimensional manifold embedded in 110×80-dimensional space.

Suppose that we are interested in the rotation angle and would like to create a low-dimensional embedding of the data that highlights this aspect of image variation, while also maintaining variations in other factors. To this end, we set up a regression problem whose covariates are image pixel intensities and whose responses are the rotation angles of the images.

We applied the mKDR algorithm to the data set and the associated regression problem, using $M = 100$ eigenvectors of the graph Laplacian. The algorithm resulted in a central space whose first direction correlates fairly well with the rotation angle, as shown in Fig. 6(b). In Fig. 7(a), we visualize the data in $\mathbb{R}^3$ by using the top three predictive eigenvectors. We use colors to encode the rotation angle of each point. The data clustering pattern in the rotation angles are clearly present in the embedding. As a contrast, we used the bottom three eigenvectors to visualize the data as typically practiced in unsupervised manifold learning. The embedding, shown in Fig. 7(b) with colors encoding rotation angles, shows no clear structure or pattern, in terms of rotation. In fact, the nearly one-dimensional embedding reflects closely the tilt angle, which tends to cause large variations in image pixel intensities.

5. Related Work

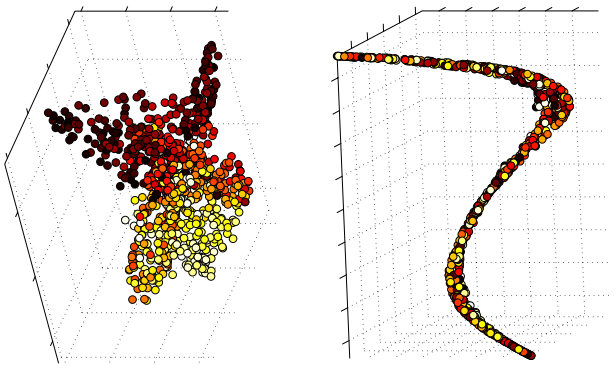
There has been relatively little previous work on the supervised manifold learning problem. Sajama and Orlitsky (2005) recently proposed a dimensionality algorithm that exploits supervised information to tie the parameters of Gaussian mixture models. This contrasts to the model-free assumption of the SDR framework. This is also the case for the manifold regularization framework (Belkin et al., 2006), which implements semi-supervised learning with regularization terms controlling the complexity both in the ambient and intrinsic geometries. Side information can also be used to achieve better low-dimensional embeddings in the "pure" setting of manifold learning (Yang et al., 2006). However, that work does not treat regression; the side information is the prior knowledge of a "correct" embedding, not the predictive information coded by responses. Finally, it is important to emphasize that despite the similarity in terminology, the work on "sufficient dimensionality reduction" of Globerson and Tishby (2003) is fundamentally different from the SDR framework studied here, in that it focuses on algorithms for compressing data in the form of two-way contingency tables. It is closely related to low-rank matrix factorization for compact representation of matrices; it is not a regression methodology.

6. Conclusions

Dimensionality reduction is an essential component of many high-dimensional data analysis procedures. This paper presents a new method for dimensionality reduction that is appropriate when supervised information is available to guide the process of finding manifolds of reduced dimensionality yet with high predictive power. Our approach is based on two strands of research that have hitherto not interacted: sufficient dimension reduction from the statistics literature and manifold learning from the machine learning literature. The bridge that connects these ideas is the recently proposed methodology of kernel dimension reduction.

We have proposed an algorithm of manifold kernel dimension reduction (mKDR). We have applied the algorithm to several synthetic and real-world data sets in the interest of exploratory data analysis and visualization. In these experiments, the algorithm discovered low-dimensional and predictive subspaces and revealed interesting and useful data patterns that are not accessible to unsupervised manifold learning algorithms.

The mKDR algorithm is a particular instantiation of the framework of kernel dimension reduction. Therefore, it en-



(a) (b)

Figure 7. Embedding the snowman images with predictive eigenvectors and bottom graph Laplacian eigenvectors, respectively. Color corresponds to rotation angle.

joys many of the desirable properties of kernel methods in general, including the ability to handle of multivariate response variables and non-vectorial data. We view mKDR as a promising general tool for the visualization of complex data types.

Acknowledgements

This work has been partly supported by grants from the Swedish Knowledge Foundation, AstraZeneca, Yahoo! Research and Microsoft Research.

References

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

Belkin, M., Niyogi, P., & Sindhwani, V. (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, *7*, 2399–2434.

Bengio, Y., Paiement, J.-F., & Vincent, P. (2003). *Out-of-sample extensions for LLE, Isomap, MDS, Eigenmaps, and spectral clustering* (Technical Report 1238). Département d'Informatique et Recherche Opérationnelle, Université de Montréal.

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

Coifman, R., Lafon, S., Lee, A., Maggioni, M., Nadler, B., Warner, F., & Zucker, S. (2005). Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences USA*, *102*, 7426–7431.

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

Cook, R. D., & Li, B. (1991). Discussion of Li (1991). *Journal of the American Statistical Association*, *86*, 328–332.

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

Donoho, D. L., & Grimes, C. (2003). Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences USA*, *100*, 5591–5596.

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

Fukumizu, K., Bach, F. R., & Jordan, M. I. (2006). *Kernel dimension reduction in regression* (Technical Report). Department of Statistics, University of California, Berkeley.

Globerson, A., & Tishby, N. (2003). Sufficient dimensionality reduction. *Journal of Machine Learning Research*, *3*, 1307–1331.

Ham, J., Lee, D., Mika, S., & Schölkopf, B. (2004). A kernel view of the dimensionality reduction of manifolds. *Proceedings of the 21'st International Conference on Machine Learning*. ACM.

Lafon, S. (2004). *Diffusion maps and geometric harmonics*. Doctoral dissertation, Yale University.

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

Li, K.-C. (1991). Sliced inverse regresion for dimension reduction. *Journal of the American Statistical Association*, *86*, 316–327.

Li, K.-C. (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *Journal of the American Statistical Association*, *86*, 316–342.

Remote Sensing Systems (2004). Microwave sounding units (MSU) data. sponsored by the NOAA Climate and Global Change Program. Data available at www.remss.com.

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

Sajama, & Orlitsky, A. (2005). Supervised dimensionality reduction using mixture models. *Proceedings of the 22'nd International Conference on Machine Learning* (pp. 760–767). ACM.

Sha, F., & Saul, L. (2005). Analysis and extension of spectral methods for nonlinear dimensionality reduction. *Proceedings of the 22'nd International Conference on Machine Learning* (pp. 785–792). ACM.

Tenenbaum, J., de Silva, V., & Langford, J. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, *290*, 2319–2322.

Yang, X., Fu, H., Zha, H., & Barlow, J. (2006). Semi-supervised nonlinear dimensionality reduction. *Proceedings of the 23'rd International Conference on Machine Learning* (pp. 1065–1072). ACM.