Journal of Machine Learning Research 11 (2010) 2175-2198 Submitted 7/07; Revised 2/09; Published 8/10

Learning Gradients: Predictive Models that Infer Geometry and Statistical Dependence

Qiang Wu WUQIANG@MATH.MSU.EDU
Department of Mathematics
Michigan State University
East Lansing, MI 48824, USA

Justin Guinney JUSTIN.GUINNEY@SAGEBASE.ORG
Sage Bionetworks
Fred Hutchinson Cancer Research Center
Seattle, WA 98109, USA

Mauro Maggioni[*] MAURO.MAGGIONI@DUKE.EDU
Sayan Mukherjee[†] SAYAN@STAT.DUKE.EDU
Departments of Mathematics and Statistical Science
Duke University
Durham, NC 27708, USA

Editor: Marina Meila

Abstract

The problems of dimension reduction and inference of statistical dependence are addressed by the modeling framework of learning gradients. The models we propose hold for Euclidean spaces as well as the manifold setting. The central quantity in this approach is an estimate of the gradient of the regression or classification function. Two quadratic forms are constructed from gradient estimates: the gradient outer product and gradient based diffusion maps. The first quantity can be used for supervised dimension reduction on manifolds as well as inference of a graphical model encoding dependencies that are predictive of a response variable. The second quantity can be used for nonlinear projections that incorporate both the geometric structure of the manifold as well as variation of the response variable on the manifold. We relate the gradient outer product to standard statistical quantities such as covariances and provide a simple and precise comparison of a variety of supervised dimensionality reduction methods. We provide rates of convergence for both inference of informative directions as well as inference of a graphical model of variable dependencies.

Keywords: gradient estimates, manifold learning, graphical models, inverse regression, dimension reduction, gradient diffusion maps

1. Introduction

The problem of developing predictive models given data from high-dimensional physical and biological systems is central to many fields such as computational biology. A premise in modeling natural phenomena of this type is that data generated by measuring thousands of variables lie on or near a low-dimensional manifold. This hearkens to the central idea of reducing data to only relevant

∗. Also in the Department of Computer Science.
†. Also in the Department of Computer Science and Institute for Genome Sciences & Policy.

information. This idea was fundamental to the paradigm of Fisher (1922) and goes back at least to Adcock (1878) and Edgeworth (1884). For an excellent review of this program see Cook (2007). In this paper we examine how this paradigm can be used to infer geometry of the data as well as statistical dependencies relevant to prediction.

The modern reprise of this program has been developed in the broad areas of manifold learning and simultaneous dimension reduction and regression. Manifold learning has focused on the problem of projecting high-dimensional data onto a few directions or dimensions while respecting local structure and distances. A variety of unsupervised methods have been proposed for this problem (Tenenbaum et al., 2000; Roweis and Saul, 2000; Belkin and Niyogi, 2003; Donoho and Grimes, 2003). Simultaneous dimension reduction and regression considers the problem of finding directions that are informative with respect to predicting the response variable. These methods can be summarized by three categories: (1) methods based on inverse regression (Li, 1991; Cook and Weisberg, 1991; Fukumizu et al., 2005; Wu et al., 2007), (2) methods based on gradients of the regression function (Xia et al., 2002; Mukherjee and Zhou, 2006; Mukherjee and Wu, 2006), (3) methods based on combining local classifiers (Hastie and Tibshirani, 1996; Sugiyama, 2007). Our focus is on the supervised problem. However, we will use the core idea in manifold learning, local estimation.

We first illustrate with a simple example how gradient information can be used for supervised dimension reduction. Both linear projections as well as nonlinear embeddings based on gradient estimates are used for supervised dimension reduction. In both approaches the importance of using the response variable is highlighted.

The main contributions in this paper consist of (1) development of gradient based diffusion maps, (2) precise statistical relations between the above three categories of supervised dimension reduction methods, (3) inference of graphical models or conditional independence structure given gradient estimates, (4) rates of convergence of the estimated graphical model. The rate of convergence depends on a geometric quantity, the intrinsic dimension of the gradient on the manifold supporting the data, rather than the sparsity of the graph.

2. A Statistical Foundation for Learning Gradients

The problem of regression can be summarized as estimating the function

$$f_r(x) = \mathbb{E}(Y \mid X = x)$$

from data $D = \{L_i = (Y_i, X_i)\}_{i=1}^n$ where X_i is a vector in a p-dimensional compact metric space $X \in \mathcal{X} \subset \mathbb{R}^p$ and $Y_i \in \mathbb{R}$ is a real valued output. Typically the data are drawn iid from a joint distribution, $L_i \overset{iid}{\sim} \rho(X,Y)$ thus specifying a model

$$y_i = f_r(x_i) + \varepsilon_i$$

with ε_i drawn iid from a specified noise model. We define ρ_x as the marginal distribution of the explanatory variables.

In this paper we will explain how inference of the gradient of the regression function

$$\nabla f_r = \left(\frac{\partial f_r}{\partial x^1}, \ldots, \frac{\partial f_r}{\partial x^p} \right)^T$$

provides information on the geometry and statistical dependencies relevant to predicting the response variable given the explanatory variables. Our work is motivated by the following ideas. The gradient is a local concept as it measures local changes of a function. Integrating information encoded by the gradient allows for inference of the geometric structure in the data relevant to the response. We will explore two approaches to integrate this local information. The first approach is averaging local gradient estimates and motivates the study of the gradient outer product (GOP) matrix. The GOP can be used to motivate a variety of linear supervised dimension reduction formulations. The GOP can also be considered as a covariance matrix and used for inference of conditional dependence between predictive explanatory variables. The second approach is to paste local gradient estimates together. This motivates the study of gradient based diffusion maps (GDM). This operator can be used for nonlinear supervised dimension reduction by embedding the support of the marginal distribution onto a much lower dimensional manifold that varies with respect to the response.

The gradient outer product Γ is a $p \times p$ matrix with elements

$$\Gamma_{ij} = \left\langle \frac{\partial f_r}{\partial x^i}, \frac{\partial f_r}{\partial x^j} \right\rangle_{L^2_{\rho_X}},$$

where ρ_X is the marginal distribution of the explanatory variables and $L^2_{\rho_X}$ is the space of square integrable functions with respect to the measure ρ_X. Using the notation $a \otimes b = ab^T$ for $a, b \in \mathbb{R}^p$, we can write

$$\Gamma = \mathbb{E}(\nabla f_r \otimes \nabla f_r).$$

This matrix provides global information about the predictive geometry of the data (developed in Section 2.2) as well as inference of conditional independence between variables (developed in Section 5). The GOP is meaningful in both the Euclidean as well as the manifold setting where the marginal distribution ρ_X is concentrated on a much lower dimensional manifold $\mathcal{M}$ of dimension $d_{\mathcal{M}}$ (developed in Section 4.1.2).

Since the GOP is a global quantity and is constructed by averaging the gradient over the marginal distribution of the data it cannot isolate local information or local geometry. This global summary of the joint distribution is advantageous in statistical analyses where global inferences are desired: global predictive factors comprised of the explanatory variables or global estimates of statistical dependence between explanatory variables. It is problematic to use a global summary for constructing a nonlinear projection or embedding that captures the local predictive geometry on the marginal distribution.

Random walks or diffusions on manifolds and graphs have been used for a variety of nonlinear dimension reduction or manifold embedding procedures (Belkin and Niyogi, 2003, 2004; Szummer and Jaakkola, 2001; Coifman et al., 2005a,b). Our basic idea is to use local gradient information to construct a random walk on a graph or manifold based on the ideas of diffusion analysis and diffusion geometry (Coifman and Lafon, 2006; Coifman and Maggioni, 2006). The central quantity in diffusion based approaches is the definition of a diffusion operator L based on a similarity metric W_{ij} between two points x_i and x_j. A commonly used diffusion operator is the graph Laplacian

$$L = I - D^{-1/2} W D^{-1/2}, \quad \text{where } D_{ii} = \sum_j W_{ij}.$$

Dimension reduction is achieved by projection onto a spectral decomposition of the operator L or powers of the operator L^t which corresponds to running the diffusion for some time t.

We propose the following similarity metric called the gradient based diffusion map (GDM) to smoothly paste together local gradient estimates

$$W_{ij} = W_f(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{\sigma_1} - \frac{|\frac{1}{2}(\nabla f_r(x_i) + \nabla f_r(x_j)) \cdot (x_i - x_j)|^2}{\sigma_2} \right).$$

In the above equation the first term $\|x_i - x_j\|^2$ encodes the local geometry of the marginal distribution and the second term pastes together gradient estimates between neighboring points. The first term is used in unsupervised dimension reduction methods such as Laplacian eigenmaps or diffusion maps (Belkin and Niyogi, 2003). The second term can be interpreted as a diffusion map on the function values, the approximation is due to a first order Taylor expansion

$$\frac{1}{2}(\nabla f_r(x_i) + \nabla f_r(x_j)) \cdot (x_i - x_j) \approx f_r(x_i) - f_r(x_j) \qquad \text{if } x_i \approx x_j.$$

A related similarity was briefly mentioned in Coifman et al. (2005b) and used in Szlam et al. (2008) in the context of semi-supervised learning. In Section 4.2 we study this nonlinear projection method under the assumption that the marginal distribution is concentrated on a much lower dimensional manifold $\mathcal{M}$.

2.1 Illustration of Linear Projections and Nonlinear Embeddings

The simple example in this section fixes the differences between linear projections and nonlinear embeddings using either diffusion maps or gradient based diffusion maps. The marginal distribution is uniform on the following manifold

$$X_1 = t\cos(t), \quad X_2 = 70h, \quad X_3 = t\sin(t) \text{ where } t = \frac{3\pi}{2}(1+2\theta), \theta \in [0,1], h \in [0,1],$$

and the regression function is $Y = \sin(5\pi\theta)$, see Figure 1(a). In this example a two dimensional manifold is embedded in $\mathbb{R}^3$ and the variation of the response variable can be embedded onto one dimension. The points in Figure 1(a) are the points on the manifold and the false color signifies the function value at these points. In Figure 1(b) the data is embedded in two dimensions using diffusion maps with the function values displayed in false color. It is clear that the direction of greatest variation is X_2 which corresponds to h. It is not the direction along which the regression function has greatest variation. In Figure 1(c) the data is projected onto two axes using the GOP approach and we see that the relevant dimensions X_1, X_3 are recovered. This example also shows that linear dimension reduction may still make sense in the manifold setting. In Figure 1(d) the data is embedded using gradient based diffusion maps and we capture the direction θ in which the data varies with respect to the regression function.

2.2 Gradient Outer Product Matrix and Dimension Reduction

In the case of linear supervised dimension reduction we assume the following semi-parametric model holds

$$Y = f_r(X) + \varepsilon = g(b_1^T X, \ldots, b_d^T X) + \varepsilon, \tag{1}$$

where ε is noise and $B = (b_1^T, ..., b_d^T)$ is the dimension reduction (DR) space. In this case the number of explanatory variables p is large but the the response variable Y depends on a few directions in $\mathbb{R}^p$

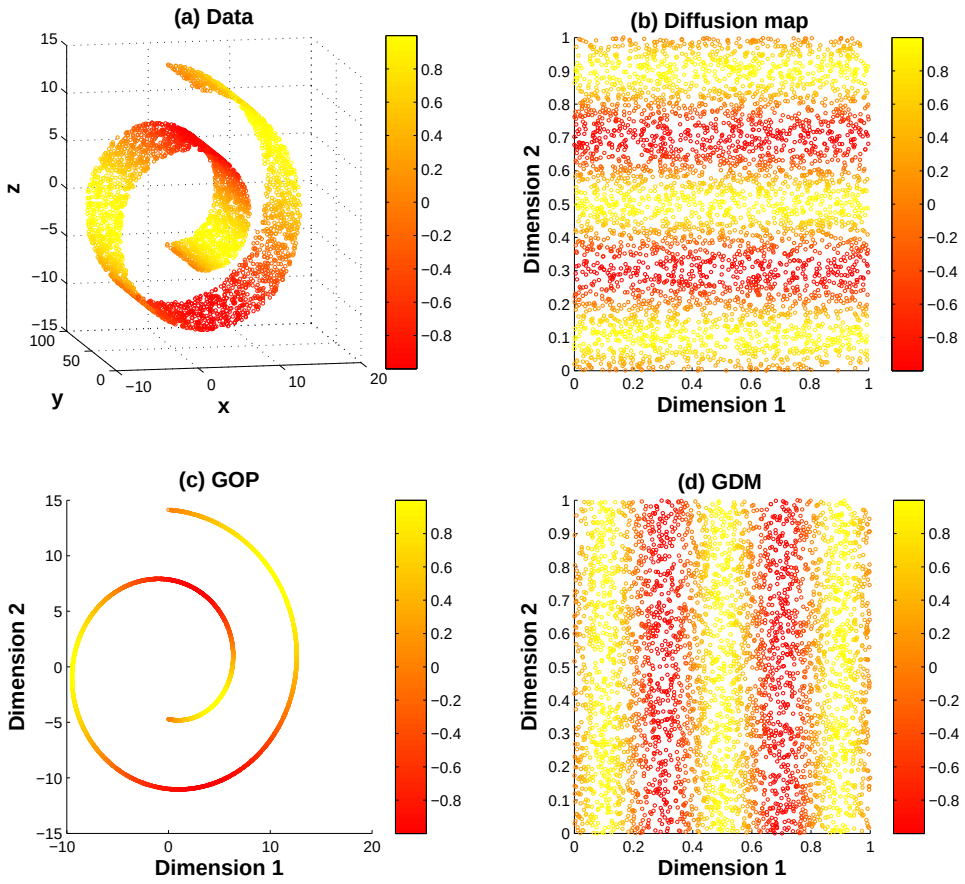


Figure 1: (a) Plot of the original three dimensional data with the color of the point corresponding to the function value; (b) Embedding the data onto two dimensions using diffusion maps, here dimensions 1 and 2 corresponds to h and θ respectively; (c) Projection of the data onto two dimensions using gradient based dimension reduction, here dimensions 1 and 2 corresponds to x and z respectively; (d) Embedding the data onto two dimensions using gradient based diffusion maps, here dimensions 1 and 2 corresponds to θ and h respectively.

and dimension reduction becomes the central problem in finding an accurate regression model. In the following we develop theory relating the gradient of the regression function to the above model of dimension reduction.

Observe that for a vector $v \in \mathbb{R}^p$, $\frac{\partial f_r(x)}{\partial v} = v \cdot \nabla f_r$ is identically zero if f_r does not depends on v and is not zero if f_r changes along the direction v. The following lemma relates the gradient outer product matrix to supervised dimension reduction.

Lemma 1 *Under the assumptions of the semi-parametric model* (1)*, the gradient outer product matrix* Γ *is of rank at most* d*. Denote by* $\{v_1, \ldots, v_d\}$ *the eigenvectors associated to the nonzero eigenvalues of* Γ*. The following holds*

$$\text{span}(B) = \text{span}(v_1, \ldots, v_d).$$

Lemma 1 states the dimension reduction space can be computed by a spectral decomposition of Γ. Notice that this method does not require additional geometric conditions on the distribution. This is in contrast to other supervised dimension reduction methods (Li, 1991; Cook and Weisberg, 1991; Li, 1992) that require a geometric or distributional requirement on X, namely that the level sets of the distribution are elliptical.

This observation motivates supervised dimension reduction methods based on consistent estimators Γ_D of Γ given data D. Several methods have been motivated by this idea, either implicitly as in minimum average variance estimation (MAVE) (Xia et al., 2002), or explicitly as in outer product of gradient (OPG) (Xia et al., 2002) and learning gradients (Mukherjee et al., 2010).

The gradient outer product matrix is defined globally and its relation to dimension reduction in Section 2.2 is based on global properties. However, since the gradient itself is a local concept we can also study the geometric structure encoded in the gradient outer product matrix from a local point of view.

2.3 Gradient Outer Product Matrix as a Covariance Matrix

A central concept used in dimension reduction is the covariance matrix of the inverse regression function $\Omega_{X|Y} = \text{cov}_Y[\mathbb{E}_X(X \mid Y)]$. The fact that $\Omega_{X|Y}$ encodes the DR directions $B = (b_1, \ldots, b_d)$ under certain conditions was explored in Li (1991).

The main result of this subsection is to relate the two matrices: Γ and $\Omega_{X|Y}$. The first observation from this relation is that the gradient outer product matrix is a covariance matrix with a very particular construction. The second observation is that the gradient outer product matrix contains more information than the covariance of the inverse regression since it captures local information. This is outlined for linear regression and then generalized to nonlinear regression. Proofs of the propositions and the underlying mathematical ideas will be developed in Section 4.1.1.

The linear regression problem is often stated as

$$Y = \beta^T X + \varepsilon, \quad \mathbb{E}\varepsilon = 0. \tag{2}$$

For this model the following relation between gradient estimates and the inverse regression holds.

Proposition 2 *Suppose* (2) *holds. Given the covariance of the inverse regression,* $\Omega_{X|Y} = \text{cov}_Y(\mathbb{E}_X(X \mid Y))$*, the variance of the output variable,* $\sigma_Y^2 = \text{var}(Y)$*, and the covariance of the input variables,* $\Sigma_X = \text{cov}(X)$*, the gradient outer product matrix is*

$$\Gamma = \sigma_Y^2 \left(1 - \frac{\sigma_\varepsilon^2}{\sigma_Y^2}\right)^2 \Sigma_X^{-1} \Omega_{X|Y} \Sigma_X^{-1}, \tag{3}$$

assuming that Σ_X *is full rank.*

The above result states that for a linear model the matrices Γ and $\Omega_{X|Y}$ are equivalent modulo a scale parameter—approximately the variance of the output variable—and a rotation—the precision matrix (inverse of the covariance matrix) of the input variables.

In order to generalize Proposition 2 to the nonlinear regression setting we first consider piecewise linear functions. Suppose there exists a non-overlapping partition of the input space

$$\mathcal{X} = \bigcup_{i=1}^{I} R_i$$

such that in each region R_i the regression function f_r is linear and the noise has zero mean

$$f_r(x) = \beta_i^T x, \quad \mathbb{E}\,\varepsilon_i = 0 \quad \text{for } x \in R_i. \tag{4}$$

The following corollary is true.

Corollary 3 *Given partitions R_i of the input space for which* (4) *holds, define in each partition R_i the following local quantities: the covariance of the input variables $\Sigma_i = \mathrm{cov}(X \in R_i)$, the covariance of the inverse regression $\Omega_i = \mathrm{cov}_Y(\mathbb{E}_X(X \in R_i \mid Y))$, the variance of the output variable $\sigma_i^2 = \mathrm{var}(Y \mid X \in R_i)$. Assuming that matrices Σ_i are full rank, the gradient outer product matrix can be computed in terms of these local quantities*

$$\Gamma = \sum_{i=1}^{I} \rho_X(R_i)\sigma_i^2 \left(1 - \frac{\sigma_{\varepsilon_i}^2}{\sigma_i^2}\right)^2 \Sigma_i^{-1}\Omega_i\Sigma_i^{-1}, \tag{5}$$

where $\rho_X(R_i)$ is the measure of partition R_i with respect to the marginal distribution ρ_X.

If the regression function is smooth it can be approximated by a first order Taylor series expansion in each partition R_i provided the region is small enough. Theoretically there always exist partitions such that the locally linear models (4) hold approximately (i.e., $\mathbb{E}\,\varepsilon_i \approx 0$). Therefore (5) holds approximately

$$\Gamma \approx \sum_{i=1}^{I} \rho_X(R_i)\sigma_i^2 \left(1 - \frac{\sigma_{\varepsilon_i}^2}{\sigma_i^2}\right)^2 \Sigma_i^{-1}\Omega_i\Sigma_i^{-1}.$$

Supervised dimension reduction methods based on the covariance of the inverse regression require an elliptic condition on X. This condition ensures that $\Omega_{X|Y}$ encodes the DR subspace but not necessarily the entire DR subspace. In the worst case it is possible that $\mathbb{E}(X \mid Y) = 0$, $\Omega_{X|Y} = \mathbf{0}$. and as a result $\Omega_{X|Y}$ contains no information. In the linear case $\Omega_{X|Y}$ will encode the full DR subspace, the one predictive direction. Since the GOP is an average the inverse covariance matrix of the locally linear models it contains all the predictive directions. This motivates the centrality of the gradient outer product.

This derivation of the gradient outer product matrix based on local variation has two potential implications. It provides a theoretical comparison between dimension reduction approaches based on the gradient outer product matrix and inverse regression. This will be explored in Section 4.1.1 in detail. The integration of local variation will be used to infer statistical dependence between the explanatory variables conditioned on the response variable in Section 5.

A common belief in high dimensional data analysis is that the data are concentrated on a low dimensional manifold. Both theoretical and empirical evidence of this belief is accumulating. In the manifold setting, the input space is a manifold $\mathcal{X} = \mathcal{M}$ of dimension $d_{\mathcal{M}} \ll p$. We assume the existence of an isometric embedding $\varphi : \mathcal{M} \to \mathbb{R}^p$ and the observed input variables $(x_i)_{i=1}^n$ are the image of points $(q_i)_{i=1}^n$ drawn from a distribution on the manifold: $x_i = \varphi(q_i)$. In this case, global

statistics are not as meaningful from a modeling perspective.[1] In this setting the gradient outer product matrix should be defined in terms of the gradient on the manifold, $\nabla_{\mathcal{M}} f_r$,

$$\Gamma = \mathbb{E}\left(d\varphi(\nabla_{\mathcal{M}} f_r) \otimes d\varphi(\nabla_{\mathcal{M}} f_r)\right) = \mathbb{E}\left(d\varphi(\nabla_{\mathcal{M}} f_r \otimes \nabla_{\mathcal{M}} f_r)(d\varphi)^T\right).$$

This quantity is meaningful from a modeling perspective because gradients on the manifold capture the local structure in the data. Note that the $d_{\mathcal{M}} \times d_{\mathcal{M}}$ matrix $\Gamma_{\mathcal{M}} = \nabla_{\mathcal{M}} f_r \otimes \nabla_{\mathcal{M}} f_r$ is the central quantity of interest in this setting. However, we know neither the manifold nor the coordinates on the manifold and are only provided points in the ambient space. For this reason we cannot compute $\Gamma_{\mathcal{M}}$. However, we can understand its properties by analyzing the gradient outer product matrix Γ in the ambient space, a $p \times p$ matrix. Details on conditions under which Γ provides information on $\Gamma_{\mathcal{M}}$ are developed in Section 4.1.2.

3. Estimating Gradients

An estimate of the gradient is required in order to estimate the gradient outer product matrix Γ. Many approaches for computing gradients exist including various numerical derivative algorithms, local linear/polynomial smoothing (Fan and Gijbels, 1996), and learning gradients by kernel models (Mukherjee and Zhou, 2006; Mukherjee and Wu, 2006). Our main focus is on what can be done given an estimate of Γ rather than estimation methods for the gradient. The application domain we focus on is the analysis of high-dimensional data with few observations, where $p \gg n$ and some traditional methods do not work well because of computational complexity or numerical stability. Learning gradients by kernel models was specifically developed for this type of data in the Euclidean setting for regression (Mukherjee and Zhou, 2006) and classification (Mukherjee and Wu, 2006). The same algorithms were shown to be valid for the manifold setting with a different interpretation in Mukherjee et al. (2010). In this section we review the formulation of the algorithms and state properties that will be relevant in subsequent sections.

The motivation for learning gradients is based on Taylor expanding the regression function

$$f_r(u) \approx f_r(x) + \nabla f_r(x) \cdot (u - x), \text{ for } x \approx u,$$

which can be evaluated at data points $(x_i)_{i=1}^n$

$$f_r(x_i) \approx f_r(x_j) + \nabla f_r(x_j) \cdot (x_i - x_j), \text{ for } x_i \approx x_j.$$

Given data $D = \{(y_i, x_i)\}_{i=1}^n$ the objective is to simultaneously estimate the regression function f_r by a function f_D and the gradient ∇f_r by the p-dimensional vector valued function $\vec{f}_D$.

In the regression setting the following regularized loss functional provides the estimates (Mukherjee and Zhou, 2006).

Definition 4 *Given the data $D = \{(x_i, y_i)\}_{i=1}^n$, define the first order difference error of function f and vector-valued function $\vec{f} = (f_1, \ldots, f_p)$ on D as*

$$\mathcal{E}_D(f, \vec{f}) = \frac{1}{n^2} \sum_{i,j=1}^{n} w_{i,j}^s \left(y_i - f(x_j) + \vec{f}(x_i) \cdot (x_j - x_i)\right)^2.$$

1. Consider $\Omega(X \mid Y)$ as an example. For any given y, the set $\{x \mid Y = y\}$ is a submanifold. The global mean will not necessarily lie on the manifold. Therefore, from a modeling perspective $\mathbb{E}(X \mid Y)$ and hence $\Omega(X \mid Y)$ may convey nothing about the manifold. Although this does not mean it is useless in practice, a theoretical justification seems impossible.

The regression function and gradient estimate is modeled by

$$(f_D, \vec{f}_D) := \arg \min_{(f,\vec{f}) \in \mathcal{H}_K^{p+1}} \left(\mathcal{E}_D(f,\vec{f}) + \lambda_1 \|f\|_K^2 + \lambda_2 \|\vec{f}\|_K^2 \right),$$

where f_D and $\vec{f}_D$ are estimates of f_r and ∇f_r given the data, $w_{i,j}^s$ is a weight function with bandwidth s, $\|\cdot\|_K$ is the reproducing kernel Hilbert space (RKHS) norm, λ_1 and λ_2 are positive constants called the regularization parameters, the RKHS norm of a p-vector valued function is the sum of the RKHS norm of its components $\|\vec{f}\|_K^2 := \sum_{t=1}^p \|\vec{f}_t\|_K^2$.

A typical weight function is a Gaussian $w_{i,j}^s = \exp(-\|x_i - x_j\|^2/2s^2)$. Note this definition is slightly different from that given in Mukherjee and Zhou (2006) where $f(x_j)$ is replaced by y_j and only the gradient estimate $\vec{f}_D$ is estimated.

In the classification setting we are given $D = \{(y_i, x_i)\}_{i=1}^n$ where $y_i \in \{-1, 1\}$ are labels. The central quantity here is the classification function which we can define by conditional probabilities

$$f_c(x) = \log \left[\frac{\rho(Y=1 \mid x)}{\rho(Y=-1 \mid x)} \right] = \arg\min \mathbb{E}\, \phi(Y f(X))$$

where $\phi(t) = \log(1+e^{-t})$ and the sign of f_c is a Bayes optimal classifier. The following regularized loss functional provides estimates for the classification function and gradient (Mukherjee and Wu, 2006).

Definition 5 *Given a sample $D = \{(x_i, y_i)\}_{i=1}^n$ we define the empirical error as*

$$\mathcal{E}_D^\phi(f,\vec{f}) = \frac{1}{n^2} \sum_{i,j=1}^n w_{ij}^s \phi\Big(y_i \big(f(x_j) + \vec{f}(x_i) \cdot (x_i - x_j) \big) \Big).$$

The classification function and gradient estimate given a sample is modeled by

$$(f_D, \vec{f}_D) = \arg \min_{(f,\vec{f}) \in \mathcal{H}_K^{p+1}} \left(\mathcal{E}_D^\phi(f,\vec{f}) + \lambda_1 \|f\|_K^2 + \lambda_2 \|\vec{f}\|_K^2 \right),$$

where f_D and $\vec{f}_D$ are estimates of f_c and ∇f_c, and λ_1, λ_2 are the regularization parameters.

In the manifold setting the above algorithms are still valid. However the interpretation changes. We state the regression case, the classification case is analogous (Mukherjee et al., 2010).

Definition 6 *Let $\mathcal{M}$ be a Riemannian manifold and $\varphi : \mathcal{M} \to \mathbb{R}^p$ be an isometric embedding which is unknown. Denote $\mathcal{X} = \varphi(\mathcal{M})$ and $\mathcal{H}_K = \mathcal{H}_K(\mathcal{X})$. For the sample $D = \{(q_i, y_i)\}_{i=1}^n \in (\mathcal{M} \times \mathbb{R})^n$, $x_i = \varphi(q_i) \in \mathbb{R}^p$, the learning gradients algorithm on $\mathcal{M}$ provides estimates*

$$(f_D, \vec{f}_D) := \arg \min_{f,\vec{f} \in \mathcal{H}_K^{p+1}} \left\{ \frac{1}{n^2} \sum_{i,j=1}^n w_{i,j}^s \left(y_i - f(x_j) + \vec{f}(x_i) \cdot (x_j - x_i) \right)^2 + \lambda_1 \|f\| + \lambda_2 \|\vec{f}\|_K^2 \right\},$$

where $\vec{f}_D$ is a model for $\mathrm{d}\varphi(\nabla_{\mathcal{M}} f_r)$ and f_D is a model for f_r.

From a computational perspective the advantage of the RKHS framework is that in both regression and classification the solutions satisfy a representer theorem (Wahba, 1990; Mukherjee and Zhou, 2006; Mukherjee and Wu, 2006)

$$f_D(x) = \sum_{i=1}^{n} \alpha_{i,D} K(x, x_i), \quad \vec{f}_D(x) = \sum_{i=1}^{n} c_{i,D} K(x, x_i), \tag{6}$$

with $c_D = (c_{1,D}, \ldots, c_{n,D}) \in \mathbb{R}^{p \times n}$, and $\alpha_D = (\alpha_{1,D}, ..., \alpha_{n,D})^T \in \mathbb{R}^p$. In Mukherjee and Zhou (2006) and Mukherjee and Wu (2006) methods for efficiently computing the minima were introduced in the setting where $p \gg n$. The methods involve linear systems of equations of dimension nd where $d \leq n$.

The consistency of the gradient estimates for both regression and classification were proven in Mukherjee and Zhou (2006) and Mukherjee and Wu (2006) respectively.

Proposition 7 *Under mild conditions (see Mukherjee and Zhou, 2006; Mukherjee and Wu, 2006 for details) the estimates of the gradients of the regression or classification function f converge to the true gradients: with probability greater than $1 - \delta$,*

$$\| \vec{f}_D - \nabla f \|_{L^2_{\rho_X}} \leq C \log \left(\frac{2}{\delta} \right) n^{-1/p}.$$

Consistency in the manifold setting was studied in Mukherjee et al. (2010) and the rate of convergence was determined by the dimension of the manifold, $d_{\mathcal{M}}$, not the dimension of the ambient space p.

Proposition 8 *Under mild conditions (see Mukherjee et al., 2010 for details), with probability greater than $1 - \delta$,*

$$||(\mathrm{d}\varphi)^* \vec{f}_D - \nabla_{\mathcal{M}} f||_{L^2_{\rho_{\mathcal{M}}}} \leq C \log \left(\frac{2}{\delta} \right) n^{-1/d_{\mathcal{M}}},$$

where where $(\mathrm{d}\varphi)^*$ is the dual of the map $\mathrm{d}\varphi$.

4. Dimension Reduction Using Gradient Estimates

In this section we study some properties of dimension reduction using the gradient estimates. We also relate learning gradients to previous approaches for dimension reduction in regression.

4.1 Linear Dimension Reduction

The theoretical foundation for linear dimension reduction using the spectral decomposition of the gradient outer production matrix was developed in Section 2.2. The estimate of the gradient obtained by the kernel models in Section 3 provides the following empirical estimate of the gradient outer product matrix

$$\hat{\Gamma} := c_D K^2 c_D^T = \frac{1}{n} \sum_{i=1}^{n} \vec{f}_D(x_i) \otimes \vec{f}_D(x_i),$$

where K is the kernel matrix with $K_{ij} = K(x_i, x_j)$ and c_D is defined in (6). The eigenvectors corresponding to the top eigenvalues of $\hat{\Gamma}$ can be used to estimate the d dimension reduction directions. The following proposition states that the estimate is consistent.

Proposition 9 *Suppose that f satisfies the semi-parametric model* (1) *and $\vec{f}_D$ is an empirical approximation of ∇f. Let $\hat{v}_1, \ldots, \hat{v}_d$ be the eigenvectors of $\hat{\Gamma}$ associated to the top d eigenvalues. The following holds*

$$\lim_{n\to\infty} \text{span}(\hat{v}_1, \ldots, \hat{v}_d) = \text{span}(B).$$

Moreover, the remaining eigenvectors correspond to eigenvalues close to 0.

Proof : Proposition 7 implies that $\lim_{n\to\infty} \hat{\Gamma}_{ij} = \Gamma_{ij}$ and hence $\lim_{n\to\infty} \Gamma = \Gamma$ in matrix norm. By perturbation theory the eigenvalues (see Golub and Loan, 1996, Theorem 8.1.4 and Corollary 8.1.6) and eigenvectors (see Zwald and Blanchard, 2006) of $\hat{\Gamma}$ converge to those of Γ respectively. The conclusions then follow from Lemma 1. ■

Proposition 9 justifies linear dimension reduction using consistent gradient estimates from a global point of view.

In the next subsection we study the gradient outer product matrix from the local point of view and provide details on the relation between gradient based methods and sliced inverse regression.

4.1.1 RELATION TO SLICED INVERSE REGRESSION (SIR)

The SIR method computes the DR directions using a generalized eigen-decomposition problem

$$\Omega_{X|Y}\beta = \nu\Sigma_X\beta. \tag{7}$$

In order to study the relation between our method with SIR, we study the relation between the matrices $\Omega_{X|Y}$ and Γ.

We start with a simple model where the DR space contains only one direction which means the regression function satisfies the following semi-parametric model

$$Y = g(\beta^T X) + \varepsilon$$

where $\|\beta\| = 1$ and $\mathbb{E}\varepsilon = 0$. The following theorem holds and Proposition 2 is a special case.

Theorem 10 *Suppose that Σ_X is invertible. There exists a constant C such that*

$$\Gamma = C\Sigma_X^{-1}\Omega_{X|Y}\Sigma_X^{-1}.$$

If g is a linear function the constant is $C = \sigma_Y^2\left(1 - \frac{\sigma_\varepsilon^2}{\sigma_Y^2}\right)^2$.

Proof It is proven in Duan and Li (1991) that

$$\Omega_{X|Y} = \text{var}(h(Y))\Sigma_X\beta\beta^T\Sigma_X$$

where $h(y) = \frac{\mathbb{E}(\beta^T(X-\mu)|y)}{\beta^T\Sigma_X\beta}$ with $\mu = \mathbb{E}(X)$ and Σ_X is the covariance matrix of X. In this case, the computation of matrix Γ is direct:

$$\Gamma = \mathbb{E}[(g'(\beta^T X))^2]\beta\beta^T.$$

By the assumption Σ_X is invertible, we immediately obtain the first relation with

$$C = \mathbb{E}[(g'(\beta^T X))^2]\text{var}(h(Y))^{-1}.$$

If $g(t) = at + b$, we have $h(y) = \frac{y-b-\beta^T\mu}{a\beta^T\Sigma_X\beta}$ and consequently

$$\mathrm{var}(h(Y)) = \frac{\sigma_Y^2}{a^2(\beta^T\Sigma_X\beta)^2}.$$

By the simple fact $\mathbb{E}(g'(\beta^T X)^2] = a^2$ and $\sigma_Y^2 = a^2\beta^T\Sigma_X\beta + \sigma_\varepsilon^2$, we get

$$C = \frac{a^4(\beta^T\Sigma_X\beta)^2}{\sigma_Y^2} = \frac{(\sigma_Y^2 - \sigma_\varepsilon^2)^2}{\sigma_Y^2} = \sigma_Y^2\left(1 - \frac{\sigma_\varepsilon^2}{\sigma_Y^2}\right)^2.$$

This finishes the proof. ■

It is apparent that Γ and $\Omega_{X|Y}$ differ only up to a linear transformation. As a consequence the generalized eigen-decomposition (7) of $\Omega_{X|Y}$ with respect to Σ_X yields the same first direction as the eigen-decomposition of Γ.

Consider the linear case. Without loss of generality suppose X is normalized to satisfy $\Sigma_X = \sigma^2 I$, we see $\Omega_{X|Y}$ is the same as Γ up to a constant of about $\frac{\sigma_Y^2}{\sigma^4}$. Notice that this factor measures the ratio of the variation of the response to the variation over the input space as well as along the predictive direction. This implies that Γ is more informative because it not only contains the information of the descriptive directions but also measures their importance with respect to the change of the response variable.

When there are more than one DR directions as in model (1), we partition the input space into I small regions $X = \bigcup_{i=1}^I R_i$ such that over each region R_i the response variable y is approximately linear with respect to x and the descriptive direction is a linear combination of the column vectors of B. By the discussion in Section 2.2

$$\Gamma = \sum_i \rho_X(R_i)\,\Gamma_i \approx \sum_{i=1}^I \rho_X(R_i)\sigma_i^2\,\Sigma_i^{-1}\,\Omega_i\Sigma_i^{-1},$$

where Γ_i is the gradient outer product matrix on R_i and $\Omega_i = \mathrm{cov}_Y(\mathbb{E}_X(X \in R_i \mid Y))$. In this sense, the gradient covariance matrix Γ can be regarded as the weighted sum of the local covariance matrices of the inverse regression function. Recall that SIR suffers from the possible degeneracy of the covariance matrix of the inverse regression function over the entire input space while the local covariance matrix of the inverse regression function will not be degenerate unless the function is constant. Moreover, in the gradient outer product matrix, the importance of local descriptive directions are also taken into account. These observations partially explain the generality and some advantages of gradient based methods.

Note this theoretical comparison is independent of the method used to estimate the gradient. Hence the same comparison holds between SIR and other gradient based methods such as mean average variance estimation (MAVE) and outer product of gradients (OPG) developed in Xia et al. (2002).

4.1.2 THEORETICAL FEASIBILITY OF LINEAR PROJECTIONS FOR NONLINEAR MANIFOLDS

In this section we explore why linear projections based on the gradient outer product matrix are feasible and have meaning when the manifold structure is nonlinear. The crux of the analysis will be demonstrating that the estimated gradient outer product matrix $\hat{\Gamma}$ is still meaningful.

Again assume there exists an unknown isometric embedding of the manifold onto the ambient space, $\varphi : \mathcal{M} \to \mathbb{R}^p$. From a modeling perspective we would like the gradient estimate from data $\vec{f}_D$ to approximate $d\varphi(\nabla_{\mathcal{M}} f_r)$ (Mukherjee et al., 2010). Generally this is not true when the manifold is nonlinear, φ is a nonlinear map. Instead, the estimate provides the following information about $\nabla_{\mathcal{M}} f_r$

$$\lim_{n\to\infty} (d\varphi)^* \vec{f}_D = \nabla_{\mathcal{M}} f_r,$$

where $(d\varphi)^*$ is the dual of $d\varphi$, the differential of φ.

Note that f_r is not well defined on any open set of $\mathbb{R}^p$. Hence, it is not meaningful to consider the gradient of ∇f_r in the ambient space $\mathbb{R}^p$. Also, we cannot recover directly the gradient of f_r on the manifold since we know neither the manifold nor the embedding. However, we can still recover the DR directions from the matrix $\hat{\Gamma}$.

Assume f_r satisfies the semi-parametric model (1). The matrix Γ is not well defined but $\hat{\Gamma}$ is well defined. The following proposition ensures that the spectral decomposition of $\hat{\Gamma}$ provides the DR directions.

Proposition 11 *If $v \perp b_i$ for all $i = 1, \ldots, d$, then* $\lim_{n\to\infty} v^T \hat{\Gamma} v = 0$.

Proof Let $\vec{f}_\lambda$ be the sample limit of $\vec{f}_D$, that is

$$\vec{f}_\lambda = \arg\min_{\vec{f}\in\mathcal{H}_K^p} \left\{ \int_{\mathcal{M}} \int_{\mathcal{M}} e^{-\frac{\|x-\xi\|^2}{2s^2}} \left(f_r(x) - f_r(\xi) + \vec{f}(x)\cdot(\xi - x) \right)^2 d\rho_{\mathcal{M}}(x) d\rho_{\mathcal{M}}(\xi) + \lambda \|\vec{f}\|_K^2 \right\}.$$

By the assumption and a simple rotation argument we can show that $v \cdot \vec{f}_\lambda = 0$.

It was proven in Mukherjee and Zhou (2006) that $\lim_{n\to\infty} \|\vec{f}_D - \vec{f}_\lambda\|_K = 0$. A result of this is for $\hat{\Xi} = c_D K c_D^T$

$$v^T \hat{\Xi} v = \|v \cdot \vec{f}_D\|_K^2 \overset{n\to\infty}{\longrightarrow} \|v \cdot \vec{f}_\lambda\|_K^2 = 0.$$

This implies that $\lim_{n\to\infty} v^T \hat{\Gamma} v = 0$ and proves the proposition. ■

Proposition 11 states that all the vectors perpendicular to the DR space correspond to eigenvalues near zero of $\hat{\Gamma}$ and will be filtered out. This means the DR directions can be still found by the spectral decomposition of the estimated gradient outer product matrix.

4.2 Nonlinear Projections: Gradient Based Diffusion Maps (GDM)

As discussed in Section 2 the gradient of the regression function can be used for nonlinear projections. The basic idea was to use local gradient information to construct a diffusion operator L based on a similarity metric W_{ij} between two points x_i and x_j. A commonly used diffusion operator is the graph Laplacian

$$L = I - D^{-1/2} W D^{-1/2}, \quad \text{where } D_{ii} = \sum_j W_{ij}.$$

Dimension reduction is achieved by projection onto a spectral decomposition of the operator L or powers $(I-L)^t$ of the operator $(I-L)$ which corresponds to running the diffusion $(I-L)$ for some time t. The gradient based diffusion map (GDM) was defined as

$$W_{ij} = W_f(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{\sigma_1} - \frac{|\frac{1}{2}(\nabla f_r(x_i) + \nabla f_r(x_j))\cdot(x_i - x_j)|^2}{\sigma_2} \right). \tag{8}$$

In the above equation the first term $\|x_i - x_j\|^2$ encodes the local geometry of the marginal distribution and the second term pastes together gradient estimates between neighboring points. The first term is used in unsupervised dimension reduction methods such as Laplacian eigenmaps or diffusion maps (Belkin and Niyogi, 2003). The second term can be interpreted as a first order Taylor expansion leading to the following approximation

$$\frac{1}{2}(\nabla f_r(x_i) + \nabla f_r(x_j)) \cdot (x_i - x_j) \approx f_r(x_i) - f_r(x_j).$$

The form (8) is closely related to the following function adapted similarity proposed in Szlam et al. (2008)

$$W_f(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{\sigma_1} - \frac{|f(x_i) - f(x_j)|^2}{\sigma_2}\right),$$

where the function evaluations $f(x_i)$ are computed based on a first rough estimate of the regression function from the data.

The utility of nonlinear dimension reduction has been shown to be dramatic with respect to prediction accuracy in the semi-supervised learning setting where a large set of unlabeled data, $\{x_1, ..., x_u\}$, drawn from the marginal distribution, were used to learn the projection and a small set of labeled data $\{(y_1, x_1), ..., (y_\ell, x_\ell)\}$ were used to learn the regression function on the projected data. Of practical importance in this setting is the need to evaluate the similarity function on out of sample data. The labeled data is used to compute the gradient estimate, which can be evaluated on out-of-sample data. Given the gradient estimate and the labeled and unlabeled data $(x_1, ..., x_\ell, x_{\ell+1}, x_{\ell+u})$ the following GDM can be defined on all the samples

$$\tilde{W}_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{\sigma_1} - \frac{|\frac{1}{2}(\vec{f}_D(x_i) + \vec{f}_D(x_j)) \cdot (x_i - x_j)|^2}{\sigma_2}\right), \quad i, j = 1, .., u + \ell.$$

An analysis of the accuracy of the GDM approach is based on how well $f(x_i) - f(x_j)$ can be estimated using the gradient estimate $\vec{f}_D$. The first order Taylor expansion on the manifold results in the following approximation

$$f(x_i) - f(x_j) \approx \nabla_{\mathcal{M}} f(x_i) \cdot v_{ij}, \text{ for } v_{ij} \approx 0,$$

where $v_{ij} \in T_{x_i}\mathcal{M}$ is the tangent vector such that $x_j = \mathrm{Exp}_{x_i}(v_{ij})$ where Exp_{x_i} is the exponential map at x_i (see do Carmo, 1992; Mukherjee et al., 2010). Since we cannot compute $\nabla_{\mathcal{M}} f$ we use $\vec{f}_D$. The following proposition states that estimates of the function value differences can be accurately estimated from gradient estimate $\vec{f}_D$.

Proposition 12 *The following holds*

$$f_r(x_i) - f_r(x_j) \approx \vec{f}_D(x_i) \cdot (x_i - x_j), \text{ for } x_i \approx x_j.$$

Proof By the fact $x_i - x_j \approx \mathrm{d}\varphi(v_{ij})$ we have

$$\vec{f}_D(x_i) \cdot (x_i - x_j) \approx \langle \vec{f}_D(x_i), \mathrm{d}\varphi(v_{ij}) \rangle = \langle (\mathrm{d}\varphi)^*(\vec{f}_D(x_i)), v_{ij} \rangle \approx \langle \nabla_{\mathcal{M}} f_r(x_i), v_{ij} \rangle$$

which implies the conclusion. ■

This proposition does not prove consistency of the GDM approach. This research program of proving convergence of the eigenvectors of a graph Laplacian to the eigenfunctions of a corresponding Laplace-Beltrami operator is a source of extensive research in diffusion maps (Belkin and Niyogi, 2005; Giné and Koltchinskii, 2006). It would be of interest to adapt these approaches to the gradient setting. The limiting operator will in general depend on how σ_1 and σ_2 approach 0 as the number of points tends to infinity. For example, it is easy to see that if $\sigma_1, \sigma_2 \to 0_+$ suitably as $n \to +\infty$, with $\sigma_1/\sigma_2 = \alpha$, then the limiting operator is the Laplacian on the manifold $(x, \alpha f(x)) \subset \mathcal{M} \times \mathbb{R}$.

4.2.1 EMPIRICAL RESULTS FOR GRADIENT BASED DIFFUSION MAPS

In this section we motivate the efficacy of the GDM approach with an empirical study of predictive accuracy in the semi-supervised setting on six benchmark data sets found in Chapelle et al. (2006). In the semi-supervised setting using the labeled as well as the unlabeled data for dimension reduction followed by fitting a regression model has often increased predictive accuracy. We used the benchmark data so we could compare the performance of DM and GDM to eleven algorithms (Chapelle et al., 2006, Table 21.11). The conclusion of our study is that GDM improves predictive accuracy over DM and that GDM is competitive with respect to the other algorithms.

For each data set twelve splits were generated with 100 samples labeled in each split. We applied DM and GDM to each of these sets to find DR directions. We projected the data (labeled and unlabeled) onto the DR directions and used a k-Nearest-Neighbor (kNN) classifier to classify the unlabeled data. The parameters of the DM, GDM, and number of neighbors were set using a validation set in each trial. The average classification error rate for the unlabeled data over the twelve splits are reported in Table 1. We also report in Table 1 the top performing algorithm for the data sets in Chapelle et al. (2006, Table 21.11). Laplacian RLS stands for Laplacian regularized least-squares, SGT stands for spectral graph transducer, Cluster-Kernel is an algorithm that uses two kernel functions, see (Chapelle et al., 2006, Chapter 11) for details.

A reasonable conclusion from Table 1 is that having label information improves the performance of diffusion operator with respect to prediction. In addition, dimension reduction using GDM followed by a simple classifier is competitive to other approaches. We suspect that integrating GDM with a penalized classification algorithm in the same spirit as Laplacian regularized least-squares can improve performance.

5. Graphical Models and Conditional Independence

One example of a statistical analysis where global inferences are desired or explanations with respect to the coordinates of the data is important is a graphical model over undirected graphs. In this setting it is of interest to understand how coordinates covary with respect to variation in response, as is provided by the GOP. Often of greater interest is to infer direct or conditional dependencies between two coordinates as a function of variation in the response. In this section we explore how this can be done using the GOP.

A natural idea in multivariate analysis is to model the conditional independence of a multivariate distribution using a graphical model over undirected graphs. The theory of Gauss-Markov graphs

Data	DM	GDM	Best
G241C	19.96%	18.61%	13.49% (Cluster-Kernel)
G241D	14.27%	13.64%	4.95% (Cluster-Kernel)
Digit1	1.8%	1.8%	1.8% (DM, GDM)
BCI	48.53%	31.36%	31.36% (GDM, Laplacian RLS)
USPS	12.85%	10.76%	4.68% (Laplacian RLS)
Text	24.71%	23.57%	23.09% (SGT)

Table 1: Error rates for DM and GDM over six data sets reported in Chapelle et al. (2006, Table 21.11). The column 'Best' reports the error rate for the algorithm with the smallest error of the 13 applied to the data.

(Speed and Kiiveri, 1986; Lauritzen, 1996) was developed for multivariate Gaussian densities

$$p(x) \propto \exp\left(-\frac{1}{2}x^T Jx + h^T x\right),$$

where the covariance is J^{-1} and the mean is $\mu = J^{-1}h$. The result of the theory is that the precision matrix J, given by $J = \Sigma_X^{-1}$, provides a measurement of conditional independence. The meaning of this dependence is highlighted by the partial correlation matrix R_X where each element R_{ij} is a measure of dependence between variables i and j conditioned on all other variables $S^{/ij}$ and $i \neq j$

$$R_{ij} = \frac{\text{cov}(X_i, X_j \mid S^{/ij})}{\sqrt{\text{var}(X_i \mid S^{/ij})}\sqrt{\text{var}(X_j \mid S^{/ij})}}.$$

The partial correlation matrix is typically computed from the precision matrix J

$$R_{ij} = -J_{ij}/\sqrt{J_{ii}J_{jj}}.$$

In the regression and classification framework inference of the conditional dependence between explanatory variables has limited information. Much more useful would be the conditional dependence of the explanatory variables conditioned on variation in the response variable. In Section 2 we stated that both the covariance of the inverse regression as well as the gradient outer product matrix provide estimates of the covariance of the explanatory variables conditioned on variation in the response variable. Given this observation, the inverses of these matrices

$$J_{X|Y} = \Omega_{X|Y}^{-1} \quad \text{and} \quad J_\Gamma = \Gamma^{-1},$$

provide evidence for the conditional dependence between explanatory variables conditioned on the response. We focus on the inverse of the gradient outer product matrix in this paper since it is of use for both linear and nonlinear functions.

The two main approaches to inferring graphical models in high-dimensional regression have been based on either sparse factor models (Carvalho et al., 2008) or sparse graphical models representing sparse partial correlations (Meinshausen and Buhlmann, 2006). Our approach differs from both of these approaches in that the response variable is always explicit. For sparse factor models

the factors can be estimated independent of the response variable and in the sparse graphical model the response variable is considered as just another node, the same as the explanatory variables. Our approach and the sparse factor models approach both share an assumption of sparsity in the number of factors or directions. Sparse graphical model approaches assume sparsity of the partial correlation matrix.

Our proof of the convergence of the estimated conditional dependence matrix $(\hat{\Gamma})^{-1}$ to the population conditional dependence matrix Γ^{-1} relies on the assumption that the gradient outer product matrix being low rank. This again highlights the difference between our modeling assumption of low rank versus sparsity of the conditional dependence matrix. Since we assume that both Γ and $\hat{\Gamma}$ are singular and low rank we use pseudo-inverses in order to construct the dependence graph.

Proposition 13 *Let Γ^{-1} be the pseudo-inverse of Γ. Let the eigenvalues and eigenvectors of $\hat{\Gamma}$ be $\hat{\lambda}_i$ and $\hat{v}_i$ respectively. If $\varepsilon > 0$ is chosen so that $\varepsilon = \varepsilon_n = o(1)$ and $\varepsilon_n^{-1}\|\hat{\Gamma} - \Gamma\| = o(1)$, then the convergence*

$$\sum_{\hat{\lambda}_i > \varepsilon} \hat{v}_i \hat{\lambda}_i^{-1} \hat{v}_i \xrightarrow{n\to\infty} \Gamma^{-1}$$

holds in probability.

Proof We have shown in Proposition 9 that $\|\hat{\Gamma} - \Gamma\| = o(1)$. Denote the eigenvalues and eigenvectors of Γ as λ_i and v_i respectively. Then

$$|\hat{\lambda}_i - \lambda_i| = O(\|\hat{\Gamma} - \Gamma\|) \qquad \text{and} \qquad \|\hat{v}_i - v_i\| = O(\|\hat{\Gamma} - \Gamma\|).$$

By the condition $\varepsilon_n^{-1}\|\hat{\Gamma} - \Gamma\| = o(1)$ the following holds

$$\hat{\lambda}_i > \varepsilon \Longrightarrow \lambda_i > \varepsilon/2 \Longrightarrow \lambda_i > 0$$

implying $\{i : \hat{\lambda}_i > \varepsilon\} \subset \{i : \lambda_i > 0\}$ in probability. On the other hand, denoting $\tau = \min\{\lambda_i : \lambda_i > 0\}$, the condition $\varepsilon_n = o(1)$ implies

$$\{i : \lambda_i > 0\} = \{i : \lambda \geq \tau\} \subset \{i : \hat{\lambda}_i \geq \tau/2\} \subset \{i : \hat{\lambda}_i > \varepsilon\}$$

in probability. Hence we obtain

$$\{i : \lambda_i > 0\} = \{i : \hat{\lambda}_i > \varepsilon\}$$

in probability.

For each $j \in \{i : \lambda_i > 0\}$ we have $\lambda_j, \hat{\lambda}_j \geq \tau/2$ in probability, so

$$|\hat{\lambda}_j^{-1} - \lambda_j^{-1}| \leq |\hat{\lambda}_j - \lambda_j|/(2\tau) \xrightarrow{n\to\infty} 0.$$

Thus we finally obtain

$$\sum_{\hat{\lambda}_i > \varepsilon} \hat{v}_i \hat{\lambda}_i^{-1} \hat{v}_i \xrightarrow{n\to\infty} \sum_{\lambda_i > 0} v_i \lambda_i^{-1} v_i^T = \Gamma^{-1}.$$

This proves the conclusion. ■

5.1 Results on Simulated and Real Data

We first provide an intuition of the ideas behind our inference of graphical models using simple simulated data. We then apply the method to study dependencies in gene expression in the development of prostate cancer.

5.1.1 SIMULATED DATA

The following simple example clarifies the information contained in the covariance matrix as well as the gradient outer product matrix. Construct the following dependent explanatory variables from standard random normal variables $\theta_1,...,\theta_5 \overset{iid}{\sim} \mathcal{N}(0,1)$

$$X_1 = \theta_1, X_2 = \theta_1 + \theta_2, X_3 = \theta_3 + \theta_4, X_4 = \theta_4, X_5 = \theta_5 - \theta_4,$$

and the following response

$$Y = X_1 + (X_3 + X_5)/2 + \varepsilon_1,$$

where $\varepsilon_1 \sim \mathcal{N}(0,.5^2)$.

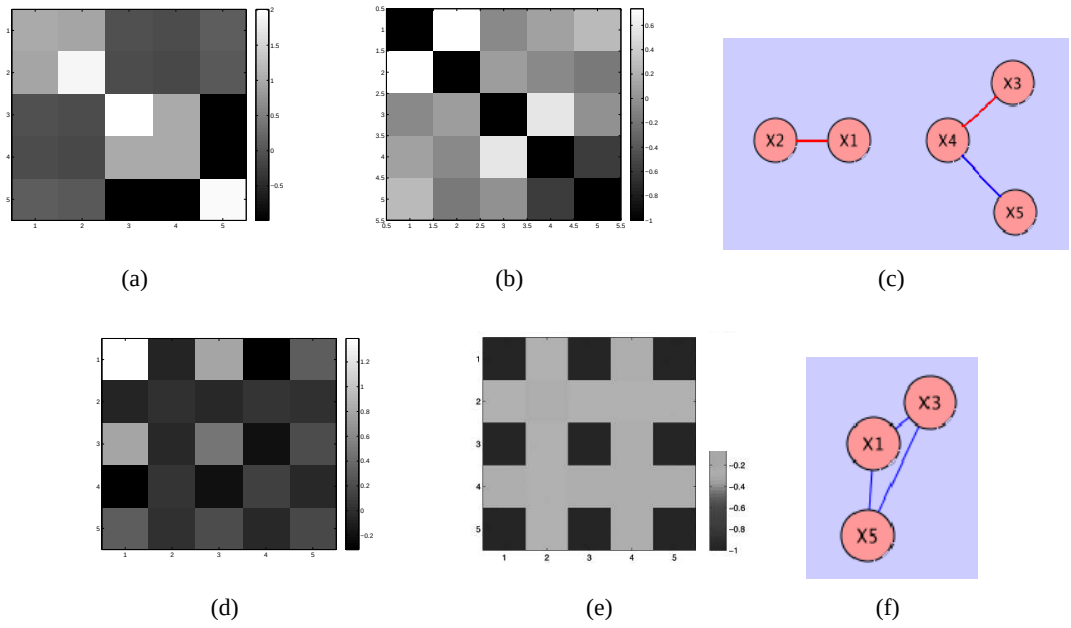
We drew 100 observations $(x_{i1},x_{i2},x_{i3},x_{i4},x_{i5},y_i)_{i=1}^{100}$ from the above sampling design. From this data we estimate the covariance matrix of the marginals $\hat{\Sigma}_X$ and the gradient outer product matrix $\hat{\Gamma}$. From $\hat{\Sigma}_X$, Figure 2(a), we see that X_1 and X_2 covary with each other and X_3, X_4, X_5 covary. The conditional independence matrix $\hat{R}_X$, Figure 2(b), provides information on more direct relations between the coordinates as we see that X_5 is independent of X_3 given X_4, $X_5 \perp\!\!\!\perp X_3 \mid X_4$. The dependence relations are summarized in the graphical model in Figure 2(c). Taking the response variable into account, we find in the gradient outer product matrix, Figure 2(d), the variables X_2 and X_4 are irrelevant while X_1, X_3, X_5 are relevant. The matrix $\hat{R}_{\hat{\Gamma}}$ is shown in Figure 2(e) and implies that any pair of X_1, X_3, X_5 are negatively dependent conditioned on the other and the response variable Y. The graphical model is given in Figure 2(f).

5.1.2 GENES DRIVING PROGRESSION OF PROSTATE CANCER

A fundamental problem in cancer biology is to understand the molecular and genetic basis of the progression of a tumor from less serious states to more serious states. An example is the progression from a benign growth to malignant cancer. The key interest in this problem is to understand the genetic basis of cancer. A classic model for the genetic basis of cancer was proposed by Fearon and Vogelstein (1990) describing a series of genetic events that cause progression of colorectal cancer.

In Edelman et al. (2008) the inverse of the gradient outer product was used to infer the dependence between genes that drive tumor progression in prostate cancer and melanoma. In the case of melanoma the data consisted of genomewide expression data from normal, primary, and metastatic skin samples. Part of the analysis in this paper was inference of conditional dependence graphs or networks of genes that drive differential expression between stages of progression. The gradient outer product matrix was used to infer detailed models of gene networks that may drive tumor progression.

In this paper, we model gene networks relevant in driving progression in prostate cancer as an illustration of how the methodology can be used to posit biological hypotheses. The objective is to understand the dependence structure between genes that are predictive of progression from benign to malignant prostate cancer. in progressing from benign to malignant prostate cancer. The data consists of 22 benign and 32 advanced prostate tumor samples (Tomlins et al., 2007; Edelman et al.,



(a) (b) (c)

(d) (e) (f)

Figure 2: (a) Covariance matrix $\hat{\Sigma}_X$; (b) Partial correlation matrix $\hat{R}_X$; (c) Graphical model representation of partial correlation matrix; (d) Gradient outer product matrix $\hat{\Gamma}$; (e) Partial correlations $\hat{R}_{\hat{\Gamma}}$ with respect to $\hat{\Gamma}$; (f) Graphical model representation of $\hat{R}_{\hat{\Gamma}}$.

2008). For each sample the expression level of over $12,000$ probes corresponding to genes were measured. We eliminated many of those probes with low variation across all samples resulting in a 4095 probes or variables. From this reduced data set we estimated the gradient outer product matrix, $\hat{\Gamma}$, and used the pseudo-inverse to compute the conditional independence matrix, $\hat{J} = (\hat{\Gamma})^{-1}$. From the conditional independence matrix we computed the partial correlation matrix $\hat{R}$ where $\hat{R}_{ij} = -\frac{\hat{J}_{ij}}{\sqrt{\hat{J}_{ii}\hat{J}_{jj}}}$ for $i \neq j$ and 0 otherwise. We again reduced the R matrix to obtain 139 nodes and 400 edges corresponding to the largest partial correlations and construct the graph seen in Figure 3.

The structure of the partial correlation graph recapitulates some known biological processes in the progression of prostate cancer. The most highly connected gene is MME (labeled green) which is known to have significant deregulation in prostate cancer and is associated with aggressive tumors (Tomlins et al., 2007). We also observe two distinct clusters annotated in yellow and purple in the graph that we call C_1 and C_2 respectively. These clusters derive their associations principally through 5 genes, annotated in light blue and dark blue in the graph. The light blue genes AMACR, ANXA1, and CD38 seem to have strong dependence with respect to the genes in C_1 wile C_2 is dependent on these genes in addition to the dark blue genes LMAN1L and SLC14A1. AMACR and ANXA1 as well as CD38 are well-known to have roles in prostate cancer progression (Jiang et al., 2004; Hsiang et al., 2004; Kramer et al., 1995). The other two genes LMAN1L and SLC14A1 are known to have tumorigenic properties and would be candidates for further experiments to better understand their role in prostate cancer.

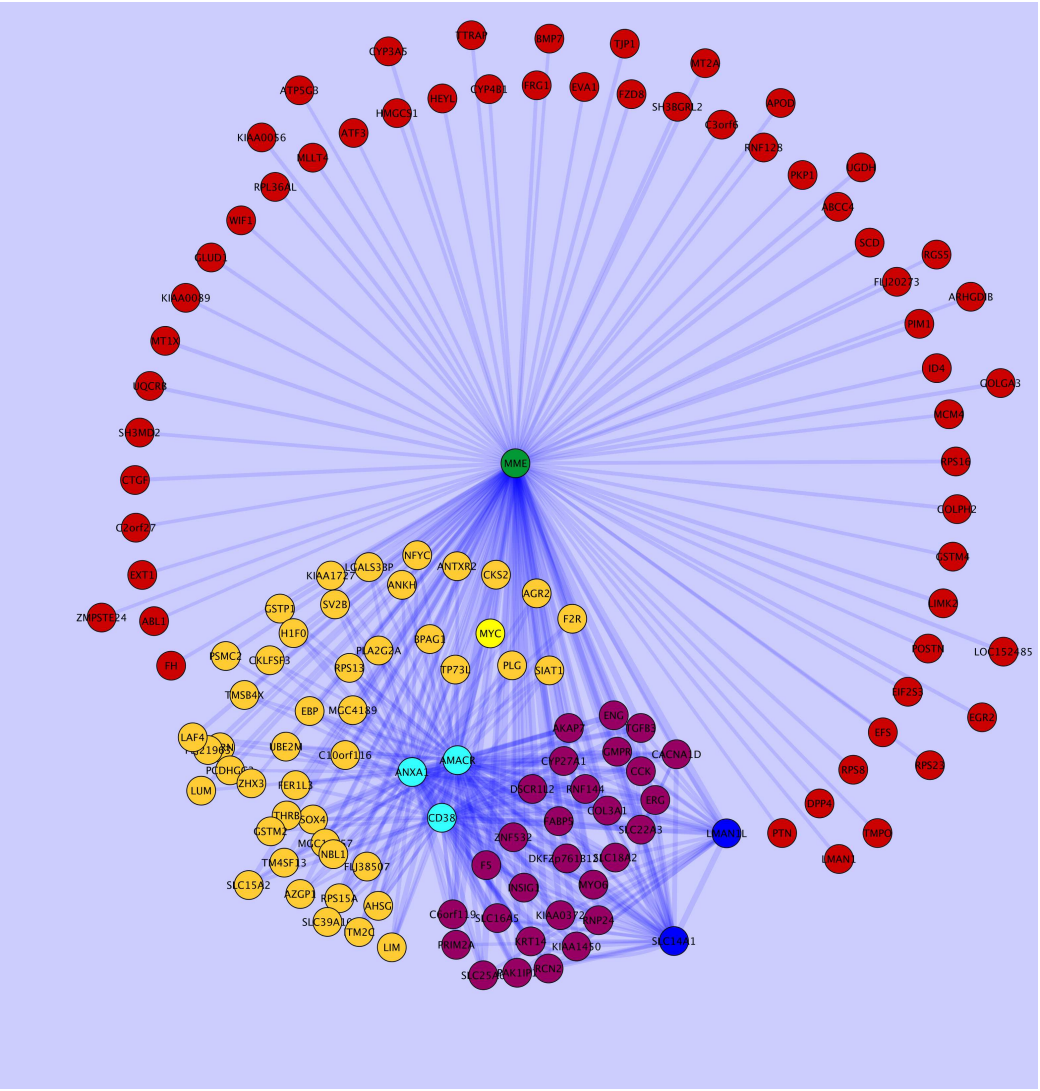


Figure 3: Graphical model of genes relevant in tumors progressing from benign to malignant prostate tissue. The edges correspond to partial correlations.

6. Discussion

The main contribution of this paper is to describe how inference of the gradient of the regression or classification function encodes information about the predictive geometry as well as the predictive conditional dependence in the data. Two methods are introduced gradient based diffusion maps and inference of conditional independence structure given gradient estimates. Precise statistical relations between different approaches to supervised dimension reduction are described. Simulated and real data are used to illustrate the utility of the methods developed. We prove convergence of the estimated graphical model to the population dependence graph. We find this direct link be-

tween graphical models and dimension reduction intriguing and suggest that the manifold learning perspective holds potential in the analysis and inference of graphical models.

Acknowledgments

This work was partially supported by NSF grant DMS-0732260, NIH Systems Biology Center Grant, and NIH R01 CA123175-01A1. MM is grateful for partial support from NSF (DMS 0650413, IIS 0803293), ONR N00014-07-1-0625, the Sloan Foundation and Duke.

References

R.J. Adcock. A problem in least squares. *The Analyst*, 5:53–54, 1878.

M. Belkin and P. Niyogi. Laplacian Eigenmaps for Dimensionality Reduction and Data representation. *Neural Computation*, 15(6):1373–1396, 2003.

M. Belkin and P. Niyogi. Semi-Supervised Learning on Riemannian Manifolds. *Machine Learning*, 56(1-3):209–239, 2004.

M. Belkin and P. Niyogi. Towards a theoretical foundation for Laplacian-based manifold methods. In *Learning Theory*, volume 3559 of *Lecture Notes in Comput. Sci.*, pages 486–500. Springer, Berlin, 2005.

C. Carvalho, J. Lucas, Q. Wang, J. Chang, J. Nevins, and M. West. High-dimensional sparse factor modelling - applications in gene expression genomics. *Journal of the American Statistical Association*, 103(484):1438–1456, 2008.

O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006. URL `http://www.kyb.tuebingen.mpg.de/ssl-book`.

R.R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21 (1):5–30, 2006.

R.R. Coifman and M. Maggioni. Diffusion wavelets. *Applied and Computational Harmonic Analysis*, 21(1):53–94, 2006.

R.R. Coifman, S. Lafon, A.B. Lee, M. Maggioni, B. Nadler, F. Warner, and S.W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences*, 102(21):7426–7431, 2005a.

R.R. Coifman, S. Lafon, A.B. Lee, M. Maggioni, B. Nadler, F. Warner, and S.W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Multiscale methods. *Proceedings of the National Academy of Sciences*, 102(21):7432–7437, 2005b.

R.D. Cook. Fisher lecture: Dimension reduction in regression. *Statistical Science*, 22(1):1–26, 2007.

R.D. Cook and S. Weisberg. Discussion of "Sliced inverse regression for dimension reduction". *J. Amer. Statist. Assoc.*, 86:328–332, 1991.

M. P. do Carmo. *Riemannian Geometry*. Birkhäuser, Boston, MA, 1992.

D. Donoho and C. Grimes. Hessian eigenmaps: new locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, 100:5591–5596, 2003.

N. Duan and K.C. Li. Slicing regression: a link-free regression method. *Ann. Statist.*, 19(2):505–530, 1991.

E.J. Edelman, J. Guinney, J-T. Chi, P.G. Febbo, and S. Mukherjee. Modeling cancer progression via pathway dependencies. *PLoS Comput. Biol.*, 4(2):e28, 2008.

F.Y. Edgeworth. On the reduction of observations. *Philosophical Magazine*, pages 135–141, 1884.

J. Fan and I. Gijbels. *Local Polynomial Modelling and its Applications*. Chapman and Hall, London, 1996.

E.R. Fearon and B. Vogelstein. A genetic model for colorectal tumorigenesis. *Cell*, 61:759–767, 1990.

R.A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Statistical Society A*, 222:309–368, 1922.

K. Fukumizu, F.R. Bach, and M.I. Jordan. Dimensionality reduction in supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5:73–99, 2005.

E. Giné and V. Koltchinskii. Empirical graph Laplacian approximation of Laplace-Beltrami operators: large sample results. In *High dimensional probability*, volume 51 of *IMS Lecture Notes Monogr. Ser.*, pages 238–259. Inst. Math. Statist., Beachwood, OH, 2006.

G.H. Golub and C.F. Va Loan. *Matrix Computations*. The Johns Hopkins University Press; 3rd edition, 1996.

T. Hastie and R. Tibshirani. Discriminant analysis by Gaussian mixtures. *J. Roy.Statist. Soc. Ser. B*, 58(1):155–176, 1996.

C-H. Hsiang, T. Tunoda, Y.E. Whang, D.R. Tyson, and D.K. Ornstein. The impact of altered annexin i protein levels on apoptosis and signal transduction pathways in prostate cancer cells. *The Prostate*, 66(13):1413–1424, 2004.

Z. Jiang, B.A. Woda BA, C.L. Wu, and X.J. Yang. Discovery and clinical application of a novel prostate cancer marker: alpha-methylacyl CoA racemase (P504S). *Am. J. Clin. Pathol*, 122(2): 275–8941, 2004.

G . Kramer, G. Steiner, D. Fodinger, E. Fiebiger, C. Rappersberger, S. Binder, J. Hofbauer, and M. Marberger. High expression of a CD38-like molecule in normal prostatic epithelium and its differential loss in benign and malignant disease. *The Journal of Urology*, 154(5):1636–1641, 1995.

S.L. Lauritzen. *Graphical Models*. Oxford: Clarendon Press, 1996.

K.C. Li. Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.*, 86:316–342, 1991.

K.C. Li. On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma. *Ann. Statist.*, 97:1025–1039, 1992.

N. Meinshausen and P. Buhlmann. High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34(2):1436–1462, 2006.

S. Mukherjee and Q. Wu. Estimation of gradients and coordinate covariation in classification. *J. Mach. Learn. Res.*, 7:2481–2514, 2006.

S. Mukherjee and DX. Zhou. Learning coordinate covariances via gradients. *J. Mach. Learn. Res.*, 7:519–549, 2006.

S. Mukherjee, D-X. Zhou, and Q. Wu. Learning gradients and feature selection on manifolds. *Bernoulli*, 16(1):181–207, 2010.

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

T. Speed and H. Kiiveri. Gaussian Markov distributions over finite graphs. *Ann. Statist.*, 14:138–150, 1986.

M. Sugiyama. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *J. Mach. Learn. Res.*, 8:1027–1061, 2007.

A. Szlam, M. Maggioni, and R. R. Coifman. Regularization on graphs with function-adapted diffusion process. *J. Mach. Learn. Res.*, 9:1711–1739, 2008.

M. Szummer and T. Jaakkola. Partially labeled classification with markov random walks. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14, pages 945–952, 2001.

J. Tenenbaum, V. de Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

S.A. Tomlins, R. Mehra, D.R. Rhodes, X. Cao, L. Wang, S.M. Dhanasekaran, S. Kalyana-Sundaram, J.T. Wei, M.A. Rubin, K.J. Pienta, R.B. Shah, and A.M. Chinnaiyan. Integrative molecular concept modeling of prostate cancer progression. *Nature Genetics*, 39(1):41–51, 2007.

G. Wahba. *Splines Models for Observational Data*. Series in Applied Mathematics, Vol. 59, SIAM, Philadelphia, 1990.

Q. Wu, F. Liang, and S. Mukherjee. Regularized sliced inverse regression for kernel models. Technical Report 07-25, ISDS, Duke Univ., 2007.

Y. Xia, H. Tong, W. Li, and L-X. Zhu. An adaptive estimation of dimension reduction space. *J. Roy.Statist. Soc. Ser. B*, 64(3):363–410, 2002.

L. Zwald and G. Blanchard. On the convergence of eigenspaces in kernel principal component analysi s. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 1649–1656. MIT Press, Cambridge, MA, 2006.