

# Learning gradients: predictive models that infer geometry and statistical dependence

**Qiang Wu**<sup>1,2,3</sup>

**Justin Guinney**<sup>3,4</sup>

**Mauro Maggioni**<sup>2,5</sup>

**Sayan Mukherjee**<sup>1,2,3</sup>

<sup>1</sup>*Department of Statistical Science*

<sup>2</sup>*Department of Computer Science*

<sup>3</sup>*Institute for Genome Sciences & Policy*

<sup>4</sup>*Program in Computational Biology and Bioinformatics*

<sup>5</sup>*Department of Mathematics*

*Duke University*

*Durham, NC 27708, USA*

QIANG@STAT.DUKE.EDU

JHG9@DUKE.EDU

MAURO.MAGGIONI@DUKE.EDU

SAYAN@STAT.DUKE.EDU

**Editor:**

## 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 and the gradient outer product. We relate the gradient outer product to standard statistical quantities such as covariances and provide a simple and precise comparison of a variety of simultaneous regression and 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. We illustrate the efficacy of the method of simulated and real data.

### Keywords:

Gradient estimates, manifold learning, graphical models, inverse regression, dimension reduction

## 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 data generated by measuring thousands of variables lies on or near a low-dimensional manifold. This hearkens to the central idea of reducing data to only relevant information that was fundamental in the paradigm of Fisher (1922) and goes back at least to Adcock (1878) and Edegworth (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 that are 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 idea of local estimates that is central to manifold learning.

The first main results in this paper are precise statistical relations between the three approaches. We will show that the gradient estimate is central to this analysis. Our second main result is the inference of graphical models based on gradient estimates. We provide rates of convergence of the estimated graphical model to its population counterpart. These rates and the underlying modeling depend not on the sparsity of the graph but on the rank of the conditional independence matrix or the intrinsic dimension of the gradient on the manifold supporting the data. Salient properties of the model and the efficacy of the gradients approach is illustrated on simulated and real data.

## 2. A statistical foundation for learning gradients

The problem of regression can be summarized as estimating the regression function

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

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 i.i.d. from a joint distribution,  $L_i \stackrel{i.i.d.}{\sim} \rho(X, Y)$ . When  $p$  is large the response variable  $Y$  may depend on a few directions in  $\mathbb{R}^p$ ,

$$Y = f_r(X) + \varepsilon = g(b_1^T X, \dots, b_d^T X) + \varepsilon, \quad (1)$$

where  $\varepsilon$  is noise and  $B = (b_1^T, \dots, b_d^T)$  is the effective dimension reduction (EDR) space. In this case 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.

### 2.1 Gradient outer product matrix and dimension reduction

The central concept of this paper is the gradient outer product matrix. Assume the regression function  $f_r(x)$  is smooth, the gradient is given by

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

and the the gradient outer product matrix  $\Gamma$  is a  $p \times p$  matrix with elements

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

where  $\rho_X$  is the marginal distribution of the explanatory variables. Using the notation  $a \otimes b = ab^T$  for  $a, b \in \mathbb{R}^p$ , we can write

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

The relation between the gradient outer product matrix and dimension reduction is illustrated by the following observation.

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

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

Lemma 1 states the effective dimension reduction space can be computed by a spectral decomposition of  $\Gamma$ . This motivates dimension reduction methods based on consistent estimators  $\Gamma_D$  of  $\Gamma$  given data  $D$ . Several methods have been motivated by this idea either implicitly (e.g. MAVE in Xia et al. (2002)) or explicitly (e.g. OPG in Xia et al. (2002) and learning gradients in Mukherjee et al. (2006)).

## 2.2 Gradient outer product matrix as a covariance matrix

The gradient outer product matrix is defined globally and its relation to dimension reduction in Section 2.1 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.

Another concept central to dimension reduction is the covariance matrix of the inverse regression function. In Li (1991) the idea of dimension reduction using inverse regression was explored. The central quantity in this approach was the covariance matrix of the inverse regression  $\Omega_{X|Y} = \text{cov}_Y[\mathbb{E}_X(X|Y)]$  which under certain conditions encodes the EDR directions  $B = (b_1, \dots, b_d)$ .

The main result of this subsection is to relate the two matrices:  $\Gamma$  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 inherently takes local information into account and thus contains more information and is more central to dimension reduction than the covariance of the inverse regression. 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

$$y = \beta^T x + \varepsilon, \quad \mathbb{E} \varepsilon = 0. \quad (3)$$

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

**Proposition 2** *Suppose (3) holds. Given the covariance of the inverse regression,  $\Omega_{X|Y} = \text{cov}_Y(\mathbb{E}_X(X|Y))$ , 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*

$$\Gamma = \sigma_Y^2 \left(1 - \frac{\sigma_\varepsilon^2}{\sigma_Y^2}\right)^2 \Sigma_X^{-1} \Omega_{X|Y} \Sigma_X^{-1}, \quad (4)$$

assuming that  $\Sigma_X$  is full rank.

The above result states that the matrices  $\Gamma$  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. We argue that the gradient outer product matrix is of greater importance since it contains more information than the covariance of the inverse regression. It is well known (Li, 1991; Duan and Li, 1991) that  $\Omega_{X|Y}$  contains information about the predictive direction  $\beta/\|\beta\|$  but  $\Gamma$  also reflects the importance of this direction weighted by the variance of the output variable  $y$ .

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

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

such that in each region  $R_i$  the regression function  $f_r$  is linear

$$f_r(x) = \beta_i^T x + \varepsilon_i, \quad \mathbb{E} \varepsilon_i = 0 \quad \text{for } x \in R_i. \quad (5)$$

The following corollary is true.

**Corollary 3** *Given partitions  $R_i$  of the input space for which (5) holds with  $\mathbb{E} \varepsilon_i = 0$ , define in each partition  $R_i$  the following local quantities: the covariance of the input variables  $\Sigma_i = \text{cov}(X \in R_i)$ , the covariance of the inverse regression  $\Omega_i = \text{cov}(\mathbb{E}(X \in R_i|Y))$ , the variance of the output variable  $\sigma_i^2 = \text{var}(Y|X \in R_i)$ . Assuming that matrices  $\Sigma_i$  are full rank, the gradient outer product matrix can be computed in terms of these local quantities*

$$\Gamma = \sum_{i=1}^{\mathcal{I}} \rho_X(R_i) \sigma_i^2 \left(1 - \frac{\sigma_{\varepsilon_i}^2}{\sigma_i^2}\right)^2 \Sigma_i^{-1} \Omega_i \Sigma_i^{-1}, \quad (6)$$

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

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 equation (6) holds approximately (i.e.,  $\mathbb{E} \varepsilon_i \approx 0$ ), then

$$\Gamma \approx \sum_{i=1}^{\mathcal{I}} \rho_X(R_i) \sigma_i^2 \left(1 - \frac{\sigma_{\varepsilon_i}^2}{\sigma_i^2}\right)^2 \Sigma_i^{-1} \Omega_i \Sigma_i^{-1}, \quad (7)$$

This illustrates the centrality of the gradient outer product in the following sense: it contains not only information on all the predictive directions but also their importance by weighting

them with respect to the variance of the output variables. It is well known the covariance of the inverse regression may contain only partial information on predictive directions and in degenerate cases where  $\mathbb{E}(X|Y) = 0$  and hence  $\Omega_{X|Y} = \mathbf{0}$  contains no information.

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 on 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  $\varphi : \mathcal{M} \rightarrow \mathbb{R}^p$  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 such as  $\Omega_{X|Y}$  are not as meaningful from a modeling perspective. Instead, all one can expect is inference of local structure. The gradient outer product matrix defined in terms the gradient on the manifold

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

is still meaningful from a modeling perspective because gradients on the manifold are defined by local structure. Note that the  $d_{\mathcal{M}} \times d_{\mathcal{M}}$  matrix  $\Gamma_{\mathcal{M}} = \nabla_{\mathcal{M}}f_r \otimes \nabla_{\mathcal{M}}f_r$  has a central meaning in our problem. However, we know neither the manifold nor the coordinates on the manifold but are only provided points in the ambient space. For this reason we cannot compute  $\Gamma_{\mathcal{M}}$  but we can understand its properties by analyzing the gradient outer product matrix  $\Gamma$  in the ambient space, a  $p \times p$  matrix. Details on conditions under which  $\Gamma$  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  $\Gamma$ . Many approaches for computing gradients exist including various numerical derivative algorithms, local linear smoothing (Fan and Gijbels, 1996), and learning gradients by kernel models (Mukherjee and Zhou, 2006; Mukherjee and Wu, 2006). The application domain we focus on is the analysis of high-dimensional data with few observations, where  $p \gg n$ . 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. (2006). 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

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

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

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

The objective of learning gradients is given data  $D = \{(y_i, x_i)\}_{i=1}^n$  simultaneously estimate the regression function  $f_r$  by a function  $f_D$  and the gradient  $\nabla f_r$  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  $\vec{f} = (f_1, \dots, f_p)$  on  $D$  as

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

The regression function and gradient estimate is modeled by

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

where  $f_D$  and  $\vec{f}_D$  are estimates of  $f_r$  and  $\nabla f_r$  given the data,  $w_{i,j}^s$  is a weight function with bandwidth  $s$ ,  $\|\cdot\|_K$  is the reproducing kernel Hilbert space (RKHS) norm,  $\lambda_1$  and  $\lambda_2$  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

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

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

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

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  $\lambda_1$  and  $\lambda_2$  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., 2006).

**Definition 6** *Let  $\mathcal{M}$  be a Riemannian manifold and  $\varphi : \mathcal{M} \rightarrow \mathbb{R}^p$  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*

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

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

$$f_D(x) = \sum_{i=1}^n \alpha_{i,D} K(x, x_i), \quad \vec{f}_D(x) = \sum_{i=1}^n c_{i,D} K(x, x_i), \quad (8)$$

with  $c_D = (c_{1,D}, \dots, c_{n,D}) \in \mathbb{R}^{p \times n}$ , and  $\alpha_D = (\alpha_{1,D}, \dots, \alpha_{n,D})^T \in \mathbb{R}^p$ . 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  $1 - \delta$ ,*

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

Consistency in the manifold setting was studied in Mukherjee et al. (2006) 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. (2006) for details), with probability greater than  $1 - \delta$ ,*

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

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 product matrix was developed in Section 2.1. The estimate of the gradient obtained by the kernel models in Section 3 provides the following empirical estimate of the gradient outer product matrix

$$\hat{\Gamma} = c_D K^2 c_D^T = \frac{1}{n} \sum_{i=1}^n \vec{f}_D(x_i) \otimes \vec{f}_D(x_i), \quad (9)$$

where  $K$  is kernel matrix with  $K_{ij} = K(x_i, x_j)$ . The eigenvectors corresponding to the top eigenvalues of  $\hat{\Gamma}$  can be used to estimate the  $d$  effective 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  $\nabla f$ . Let  $\hat{v}_1, \dots, \hat{v}_d$  be the eigenvectors of  $\hat{\Gamma}$  associated to the top  $d$  eigenvalues. The following holds*

$$\text{span}(\hat{v}_1, \dots, \hat{v}_d) \longrightarrow \text{span}(B).$$

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

**Proof** By Proposition 7  $\hat{\Gamma}_{ij} \rightarrow \Gamma_{ij}$  and hence  $\hat{\Gamma} \rightarrow \Gamma$  in matrix norm. By perturbation theory, the eigenvalues and eigenvectors of  $\hat{\Gamma}$  converge to the eigenvalues and eigenvectors of  $\Gamma$  respectively. The conclusions then follows from Lemma 1.  $\blacksquare$

Proposition (9) is a justification of 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 EDR directions using a generalized eigen-decomposition problem

$$\Omega_{x|Y} \beta = \nu \Sigma_X \beta. \quad (10)$$

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

We start with a simple model where the EDR 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  $\Sigma_X$  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  $\Sigma_X$  is the covariance matrix of  $X$ . In this case, the computation of matrix  $\Gamma$  is direct:

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

By the assumption  $\Sigma_X$  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

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

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  $\Gamma$  and  $\Omega_{X|Y}$  differ only up to a linear transformation. As a consequence the generalized eigen-decomposition (10) of  $\Omega_{X|Y}$  with respect to  $\Sigma_X$  yields the same first direction as the eigen-decomposition of  $\Gamma$ .

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  $\Gamma$  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  $\Gamma$  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  $y$ .

When there are more than one EDR directions as in model (1), we partition the input space into  $\mathcal{I}$  small regions  $X = \bigcup_{i=1}^{\mathcal{I}} R_i$  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

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

where  $\Gamma_i$  is the gradient outer product matrix on  $R_i$  and  $\Omega_i = \text{cov}(\mathbb{E}(X \in R_i|Y))$ . In this sense, the gradient covariance matrix  $\Gamma$  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,  $\varphi : \mathcal{M} \rightarrow \mathbb{R}^p$ . 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., 2006). Generally this is not true when the manifold is nonlinear,  $\varphi$  is a nonlinear map. Instead, the estimate provides the following information about  $\nabla_{\mathcal{M}} f_r$ :

$$(d\varphi)^* \vec{f}_D \longrightarrow \nabla_{\mathcal{M}} f_r \quad \text{as} \quad n \rightarrow \infty,$$

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

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  $\nabla f_r$  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 EDR directions from the matrix  $\hat{\Gamma}$ .

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

**Proposition 11** *If  $v \perp b_i$  for all  $i = 1, \dots, d$ , then  $v^T \hat{\Gamma} v \rightarrow 0$ .*

**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  $v \cdot f_\lambda = 0$ .

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

$$v^T \hat{\Xi} v = \|v \cdot \vec{f}_D\|_K^2 \rightarrow \|v \cdot \vec{f}_\lambda\|_K^2 = 0.$$

This implies that  $v^T \hat{\Gamma} v \rightarrow 0$  and proves the proposition. ■

Proposition 11 states that all the vectors perpendicular to the EDR space correspond to eigenvalues near zero of  $\hat{\Gamma}$  and will be filtered out. This means the EDR 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)

We have focused on linear projections since these are implied by the semi-parametric model in (1). In this section we adapt the gradient learning method to nonlinear projections based on diffusions on graphs. The 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). Random walks on graphs have been used in many dimension reduction methods (Belkin and Niyogi, 2003, 2004; Szummer and Jaakkola, 2001; Coifman et al., 2005a,b).

The central quantity in all of these approaches is a diffusion operator on the graph which we designate as  $L$ . This operator is constructed from a similarity matrix  $W$  representing a weighted undirected graph with the edges corresponding to the similarity between observations. 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  $t$ . Note that this approach is not operating in the space of the explanatory variables but in the space of observations and so the projection is onto a linear combination of observations.

In the case of unsupervised dimension reduction methods such as Laplacian eigenmaps (Belkin and Niyogi, 2003) similarity functions of the following form are used

$$W(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{\sigma_1}\right).$$

In the supervised setting Szlam et al. (2007) proposed the following data or function adapted similarity function

$$W_f(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{\sigma_1} - \frac{|f(x_i) - f(x_j)|^2}{\sigma_2}\right), \quad (11)$$

where the function evaluations  $f(x_i)$  are computed based on a 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, drawn from the marginal distribution, were used to learn the projection and a small set of labeled data 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.

Gradient estimates can be used to compute function adapted similarity functions. We call the diffusion maps defined on these similarity functions *gradient based diffusion maps (GDM)*. The labeled data is used to compute the gradient estimate,  $\vec{f}_D$ . Given the gradient estimate the pairwise differences between function values on all the labeled and unlabeled data is evaluated based on the following symmetrized estimate of the Taylor series

$$f(x_i) - f(x_j) \approx \frac{1}{2}(\nabla f(x_i) + \nabla f(x_j)) \cdot (x_i - x_j), \quad \text{for } x_i \approx x_j.$$

This evaluation is possible on both the labeled as well as unlabeled data because we can evaluate the gradient at any point. The differences are then used to evaluate equation (11).

In Appendix A we provide empirical comparisons that show that the GDM performs much better than the unsupervised diffusion maps with respect to classification accuracy on a benchmark data set. We also find the performance of the GDM is comparable to that of the function adapted diffusion maps in Szlam et al. (2007).

The feasibility of the gradient based diffusion map is based on the accuracy of the following first order Taylor expansion on the manifold

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

where  $v_{ij} \in T_{x_i}\mathcal{M}$  is the tangent vector such that  $x_j = \text{Exp}_{x_i}(v_{ij})$  where  $\text{Exp}_{x_i}$  is the exponential map at  $x_i$  (see do Carmo (1992); Mukherjee et al. (2006) for details). We cannot compute  $\nabla_{\mathcal{M}} f$  so use  $\vec{f}_D$  instead. The following proposition states that estimates of the function value differences between two can be accurately estimated from the 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  $x_i - x_j \approx d\varphi(v_{ij})$  we have

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

which implies the conclusion. ■

## 5. Graphical models and conditional independence

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 (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  $i \neq j$

$$R_{ij} = \frac{\text{cov}(x_i, x_j | S^{/ij})}{\sqrt{\text{var}(x_i | S^{/ij})} \sqrt{\text{var}(x_j | S^{/ij})}}.$$

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

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

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 response variables. Our approach and the sparse factor models approach both share an assumption of sparsity in the number of factors or directions but not of sparsity of the partial correlations between variables in the sparse graphical models approach.

Our proof of the convergence of the estimated conditional dependence matrix  $(\hat{\Gamma})^{-1}$  to the population conditional dependence matrix  $\Gamma^{-1}$  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  $\Gamma$  and  $\hat{\Gamma}$  are singular and low rank we use pseudo-inverses in order to construct the dependence graph.

**Proposition 13** *Let  $\Gamma^{-1}$  be the pseudo-inverse of  $\Gamma$ . 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*

$$\sum_{\hat{\lambda}_i > \varepsilon} \hat{v}_i \hat{\lambda}_i^{-1} \hat{v}_i \rightarrow \Gamma^{-1}$$

holds in probability.

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

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

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

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

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

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

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. Then  $|\hat{\lambda}_j^{-1} - \lambda_j^{-1}| \leq |\hat{\lambda}_j - \lambda_j|/(2\tau) \rightarrow 0$ . Thus we finally obtain

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

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  $\theta_1, \dots, \theta_5 \stackrel{iid}{\sim} \mathcal{N}(0, 1)$

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

and the following response

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

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

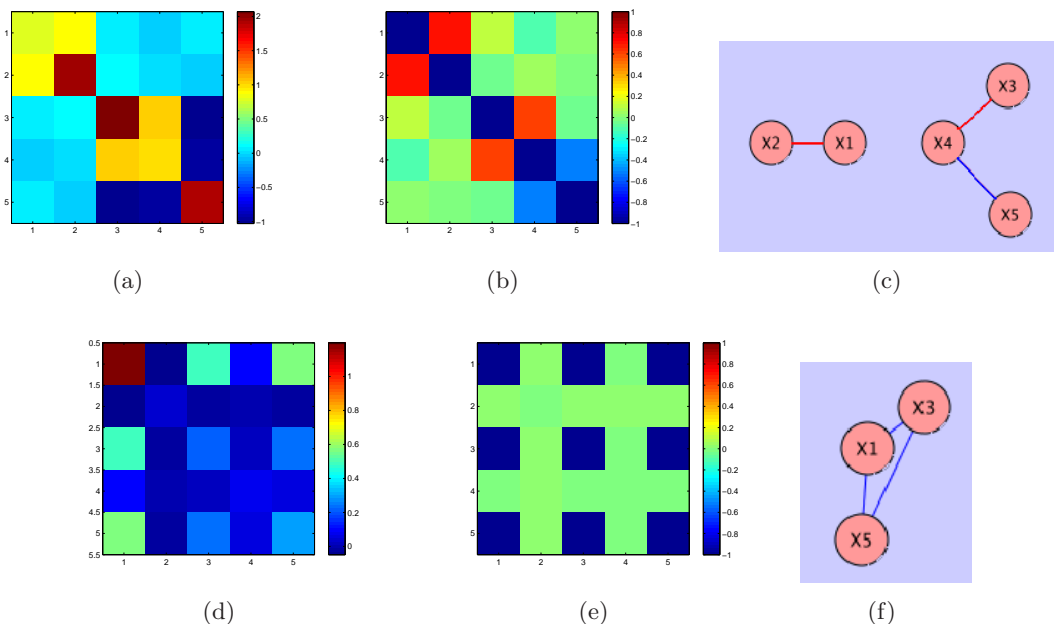


Figure 1: (a) Covariance matrix  $\hat{\Sigma}_X$ ; (b) Partial correlation matrix  $\hat{R}_X$ ; (c) Graphical model representation of partial correlation matrix; (d) Gradient outer product matrix  $\hat{\Gamma}$ ; (e) Partial correlations  $\hat{R}_{\hat{\Gamma}}$  with respect to  $\hat{\Gamma}$ ; (f) Graphical model representation of  $\hat{R}_{\hat{\Gamma}}$ .

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 1(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 1(b), provides information on more direct relations between the coordinates as we see that  $X_5$  is independent of  $X_3$  given  $X_4$ ,  $X_5 \perp\!\!\!\perp X_3 | X_4$ . The dependence relations are summarized in the graphical model in Figure 1(c). Taking the response variable into account, we find in the gradient outer product matrix (Figure 1(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 1(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 1(f).

### 5.1.2 GENES DRIVING PROGRESSION OF PROSTATE CANCER

A major motivation for the graphical model inference we propose is to infer dependencies between the expression of genes that are relevant to the appearance of a complex phenotype. This method was used to infer detailed biological models of tumor progression to infer gene dependencies or putative networks as well dependencies between gene sets or pathways in Edelman et al. (2008). We use progression in prostate cancer to illustrate how this method can be used to posit biological hypotheses.

Our objective will be to understand the dependence structure between genes conditioned on their differential expression in progressing from benign to malignant prostate cancer. The data consists of 22 benign and 32 advanced prostate tumor samples. For each sample the expression level of over 12,000 probes was measured where the probes correspond to genes. 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 2.

The structure of the partial correlation graph recapitulates some know 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  $C_1$  while  $C_2$  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.

## 6. Discussion

The two key ideas in this paper are a precise statistical interpretation of learning gradients which we use to relate various supervised dimension reduction methods and the inference of graphical models based on gradient estimates. We focus mainly on linear projections and explain why linear projections can be appropriate even in the nonlinear manifold setting. A brief explanation of how gradient estimates can be used for nonlinear projections is also provided. We use simulated and real data illustrate the utility of our approach in the inference of conditional dependence graphs for predictive explanatory variables given data. We also prove convergence of the estimated graphical model to the population dependence graph. We find this direct link between graphical models and dimension reduction intriguing and suggest that the manifold learning perspective holds potential in the analysis and inference of graphical models.

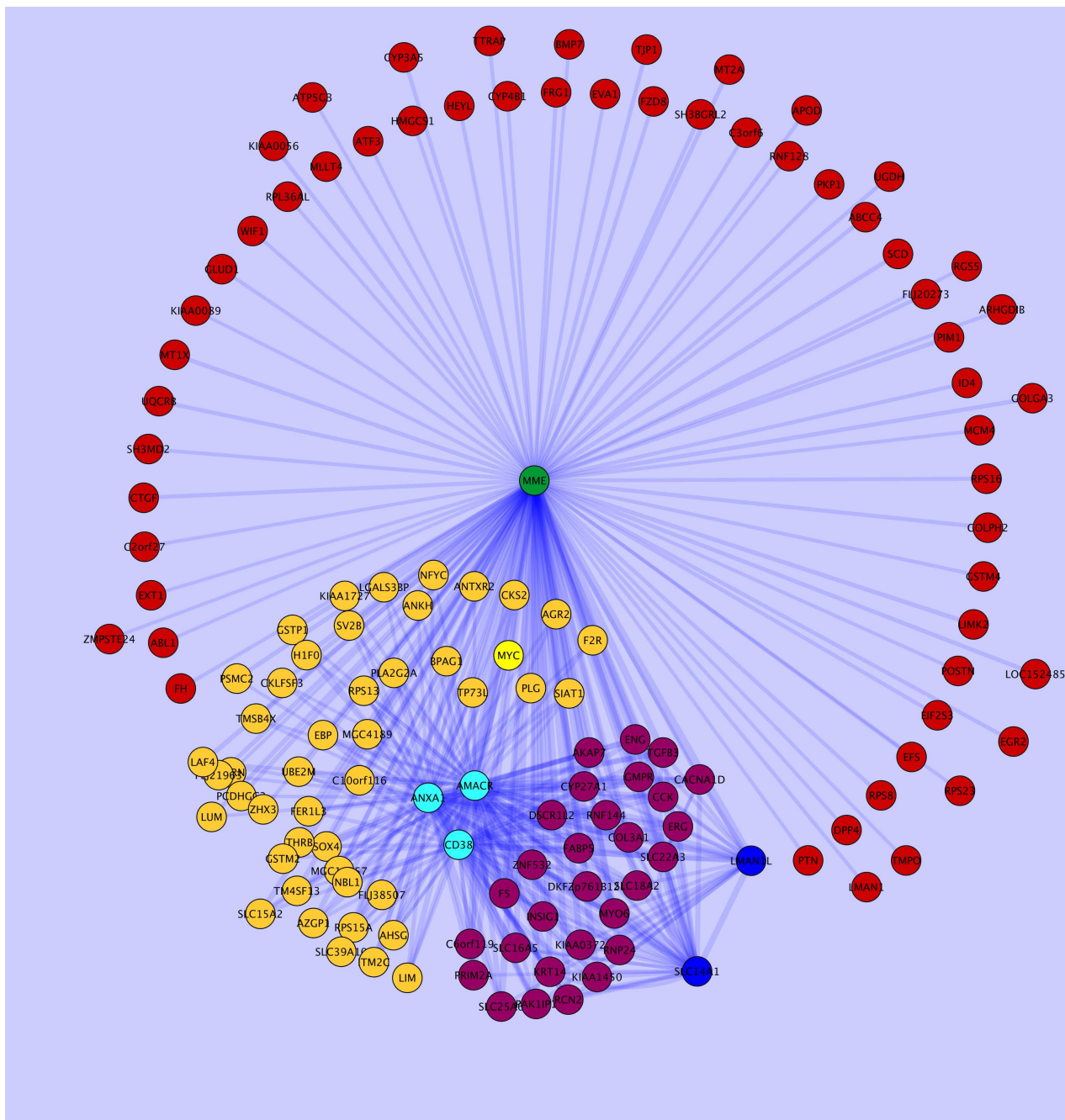


Figure 2: Graphical model of genes relevant in tumors progressing from benign to malignant prostate tissue. The edges correspond to partial correlations.

## Appendix A. Empirical results for gradient based diffusion maps

We show that having label information improves the performance of diffusion operator with respect to prediction by comparing nonlinear dimension reduction using the gradient based diffusion map (GDM) to unsupervised diffusion maps (DM). The comparison is on a subset of the MNIST digits data set (Y. LeCun, <http://yann.lecun.com/exdb/mnist/>). The data set contains 60,000 images of handwritten digits  $\{0, 1, 2, \dots, 9\}$ , where each image consists of  $p = 28 \times 28 = 784$  gray-scale pixel intensities. We will focus on discriminating a “6” from a “9”, the “69” data set, and “3” from a “8”, the “38” data set. The first is one of the more difficult pairwise comparisons.

For both data sets the following procedure was repeated 50 times. Randomly select 1000 samples from each of the two digits and then provide label information for a few samples. Then we apply GDM and DM for dimension reduction. For DM, we use the labeled training data as well as the unlabeled test data and the kernel (11) to build the graph where the kernel parameter  $\sigma_1$  is self-tuned according to Szlam et al. (2007). For GDM we first learn the gradient on the labeled training data and then build the graph using the function adapted kernel with the parameter  $\sigma_2$  chosen by cross-validation. We also use cross-validation to select dimensions, between 1 – 5. We then use the 5-nearest neighbor algorithm on the low dimensional training data to classify the independent test data and compute the error rate. The average over the 50 iterations is displayed in Tables 1. The results indicate that the response dependent method always outperforms the response independent methods, though the difference may be slight when the data is well separable (e.g. classifying “6” versus “9”).

	Labeled points	20	40	60	100
5 vs. 8	DM	8.05%	4.44%	4.15%	3.77%
	GDM	7.24%	3.91%	3.73%	3.43%
6 vs. 9	DM	1.75%	0.09%	0.05%	0.05%
	GDM	1.08%	0.09%	0.05%	0.06%

Table 1: Error rates for digits classification by GDM and DM.

## References

- R.J. Adcock. A problem in least squares. *The Analyst*, 5:53–54, 1878.
- M. Belkin and P. Niyogi. Semi-Supervised Learning on Riemannian Manifolds. *Machine Learning*, 56(1-3):209–239, 2004.
- M. Belkin and P. Niyogi. Laplacian Eigenmaps for Dimensionality Reduction and Data representation. *Neural Computation*, 15(6):1373–1396, 2003.
- 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*, 2008.

- 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*, page in press, 2007.
- R.D. Cook and S. Weisberg. Discussion of "sliced inverse regression for dimension reduction". *J. Amer. Statist. Assoc.*, 86:328–332, 1991.
- Manfredo P. do Carmo. *Riemannian Geometry*. Birkhäuser, Boston, MA, 1992.
- D. Donoho and C. Grimes. Hessian eigenmaps: new locally linear embedding techniques for highdimensional 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.
- F.Y. Edegworth. On the reduction of observations. *Philosophical Magazine*, pages 135–141, 1884.
- 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.
- J. Fan and I. Gijbels. *Local Polynomial Modelling and its Applications*. Chapman and Hall, London, 1996.
- 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.
- 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: Clarendo Press, 1996.
- K.C. Li. Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.*, 86: 316–342, 1991.
- 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. Technical Report 06-20, ISDS, Duke Univ., 2006.
- 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.
- Masashi 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.*, 2007. accepted.
- Martin Szummer and Tommi 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. Kalyanasundaram, 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.