

Learning gradients: predictive models that infer geometry and dependence

Qiang Wu^{1,2,3}

Justin Guinney^{3,4}

Mauro Maggioni^{2,5}

Sayan Mukherjee^{1,2,3}

¹*Department of Statistical Science*

²*Department of Computer Science*

³*Institute for Genome Sciences & Policy*

⁴*Program in Computational Biology and Bioinformatics*

⁵*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

This paper develops and discusses a modeling framework called learning gradients that allows for predictive models that simultaneously infer the geometry and statistical dependencies of the input space relevant for prediction. The geometric relations addressed in this paper hold for Euclidean spaces as well as the manifold setting. The central quantity in this framework is an estimate of the gradient of a regression or classification function, which is computed by a discriminative approach. We relate the gradient to the problem of inverse regression which in the machine learning community is typically addressed by generative models. A result of this relation is a simple and precise comparison of a variety of simultaneous regression and dimensionality reduction methods from the statistics literature. The gradient estimate is applied to a variety of problems central to machine learning: variable selection, linear and nonlinear dimension reduction, and the inference of a graphical model of the dependencies of the input variables that are relevant to prediction.

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).

The modern reprise of this program in the machine learning literature has been the idea of “manifold learning”. This has given rise to a variety of algorithms: isometric feature mapping (ISOMAP) (Tenenbaum et al., 2000), local linear embedding (LLE) (Roweis and Saul, 2000), Hessian Eigenmaps (Donoho and Grimes, 2003), and Laplacian Eigenmaps (Belkin and Niyogi, 2003) are all formulated to estimate a low-dimensional underlying manifold from high-dimensional sparse input or explanatory variables. However, these approaches are unsupervised and so do not use output or response variates in the models or algorithms and hence may be suboptimal with respect to predicting response. In statistics the ideas developed in sliced inverse regression (SIR) (Li, 1991), Kernel dimensionality reduction (KSIR) (Fukumizu et al., 2005), (conditional) minimum average variance estimation (MAVE) (Xia et al., 2002), and sliced average variance estimation (SAVE) (Cook and Weisberg, 1991) consider simultaneous dimensionality reduction and regression. The response variates are taken into account and the focus is on linear subspaces. These approaches do not extend to the manifold paradigm. In series of papers Mukherjee and Zhou (2006); Mukherjee and Wu (2006); Mukherjee et al. (2006), the method of learning gradients was developed to allow for simultaneous dimension reduction and regression in the manifold setting.

In this paper we provide a general statistical framework based on learning gradients that simultaneously infers a predictive model and estimates the geometry and statistical dependence of the input variables relevant to prediction.

In Section 2 we provide a statistical foundation for the centrality of the gradient estimate and its relation to inverse regression. The central quantity in this theory will be the gradient outer product (GOP) and this theory will provide a comparison of a variety of statistical methods for simultaneous dimension reduction and regression (Li, 1991; Xia et al., 2002; Cook and Weisberg, 1991; Li, 1992). An algorithm for estimating the gradient is stated in Section 3. The remaining sections illustrate how the GOP can be used for a variety of machine learning tasks. In Section 4 we discuss variable selection and feature construction. For feature construction we describe methods for linear and nonlinear dimensionality reduction. The nonlinear methods are based on a local notion of the GOP and diffusion maps (Coifman et al., 2005a,b; Szlam et al., 2007). In Section 5 the GOP is used to infer a graphical model of dependencies between the input variables that are relevant in predicting the response variable. The methods developed are applied to simulated data as well as real data. Linear and nonlinear dimension reduction was applied to the classification of handwritten digits and graphical models were inferred on gene expression data providing putative gene networks. We close with a discussion.

2. A statistical foundation for learning gradients

The standard regression problem considers data $D = \{L_i = (Y_i, X_i)\}_{i=1}^n$ where X_i is an input variable 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, $L_i \stackrel{i.i.d.}{\sim} \rho(X, Y)$.

The two standard settings in the machine learning literature to model this data are generative models and discriminative or regression models. The discriminative approach stresses accurate predictions of the labels given the input $Y|X$. This is a regression approach. A common formulation of the generative approach is to model the distribution that generates the multivariate inputs given the output class, $X|Y$. Sometimes this approach is

called inverse regression. A reasonable summary of the above approaches is that generative models are richer in terms of the information they contain, however accurate inference is more difficult for generative models and therefore regression based approaches are preferred in prediction.

In this paper we will advocate the simultaneous estimation of the regression function

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

as well as the covariation of the inverse regression

$$\Omega_{X|Y} = \text{cov}(\mathbb{E}(X|Y))$$

as summary statistics. The regression provides a predictive model. The covariation of the inverse regression provides a great deal of information about the geometry as well as statistical dependencies of input variables relevant to prediction. We will see that an estimate of the covariation is very useful and we will not attempt to model the full distribution of the inverse regression.

The other concept central to this paper is given data $D = \{L_i = (Y_i, X_i)\}_{i=1}^n$ the simultaneous estimation of the regression function, $f_r(x)$, and its gradient $\nabla f_r = \left(\frac{\partial f_r}{\partial x^1}, \dots, \frac{\partial f_r}{\partial x^p}\right)^T$. Central to understanding the geometry and dependencies of the input data is the gradient outer product (GOP) matrix Γ with elements

$$\Gamma_{ij} = \left\langle \frac{\partial f_r}{\partial x^i}, \frac{\partial f_r}{\partial x^j} \right\rangle_{L^2_{\rho_X}}. \quad (1)$$

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 main result of this section is relating the two matrices Γ and $\Omega_{X|Y}$ to each other and explaining why the matrix Γ contains greater information and is of greater centrality in modeling relations between relevant variables. This is outlined for a linear setting and then generalized to nonlinear settings. Proofs of the propositions as well as the mathematical ideas underlying these results will be developed in Section 4.4.

The linear regression problem is typically stated as

$$y = \beta \cdot x + \epsilon, \quad \epsilon \sim \text{No}(0, \sigma_\epsilon^2). \quad (2)$$

In this case the following relation between gradient estimates and the inverse regression holds.

Proposition 1 *Suppose (2) holds. Given the covariance of the inverse regression, $\Omega_{X|Y} = \text{cov}(\mathbb{E}(X|Y))$, 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 GOP matrix is*

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

assuming that Σ_X is full rank.

The above result states that 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. We argue that the GOP 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 of the predictive direction $\beta/\|\beta\|$. But Γ also reflects the importance of this direction weighted by the variance of the output variable y .

We now generalize Proposition 1 to the general regression setting. The key idea will be that after partitioning the input space into small regions

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

the same relation holds approximately in each region between the covariance of inverse regression and the GOP. Assume that for each partition R_i the function is approximately linear by a Taylor expansion

$$f_r(x) = \beta_i \cdot x + \varepsilon_i, \quad \forall x \in R_i \quad (4)$$

where ε_i is a second order term in a Taylor expansion. This is always possible assuming the function is smooth.

In this case the following corollary is apparent.

Corollary 2 *Given partitions R_i of the input space for which (4) 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 $\Omega_i = \text{cov}(X \in R_i|Y)$, the variance of the output variable $\sigma_i^2 = \text{var}(Y|X \in R_i)$. Assuming that matrices Σ_i are full rank, the GOP can be defined in terms of these local quantities*

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

where $\rho_X(R_i)$ is the measure of partition R_i with respect to the marginal distribution ρ_X . In general (4) only holds approximately as does (5).

Corollary 2 illustrates the centrality of the GOP 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 usually contains only partial information on predictive directions and in degenerate cases (where $\mathbb{E}(X|Y) = 0$) may contain no information.

Estimates of either Γ or $\Omega_{X|Y}$ can be applied to the following three applications explored in this paper: dimension reduction, variable selection, and estimates of covariance of input variables that are predictive. In Section 4.4.2 we will see that for dimension reduction both Γ and $\Omega_{X|Y}$ can be used for dimension reduction in the linear setting. For nonlinear functions or when $\Omega_{X|Y}$ is degenerate the GOP is well defined and can be robustly estimated while inverse regression is problematic. For variable selection the GOP can be used but the inverse

regression cannot. The situation for estimating the covariation of relevant input variables is the same as dimension reduction, both Γ and $\Omega_{X|Y}$ can be used for dimension reduction in the linear setting and only the GOP has meaning for the nonlinear models or when $\Omega_{X|Y}$ is degenerate.

Methods presented in the statistics literature for simultaneous regression and dimension reduction can be divided into those estimating the GOP (Xia et al., 2002; Mukherjee et al., 2006) and those estimating the covariance of the inverse regression (Li, 1991; Cook and Weisberg, 1991). In SIR (Li, 1991) the covariance of the inverse regression is defined as the outer product of each (sliced) regression

$$\Omega_{X|Y} = \mathbb{E}((\mathbb{E}[X|Y] - \mathbb{E}X) \otimes (\mathbb{E}[X|Y] - \mathbb{E}X)).$$

In SAVE (Cook and Weisberg, 1991) the average variance estimate is considered and the following matrix

$$S_{X|Y} = \mathbb{E}(\text{cov}(X) - \text{cov}(X|Y))^2$$

is considered. Note that

$$\Omega_{X|Y} = \mathbb{E}(\text{cov}(X) - \text{cov}(X|Y)).$$

In the definition of S the square is taken to prevent degeneracy. In both MAVE (Xia et al., 2002) and learning gradients (Mukherjee et al., 2006) the GOP is estimated. The difference is the underlying model and the method of estimation. MAVE is a semi-parametric model, estimates are made via maximum likelihood, and the method is designed for the classical regression setting where $n > p$. Learning gradients uses a non-parametric model and estimation is based on regularization methods and is designed for the high-dimensional setting $p \gg n$. We provide more detailed relations between learning gradients and these methods in Section 4.4.

A more general regression setting for high-dimensional data is when the marginal distribution ρ_X is concentrated on a d -dimensional manifold \mathcal{M} with $d \ll p$. The input space is the manifold, $\mathcal{X} = \mathcal{M}$. 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: $x_i = \varphi(q_i)$.

A generative model or inverse regression in this case is still meaningful $X|Y$. However, a global covariance matrix $\Omega_{X|Y}$ is not so meaningful from a modeling perspective since all one can expect in this setting is local linear structure. The GOP defined in terms the gradient on the manifold

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

is still meaningful from a modeling perspective because gradients on the manifold take local structure into account. Note that the $d \times d$ matrix $\nabla_{\mathcal{M}}f_r \otimes \nabla_{\mathcal{M}}f_r$ has central meaning in our problem. However, we know neither the manifold nor the coordinates on the manifold but the points in the ambient space. For this reason we can only study its properties by investigating the GOP matrix Γ in the ambient space, a $p \times p$ matrix. Details on conditions under which Γ provides information on $\nabla_{\mathcal{M}}f_r \otimes \nabla_{\mathcal{M}}f_r$ are developed in Section 4.3.

An interesting direction that we have not pursued is to consider the partitions separately and not integrate over all partitions. This relates to the idea that different parts of the space have different geometry or statistical dependence.

Remark 3 *The relation between inverse regression and the gradient outer product we present is for regression. A very similar result holds for the classification setting due to the fact that the function $g(\cdot)$ in the semi-parametric model (7) that the results are based on can be a link function such as a logistic or probit function.*

3. Estimating gradients

Algorithms for learning gradients were developed 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 idea behind 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 ∇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 f and \vec{f} 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 ∇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 , λ_2 , and s 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 the Gaussian weights $w_{i,j}^s = \exp(-\|x_i - x_j\|^2/2s^2)$. 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 the 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. We are not given the value of the classification function $f_c(x_i)$. Instead we are given the label $y_i \in \{-1, 1\}$. So we must estimate the functions simultaneously. 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

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

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 λ_1, λ_2 and s are the regularization parameters.

In the manifold setting the above algorithms are still valid. However the interpretation is different. 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 $d\varphi(\nabla_{\mathcal{M}} f_\tau)$ and f_D is a model for f_τ .

From a computational perspective the advantage of the 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)

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

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 involved linear systems of

equations of dimension nd where $d \leq n$. A result of this representation is the following empirical estimate of the GOP

$$\hat{\Gamma} = c_D K^2 c_D^T = \frac{1}{n} \sum_{i=1}^n \vec{f}_D(x_i) \otimes \vec{f}_D(x_i), \quad (6)$$

where K is kernel matrix with $K_{ij} = K(x_i, x_j)$.

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. (2006) and the rate of convergence was determined by the $d_{\mathcal{M}}$, the dimension of the manifold, not the dimension of the ambient space p .

Proposition 8 *Under mild conditions (see Mukherjee et al. (2006) for details), with probability greater than $1 - \delta$,*

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

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

Remark 9 *The idea of learning gradients has been well studied in the numerical analysis literature in terms of computing numerical derivatives, a classical ill-posed problem. For low dimensional data, for example $p = 1$, our gradient estimates are accurate. This is due to the fact that our method is effectively a spline model and spline models have been used with success in numerically computing derivatives (Wahba and Wendelberger, 1980).*

Our methods have been specifically designed for high-dimensional data. In this situation the estimates are rough but still useful for dimension reduction problems as shown in Mukherjee and Zhou (2006); Mukherjee and Wu (2006); Mukherjee et al. (2006).

4. Simultaneous dimension reduction and regression

In this section we study simultaneous dimension reduction and variable and feature selection under various settings. We also relate learning gradients to previous approaches for dimension reduction in regression. We close with an empirical study of the efficacy of some of the methods discussed in this section on real data.

4.1 Variable selection and linear feature construction

Variable selection and linear feature construction are two problems where knowledge of the geometry and statistical dependencies of relevant variables is important. In this section we discuss how estimates of the GOP can be used to address these two problems. We use the formalism in Guyon and Elisseeff (2003) where variable selection refers to the selection of “raw” input variables. We also use “linear feature” to refer to new variables constructed from linear combinations of the input variables. Feature and variable selection are not well defined problems and mathematical or statistical formulations of the problems vary (Guyon and Elisseeff, 2003).

4.1.1 VARIABLE SELECTION

The problem of variable selection is typically stated as finding a small subset or the smallest subset of predictors that are most relevant to predicting the label or outcome (Kohavi and John, 1997; Guyon and Elisseeff, 2003). This idea has been formalized in the work on sparse signal reconstruction (Candès et al., 2005) and one-norm penalization methods (Tibshirani, 1996; Chen et al., 1999).

Learning gradients can be used for the variable selection problem. The diagonal elements of the empirical GOP, $\hat{\Gamma}_{ii} \approx \left\| \frac{\partial f_r}{\partial x^i} \right\|_{L^2_{\rho_X}}^2$, provide a criteria for the importance of each variable – the norm of the partial derivative. The intuition underlying this idea is if a variable is important for prediction, then the target function (f_ρ or f_c) changes quickly along the corresponding coordinate and the norm of the partial derivative should be large.

In Mukherjee and Zhou (2006); Mukherjee and Wu (2006) the RKHS norms of the empirical approximation $\|\vec{f}_{D,i}\|_K$ are used. In this case, more formally the logic is that if $\|\vec{f}_{D,i}\|_K$ is small, then $\left\| \frac{\partial f}{\partial x^{(i)}} \right\|_\infty$ is small and $f(x)$ changes slowly along the i -th coordinate which is therefore less important for prediction. This criteria was shown to be effective in Mukherjee and Zhou (2006); Mukherjee and Wu (2006).

We have observed in simulations that the norm used does not have much of an effect on the rankings of the coordinates, especially in the small sample setting. Intuitively, it seems that for regression using the L_1 or L_2 norm is a good idea since they dominate the total variation of the function. For classification an argument is made for the L_∞ or RKHS norm in Mukherjee et al. (2006). We use the RKHS norm due to computational as well as interpretational advantages.

The key issue in most formulations of variable selection is the colinearity of variables or more generally the redundancy of information provided by a variable. The variety of definitions of variable selection correspond to statistical or algorithmic objectives to address the redundancy. In many applications where the only objective is predictive accuracy the sparsest predictive regression model is desired. Our method is not optimal with respect to this criteria since highly correlated variables will have similar RKHS norms. However, the off-diagonal elements of the empirical GOP matrix, $\hat{\Gamma}_{ij} \approx \left\langle \frac{\partial f_r}{\partial x^i}, \frac{\partial f_r}{\partial x^j} \right\rangle$, does provides a measure of how two variables covary and it may be possible to use the entire GOP matrix to find a sparse subset of variables. In the next sections we will show how a sparse set

of features or combinations of variables can be constructed that are in some sense optimal with respect to prediction.

4.1.2 LINEAR FEATURE CONSTRUCTION

The idea behind linear feature construction is to find a small set of linear combinations of variables most relevant to prediction. These linear combinations called bases or factors span a subspace V of much lower dimensionality than the data and projecting the data onto V results in a regression independent of the original space, $Y|V \perp\!\!\!\perp X$. In this section we will explore the idea of linear projections for feature selection. This idea applies to linear as well as nonlinear functions and will be generalized to the manifold setting in Section 4.3.

In Mukherjee et al. (2006), the following concept of *sensitivity* along a unit vector was introduced as a feature selection criteria.

Definition 10 *Let f be a smooth function on \mathbb{R}^p with gradient ∇f . The sensitivity of the function f along a (unit normalized) direction u is $\|u \cdot \nabla f\|_\infty$.*

Given data D and assuming that $\vec{f}_D \in \mathcal{H}_K^p$ is an approximation of ∇f , the empirical sensitivity along u is defined as $\|u \cdot \vec{f}_D\|_K$.

When the gradient is estimated by the algorithms in Section 3, the top empirical sensitive linear features can be easily computed (Mukherjee et al., 2006) by a spectral decomposition of the *empirical gradient covariance matrix* (EGCM), $\Xi = c_D K c_D^T$ where

$$\Xi_{ij} = \langle f_{D,i}, f_{D,j} \rangle_K.$$

Proposition 11 *Let u_i be the eigenvector of Ξ corresponding to the i -th large eigenvalue. The d most sensitive features are those $\{u_1, \dots, u_d\}$ that are orthogonal to each other and maximize $\|u_i \cdot \vec{f}_D\|_K$.*

Projecting the data onto these d eigenvectors is linear dimension reduction and was proposed and studied in Mukherjee et al. (2006).

The remainder of this section will explore the relation between sensitive features for linear dimension reduction and other statistical methods for linear dimension reduction including SIR and MAVE. These methods focus on the following semi-parametric model:

$$f(x) = g(B^T x) \tag{7}$$

where $B = [b_1, \dots, b_d]$ is a $p \times d$ matrix. The linear subspace spanned by $\{b_i\}_{i=1}^d$, denoted by $\text{span}(B)$, is called the *effective dimension reduction (EDR)* space and $\{b_i\}_{i=1}^d$ are the EDR directions.

A spectral decomposition of the empirical GOP matrix $\hat{\Gamma}$ can be used to compute the d EDR directions. This is shown by the following two results.

Lemma 12 *If f satisfies the semi-parametric model (7), then the matrix Γ is of rank d . Denote by $\{v_1, \dots, v_d\}$ the eigenvectors associated to the nonzero eigenvalues of Γ , it holds that*

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

Proposition 13 *Suppose that f satisfies the semi-parametric model (7) and \vec{f}_D is an empirical approximation of ∇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 Γ respectively. The conclusions then follows from Lemma 12. \blacksquare

Using the empirical GOP matrix Γ to search the EDR space provides a new approach of linear dimension reduction. However, we will show it is closely related to the sensitive feature construction using the EGCM Ξ . We investigate this from a computational perspective. Recall

$$\hat{\Gamma} = c_D K^2 c_D^T \quad \text{and} \quad \Xi = c_D K c_D^T. \quad (8)$$

Without loss of generality, assume K is invertible and has the spectral decomposition $K = QDQ^T$. Then

$$\hat{\Gamma} = c_D Q D^2 Q^T c_D^T.$$

Suppose there is no noise which corresponds to the sample limit case. Then $c_D Q$ will be of rank d and the EDR space is $\text{span}(c_D Q)$ by Proposition 13. For the EGCM

$$\Xi = c_D Q D Q^T c_D^T$$

the top d empirical sensitive features also span $\text{span}(c_D Q)$, coinciding with the EDR space. Therefore, the empirical sensitive features can be regarded as a re-weighting of the EDR directions with the weight depending on the kernel. Therefore, both methods have similar performance.

4.2 Nonlinear dimension reduction: gradient based diffusion maps (GDM)

Several recent methods for nonlinear dimension reduction have exploited the idea of random walks on graphs and manifolds (Belkin and Niyogi, 2003, 2004; Szummer and Jaakkola, 2001; Coifman et al., 2005a,b). A mathematical formalization of this was developed in the ideas of diffusion analysis and diffusion geometry (Coifman and Lafon, 2006; Coifman and Maggioni, 2006).

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 . This matrix represents a weighted undirected graph with the nodes as data points and edges corresponding to the similarity between two points. Given this similarity matrix W two common diffusion operators are the graph Laplacian

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

and a local averaging filter

$$L = D^{-1}W.$$

Nonlinear dimension reduction is achieved by constructing bases based on spectral decompositions of the diffusion operator K or powers of the diffusion operator, K^t , running the diffusion process for some time t . For most unsupervised spectral methods such as Laplacian eigenmaps (Belkin and Niyogi, 2003) kernels of the following form are used to construct the diffusion operator

$$W(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{\sigma_1}\right). \quad (9)$$

In Szlam et al. (2007) the following data or function adapted kernel was proposed to construct the diffusion operator

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

The idea was to first learn a rough estimate of the target function and then plug this into the kernel to construct the graph. This method was shown to work well.

The utility of these diffusion and spectral methods was most dramatic in the semi-supervised learning setting where in addition to a small labeled dataset D we are given many unlabeled data U drawn from the marginal distribution, $U = \{x_i\}_{i=1}^u$. The typical idea was to use the labeled and unlabeled data to learn the bases to project the data onto and then use the labeled data to learn a regression function in this low dimensional projection.

We will propose to use gradient estimates in the function adapted kernel, equation (9), to construct *gradient based diffusion maps (GDM)*. The idea will be to use the labeled data to compute a gradient estimate, \vec{f}_D , and then use the gradient estimate to evaluate equation (9) on all the labeled and unlabeled data to construct the diffusion map. The fact that our learning gradient algorithms provide gradient estimates at any point in the input space and not just on the labeled training data is vital in this process.

For the function based diffusion maps we need to evaluate the functional difference between any two points. Gradient estimates can be used to evaluate this for any two points close to each other based on the following Taylor expansion

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

This estimate is not reliable for points that are far apart but this is not an issue since $W_f(x_i, x_j) \approx 0$ for these points due to the first term in equation (9). We find in practice symmetrizing the estimates alleviates numeric issues and improves performance

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

Empirical comparisons of linear and nonlinear methods including GDM on real data are detailed in Section 4.5.

4.3 The manifold setting

A premise of the manifold setting is that high dimensional data is usually concentrated on a low dimensional manifold. This idea was central in developing learning gradients on manifolds (Mukherjee et al., 2006). The common setting of manifold learning is to assume that there is an isometric mapping into the ambient space, $\varphi : \mathcal{X} \rightarrow \mathbb{R}^p$, and the observations x_i are the images of points of \mathcal{X} onto \mathbb{R}^p under φ , $x_i = \varphi(q_i)$. The idea is analogous to the ambient space formulation except the Taylor expansion is now on the Riemannian manifold. The explicit formula for the regression case is given in Definition 6. Note that the algorithms for the ambient case and the manifold case are identical.

From a modeling perspective we would like the gradient estimate provided by the algorithm \vec{f}_D to approximate $d\varphi(\nabla_{\mathcal{M}} f_\rho)$ (Mukherjee et al., 2006). Generally this is not true when the manifold is nonlinear (that is, φ is a nonlinear map). Instead, the estimate provides the following information on $\nabla_{\mathcal{M}} f_\rho$:

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

where $(\mathrm{d}\varphi)^*$ is the dual of $\mathrm{d}\varphi$, 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 . We cannot recover directly the gradient of f_R on the manifold since we we know neither the manifold nor the embedding. The question becomes what can we do with the gradient estimate \vec{f}_D the empirical GOP $\hat{\Gamma}$ or the EGCM, Ξ . We will see that these quantities are indeed meaningful and we discuss this more formally below. When the manifold has a linear structure then \vec{f}_D has a geometric interpretation as the gradient of the canonical extension of f_r (Mukherjee et al., 2006). In the following we consider the more general setting of a nonlinear manifold.

4.3.1 FEASIBILITY OF LINEAR FEATURE CONSTRUCTION

In this subsection we explain why linear feature construction selection is still feasible. Assume f_r satisfies the semi-parametric model (7). The matrix Γ is not well defined but $\hat{\Gamma}$ and Ξ are. In order to show spectral decompositions of $\hat{\Gamma}$ or Ξ provide the EDR directions, it is enough to notice the following result.

Proposition 14 *If $v \perp b_i$ for all $i = 1, \dots, d$, then $v^T \Xi v \rightarrow 0$ and $v^T \hat{\Gamma} v \rightarrow 0$.*

Proof Let \vec{f}_λ 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 $v \cdot f_\lambda = 0$.

It was proven in Mukherjee and Zhou (2006) that $\|\vec{f}_D - \vec{f}_\lambda\|_K \rightarrow 0$ implying

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

as well as $v^T \hat{\Gamma} v \rightarrow 0$. This proves the conclusion. ■

Proposition 14 states that all the vectors perpendicular to the EDR space correspond to eigenvalues near zero of Ξ and will be filtered out. This means the EDR directions can be still found by the spectral decomposition of the EGCM.

4.3.2 FEASIBILITY OF GRADIENT BASED DIFFUSION MAPS

For the gradient based diffusion map to be feasible, the key is to use the gradient estimate to evaluate the difference between points. On a manifold this should be done by 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 Exp_{x_i} is the exponential map at x_i ; see do Carmo (1992); Mukherjee et al. (2006). As before, $\nabla_{\mathcal{M}} f$ is not computable and we need use \vec{f}_D instead. The following result shows that this is reasonable.

Proposition 15 *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. ■

These results verify that the dimension reduction methods we propose using spectral decompositions or gradient based diffusion maps still make sense when the underlying input space is a nonlinear manifold.

4.4 Relation to previous work

In this section we relate gradient based dimension reduction methods with several well known methods including SIR, and MAVe.

4.4.1 RELATION TO MINIMUM AVERAGE VARIANCE ESTIMATION (MAVE) AND OUTER PRODUCT OF GRADIENTS (OPG)

In Xia et al. (2002) the MAVe and OPG methods were proposed. These methods share with learning gradients the idea of using a first order Taylor expansion and minimizing the empirical variance. MAVe starts from the semi-parametric model (7) and estimates the EDR space directly. The gradient is used implicitly in the estimation process. The OPG method estimates the gradient only at the sample points and computes an empirical approximation of the GOP matrix Γ . Learning gradients has two main advantages over these approaches:

1. the estimate of the gradient is a vector valued function that can be evaluated anywhere in the input space and not just at sample points;

2. a regularization term or shrinkage prior is added for numerical stability and statistical robustness.

A result of the first point is that learning gradients can be used not only linear dimension reduction by constructing an empirical GOP matrix but also for response driven nonlinear methods such as GDM. Recall that estimate of gradient is a typical ill-posed inverse problem. A result of the second point is that although our gradient based linear feature construction method performance similarly to MAVE and OPG for large datasets, when n is small or p is very large MAVE or OPG may fail while our method still works because of the regularization.

4.4.2 RELATION TO SLICED INVERSE REGRESSION (SIR)

The SIR method searches the EDR directions by 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 Γ .

We start with a simple model where the EDR space contains only one direction, the regression function satisfies this semi-parametric model

$$y = g(\beta^T x) + \epsilon$$

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

Theorem 16 *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 $C = \sigma_Y^2 \left(1 - \frac{\sigma_\epsilon^2}{\sigma_Y^2}\right)^2$.

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

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

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

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

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 (10) 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 $\approx \frac{\sigma_Y^2}{\sigma^4}$. Notice that this factor measures the change ratio of the response variable over the input space as well as along the predictive direction. So we can say 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 y .

When there are more than one EDR directions as in model (7), we partition the input space into several small regions $X = \bigcup_i R_i$ 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 Corollary 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 Γ_i is the gradient out product matrix on R_i and $\Omega_i = \text{cov}(X|R_i)$. In this sense, the gradient covariance matrix Γ can be regarded as the weighted sum of the local covariance matrix 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 cannot 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 a few advantages of gradient based methods.

4.5 Empirical comparisons on real data

In this section we empirically examine nonlinear dimension reduction using the gradient based diffusion map (GDM). The efficacy of linear dimension reduction using gradient based linear feature selection (GLFS) was studied in simulations as well as real data in Mukherjee et al. (2006). Our approach will be to compare four dimension reduction methods: principal components analysis (PCA), gradient based linear feature selection (GLFS), (unsupervised) diffusion maps (DM), and gradient based diffusion maps (GDM). From the training data we first learn the dimension reduction and then use the 5-nearest neighbor algorithm as a classifier. The test error will serve as the criteria for the accuracy of the dimension reduction.

The comparison is on a subset of the MNIST digits dataset, a standard benchmark dataset used in the machine learning community (Y. LeCun, <http://yann.lecun.com/exdb/mnist/>).

The dataset 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” dataset, and “3” from a “8”, the “38” data set. The first is one of the more difficult pairwise comparisons.

For both datasets the following procedure was repeated 50 times. Randomly selection 1000 samples from each of the two digits. Select a small number of training samples (see Tables 1 and 2 for the exact numbers). Use this training data to learn the dimension reduction. Cross-validation is used to select of dimensions, between 1 – 5, and the parameters of the dimension reduction method. We then use the 5-nearest neighbor algorithm on the low dimensional training data to classify the independent testing data and record the error rate. The average over the 50 iterations is displayed in Tables 1 and 2.

For PCA and GLFS, we use only the labeled training data to construct the linear features and train the classifier. For DM, we use the labeled training data as well as the unlabeled test data and the kernel (9) to build the graph where the kernel parameter σ_1 is a 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 σ_2 chosen by cross-validation. As a reference or best case error rate we list in the final row of the two tables results for optimal parameters for GDM.

The results indicate that the nonlinear methods outperform linear methods when unlabeled data is available. The response dependent methods outperform the response independent methods. However, note that linear methods perform well when the underlying structure of the marginal distribution is approximately linear, as may be the case for for “6” versus “9”.

Labeled points	20	40	60	100	200
PCA	18.85%	15.81%	14.88%	13.08%	11.91%
GLFS	14.74%	10.49%	9.49%	8.16%	6.16%
DM	8.05%	4.44%	4.15%	3.77%	3.46%
GDM	7.24%	3.91%	3.73%	3.43%	3.21%
GDMopt	5.76%	3.42%	3.25%	2.86%	2.61%

Table 1: Comparison of error rates for the 3 vs 8 classification for the various dimension reduction methods.

5. Graphical models and conditional independence

A key idea in Section 4.1.2 was that the estimate of the GOP matrix $\hat{\Gamma}$ provides information about the geometry and covariance of relevant input variables. This matrix is the covariance matrix of a multivariate Gaussian due to the RKHS construction, see equations (8).

A very natural idea is to use a graphical model over undirected graphs to model conditional independence of the multivariate distribution characterized by $\hat{\Gamma}$. The theory of Gauss-Markov graphs (Speed and Kiiveri, 1986; Lauritzen, 1996) was developed for multi-

labeled points	20	40	60	100
PCA	3.43%	1.79%	1.32%	1.03%
GLFS	2.35%	1.32%	1.12%	0.86%
DM	1.75%	0.09%	0.05%	0.05%
GDM	1.08%	0.09%	0.05%	0.06%
GDMopt	0.23%	0.06%	0.05%	0.05%

Table 2: Comparison of error rates for 6 vs. 9 classification for the various dimension reduction methods.

variate Gaussian densities

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

where the covariance is J^{-1} and the mean is $\mu = J^{-1}h$. The result of the theory is the precision matrix J is the conditional independence matrix and each element J_{ij} is a measure of dependence between variables i and j conditioned on all other variables.

An immediate result of this is that the inverse of the GOP provides an inference on the dependence structure of the variables relevant to prediction.

5.1 Application to real and simulated data

We construct undirected graphical models using the precision matrix J for simulated and gene expression data. In practice, the covariance matrix is ill-conditioned so we use the pseudo-inverse to invert $\hat{\Gamma}$. We set the threshold for the SVD based on the decay of the eigenvalues, specifically jumps or eigen-gaps.

For our simulation, we follow the example in Mukherjee and Zhou (2006) with some minor modifications. We consider the regression problem by sampling 30 points from three linear functions in \mathbb{R}^{80} :

$$y_i = x_i \cdot w_1 + \text{No}(0, \sigma_y)$$

$$y_i = x_i \cdot w_2 + \text{No}(0, \sigma_y)$$

$$y_i = x_i \cdot w_3 + \text{No}(0, \sigma_y)$$

We draw ten points from each of the 3 functions over different partitions of the space, where

$$\{x_i\}_{i=1}^{10} = \mathbf{1}, \{x_i\}_{i=11}^{20} = \mathbf{1}, \{x_i\}_{i=21}^{30} = -\mathbf{1}$$

and with slopes

$$w_1 = 2 \text{ for } j=1, \dots, 10 \text{ and } 0 \text{ otherwise,}$$

$$w_2 = -2 \text{ for } j=11, \dots, 20 \text{ and } 0 \text{ otherwise,}$$

$$w_3 = -2 \text{ for } j=41, \dots, 50 \text{ and } 0 \text{ otherwise.}$$

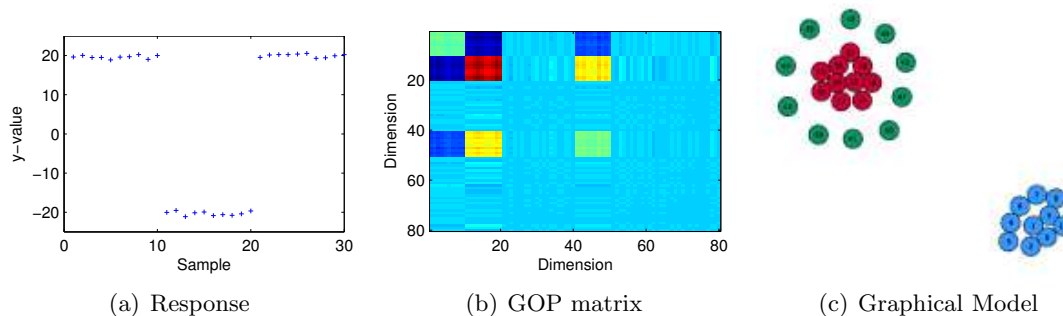


Figure 1: Regression simulation: (a) the response variable, (b) the GOP matrix, (c) graphical model inferred.

The response values and the GOP matrix are shown in Figure (1a,b) respectively. We build a graphical model from $\hat{\Gamma}$ using the precision matrix J , displayed in Figure (1c). Edges are removed for visual clarity. Distances between nodes reflect the magnitude of the conditional dependence, with large distances reflecting strong negative correlation, and small distances showing strong positive correlation. Nodes that are strongly independent to all other nodes (weights ≈ 0) are removed from the graph. We notice the strongest dependencies in the block $i, j \in \{11, \dots, 20\}$, corresponding to the dense red nodes in our graph. This is consistent with our intuition, as these variables are associated with the largest change in y in the regression problem. The dark blue block and bright yellow block in Figure (1b) reflect the strong negative and positive dependencies we expect from this data. Likewise, the large distance between the blue and red nodes, and the smaller distance between the green and red nodes reflects this dependency structure.

We next explore graphical models using mRNA micro-array expression data. We obtain gene expression data derived from oncogenic cell lines, in which a specific oncogene is knocked-out (silenced). Oncogenes are cancer catalysts, triggering biological pathways that are implicated in cancer genesis and progression. For our experiment we look at two cell lines, the Ras and Myc oncogenic cell lines (Bild et al., 2006), classes 1 and -1 respectively. The expression values reflect the downstream perturbations of the gene network when the oncogene is silenced. The underlying question we wish to address is how do these perturbations covary within the context of a discriminative model that distinguishes the Ras and Myc specific pathways.

We apply the classification methods described in Mukherjee and Wu (2006) to obtain our GOP matrix. Because of the large number of variables in this problem, we set a threshold to filter out highly independent genes ($\hat{\Gamma}_{ij} < \epsilon$) to obtain a graph of $\bar{1}500$ nodes, Figure (2). Highlighted on this graph are the nodes that correspond to the unique pathway signatures for Ras (blue) and Myc (red) obtained in Bild et al. (2006). The most striking observation of this graph is the presence of 3 tight clusters, with strong negative conditional dependence of two of the clusters with the third. We also observe that many Myc genes show stronger conditional dependence with the bottom, right cluster, and the Ras genes exhibit a conditional dependence distributed between the two bottom clusters. We believe

this graph captures many interesting gene dependencies that could become the basis for future biological experiments.

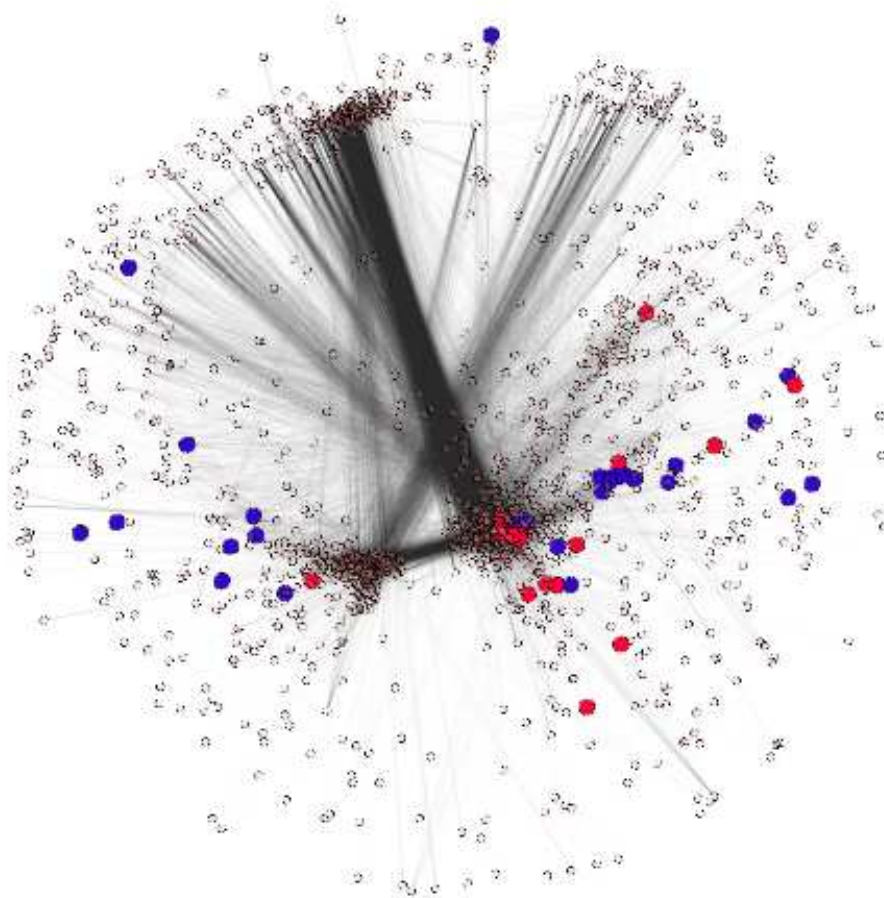


Figure 2: Graphical Model: Myc/Ras Gene Expression.

6. Discussion

In this paper we provide a general statistical framework based on learning gradients that simultaneously infers a predictive model and estimates the geometry and statistical dependence of the input variables relevant to prediction.

The development of this framework argues for the centrality of the gradient outer product (GOP) and relates a variety of methods for simultaneous dimension reduction and regression (Li, 1991; Xia et al., 2002; Cook and Weisberg, 1991; Li, 1992). We discuss how

the gradient estimates can be applied to a variety of problems central to machine learning: variable selection, linear and nonlinear dimension reduction, and the inference of a graphical model of the dependencies of the input variables that are relevant to prediction. Empirical results for dimension reduction and inference of graphical models show the efficacy of this approach.

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.
- A.H. Bild, G. Yang, J.T. Wang, Q. Wang, A. Potti, D. Chasse, M. Joshi, D. Harpole, J.M. Lancaster, A. Berchuck, J.A. Olsen Jr, J.R. Marks, H.K. Dressman, M. West, and J.R. Nevins. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*, 439:353–357, 2006.
- E. J. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm Pure Appl Math*, 59:1207–1223, 2005.
- S.S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20,(1):33–61, 1999.
- 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". *jasa*, 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. Stat.*, 19(2): 505–530, 1991.
- F.Y. Edegworth. On the reduction of observations. *Philosophical Magazine*, pages 135–141, 1884.
- R.A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the ROyal Statistical Society A*, 222:309–368, 1922.
- K Fukumizu, FR Bach, and MI Jordan. Dimensionality reduction in supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5:73–99, 2005.
- I Guyon and A Ellsseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.
- R. Kohavi and G.H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97 (1-2):273–324, 1997.
- S.L. Lauritzen. *Graphical Models*. Oxford: Clarendo Press, 1996.
- K.C. Li. Sliced inverse regression for dimension reduction. *jasa*, 86:316–342, 1991.
- K.C. Li. On principal hessian directions for data visualization and dimension reduction: another application of Stein’s lemma. *annals*, 97:1025–1039, 1992.
- 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.
- 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.
- R. Tibshirani. Regression shrinkage and selection via the lasso. *J Royal Stat Soc B*, 58(1): 267–288, 1996.
- G. Wahba. *Splines Models for Observational Data*. Series in Applied Mathematics, Vol. 59, SIAM, Philadelphia, 1990.
- G. Wahba and J. Wendelberger. Some new mathematical methods for variational objective analysis using splines and cross-validation. *Monthly Weather Rev.*, 108:1122–1145, 1980.
- Y. Xia, H. Tong, W. Li, and L-X. Zhu. An adaptive estimation of dimension reduction space. *jrssb*, 64(3):363–410, 2002.