Variable Selection and Dimension Reduction by Learning Gradients

Qiang Wu and Sayan Mukherjee

August 6, 2008

1 Introduction

High dimension data analysis has become a challenging problem in modern sciences. The diagnosis by gene expression or SNPs data arising in medical and biological sciences is a typical example and may be the most important focus of the research in the past decade. The number of variables in these data sets may be tens to hundreds of thousands. The understanding of the data structure and inference are much difficult because of the curse of the dimension. This has driven the rapid advances in the research of variable selection and dimension reduction techniques in machine learning and statistical communities that show great advantages.

Variable selection is closely related to the relevance study and dates back at least to the middle of 1990's (Tibshirani, 1996; Blum and Langley, 1997; Kohavi and John, 1997). Since then it has been rapidly developed, especially after the microarray data and text categorization draw attention of researchers. A special issue on this topic is published by *Journal of Machine Learning Research* in 2003. In Guyon and Elisseeff (2003) the main benefits of of variable selection were summarized to be 3-fold: improving the inference performance, providing faster and cost-effective predictors, and better understanding of underlying process that generates the data. The many methods that have been proposed in the literature includes various correlation and information criteria (see Guyon and Elisseeff (2003) for a

good review); the ordinary least square based algorithms such as LASSO (Tibshirani, 1996) and elastic net (Zou and Hastie, 2005), SVM based algorithms (Guyon et al., 2002; Weston et al., 2001; Rakotomamonjy, 2003; Zhang, 2006), and gradient based algorithms (Mukherjee and Zhou, 2006; Mukherjee and Wu, 2006; Cai and Ye, 2008; Ying and Campbell, 2008).

Dimension reduction is another major component arising in high-dimensional data analysis that can be used for visualization, statistical modeling, inference of geometry and structure, and improving predictive accuracy. The unsupervised dimension reduction does not take the response variable into account. It goes back to principal components analysis and has been extended to nonlinear setting by kernel trick (Schölkopf et al., 1997). Rooting in the belief that high dimensional data are usually concentrated on a low dimensional (possibly nonlinear) manifold, approaches of nonlinear dimension reduction by recovering the intrinsic metric have been proposed (Tenenbaum et al., 2000; Roweis and Saul, 2000; Belkin and Niyogi, 2003; Donoho and Grimes, 2003; Coifman and Lafon, 2006; Coifman et al., 2005a,b). An obvious observation is that given label or response information incorporating this information into the dimension reduction framework should be beneficial, especially for prediction. The problem of finding predictive directions or projections called supervised dimension reduction has been developed in the statistics literature by methods such as sliced inverse regression (SIR, (Duan and Li, 1991; Li, 1991)) and various extensions, principle Hessian directions (PHD, Li (1992)), gradient based methods such as minimum average variance estimation (MAVE, Xia et al. (2002)) and kernel based gradient learning (Mukherjee et al., 2006b), and so on. All of these methods were shown empirically effective and their properties have been theoretically analyzed.

The aim of this paper is to review the kernel gradient learning algorithms proposed in Mukherjee and Zhou (2006); Mukherjee and Wu (2006); Mukherjee et al. (2006b). It can be used for simultaneous feature selection and dimension reduction. The discussion will be put it into the context of feature selection by nonlinear

models. This provides a new view of point.

2 Motivations and foundations

In this section we study the motivations and foundations for variable selection and dimension reduction by learning gradients. For variable selection we will show that linear models may not be enough to retrieve all the relevant variables and algorithms based on nonlinear models are necessary. In this context gradients usually provide enough information to rank the importance of the variables and help select most relevant and predictive ones. For dimension reduction we will show that gradient outer product matrix contains all the information about how the the target function changes and could be used to recover the predictive directions without degeneracy.

Before we go into the details, we introduce the concepts and notations that will be used throughout the whole paper. Let $\mathbf{x} = (x_1, \ldots, x_p)^T \in \mathbb{R}^p$ be a random variable of p predictors. Each predictor x_i is called a variable or feature. Let y be the response variable with domain $\mathbb{R}$ for regression problem or $\{\pm 1\}$ for binary classification problem. We will focus on the additive noise model, i.e., the response variable y depends on the input variable $\mathbf{x}$ in the following way:

$$y = f_*(\mathbf{x}) + \epsilon \tag{1}$$

where ϵ is independent of $\mathbf{x}$ and $\mathbb{E}\epsilon = 0$.

2.1 Gradient and variable selection by nonlinear models

In very high dimensional data analysis, the target function may depend only on a small subset of the predictors, not all of them. Denote the set of relevant variables by $\mathbf{x}_R$ where $R \subset \{1, \ldots, p\}$. The models (1) becomes

$$y = f(\mathbf{x}_R) + \epsilon. \tag{2}$$

The aim of variable selection is to find this subset of relevant variables and rank their importance. This is important in many applications such as the medical diagnosis and prognosis. A direct benefit of feature subset selection is better understanding of the dependence between the response and the predictors. Also, though theoretically the prediction is more accurate by including all the relevant variables, in practice the prediction may be better using only several top important predictors because most learning algorithms have higher accuracy for lower dimensional inputs, especially when only a limited set of samples are available.

In the literature a variety of variable ranking methods are based on correlation and information criteria (Golub et al., 1999; Hastie et al., 2001, and so on). As reviewed in Guyon and Elisseeff (2003) they face some basic difficulties. Many correlation criteria can only detect linear dependency and may loss some relevant variables. Moreover, a common criticism on correlation criteria is the selection of a redundant subset.

Feature selection by linear models have been extensively studied. In these methods a linear regression function or classifier is trained and the variables are ranked by the coefficients. Typical examples include the LASSO (Tibshirani, 1996), LARS (Efron et al., 2004) and elastic net (Zou and Hastie, 2005) for regression and SVM for classification Vapnik (1998); Guyon et al. (2002). Next we will show that linear models are not enough in many situations. Let us consider the regression case.

In regression a linear function is fitted by ordinary least square (OLS) together with certain penalty. Given a set of samples $\{\mathbf{x}_i, y_i\}_{i=1}^n$ where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T \in \mathbb{R}^p$, the regression function is approximation by $\hat{f}(\mathbf{x}) = \hat{\mathbf{w}}_\lambda^T \mathbf{x} + \hat{b}_\lambda$ where

$$(\hat{\mathbf{w}}_\lambda, \hat{b}_\lambda) = \arg \min_{\mathbf{w} \in \mathbb{R}^p, b \in \mathbb{R}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - \mathbf{w}^T \mathbf{x}_i)^2 + \lambda \Omega(\mathbf{w}) \right\}, \tag{3}$$

where $\mathbf{w} = (w_1, \ldots, w_p)^T \in \mathbb{R}^p$, Ω is a penalty function, and $\lambda > 0$ is called the regularization parameter. Various penalties have been introduced. For instance, the ridge regression uses a L_2 penalty $\Omega(\mathbf{w}) = \sum_{j=1}^p w_j^2$, LASSO uses L_1 penalty

$\Omega(\mathbf{w}) = \sum_{j=1}^{p} |w_j|$, and bridge regression use L_q penalty $\Omega(\mathbf{w}) = \sum_{j=1}^{p} |w_j|^q$ for $q > 0$. Combined penalties are also considered such as that the combination of L_2 and L_1 penalties (Zou and Hastie, 2005) and the combination of L_0 and L_1 penalties in (Liu and Wu, 2007). Since these penalized linear regression is consistent it is suffices to consider the sample limit case to study their properties.

The penalized linear regression for the sample limit case takes the form $f(\mathbf{x}) = \mathbf{w}_\lambda^T \mathbf{x} + b_\lambda$ where $\mathbf{w}_\lambda = (w_{\lambda,1} \ldots, w_{\lambda,p})^T \in \mathbb{R}^p$ and $b_\lambda \in \mathbb{R}$ is defined as

$$(\mathbf{w}_\lambda, b_\lambda) = \arg \min_{\mathbf{w} \in \mathbb{R}^p, b \in \mathbb{R}} \Big\{ \mathbb{E}(y - \mathbf{w}^T \mathbf{x})^2 + \lambda \Omega(\mathbf{w}) \Big\}. \tag{4}$$

We have the following conclusion.

Theorem 1. *Assume the penalty function Ω satisfies $\Omega(\mathbf{w}_1) \leq \Omega(\mathbf{w}_2)$ if $w_{1j} \leq w_{2j}$ for all $j = 1, \ldots, p$ and, if in addition there exists j such that $w_{1j} = 0 < w_{2j}$, then $\Omega(\mathbf{w}_1) < \Omega(\mathbf{w}_2)$. Then we have $w_{\lambda,j} = 0$ if and only if*

$$\mathbb{E}[(x_j - \bar{x}_j)(y - \bar{y})] = 0$$

where $\bar{x}_j = \mathbb{E}x_j$ and $\bar{y} = \mathbb{E}y$.

This result indicates that all the predictors uncorrelated to y will not be selected. Note that it is possible for a variable x_j to be relevant but uncorrelated to y. So feature selection by linear models may loss important variables if y depends on the predictor nonlinearly. In fact, this situation could be very common as shown in the following corollary.

Corollary 2. *Let the penalty function satisfy the same assumption as in Theorem 1. Assume further that $\mathbf{x}$ has normal distribution. Then $w_{\lambda,j} = 0$ if and only if $\mathbb{E}[\frac{\partial f_*(\mathbf{x})}{\partial x_j}] = 0$.*

Corollary 1 follows immediately from Theorem 1 and Stein's Lemma (Stein, 1981, Lemma 4).

The failure of linear models advocates the necessity of feature subset selection by nonlinear models. Several approaches have been proposed in the literature, for

instance, the SVM based criteria in Weston et al. (2001); Rakotomamonjy (2003), the component smoothing and selection operator (COSSO, Zhang (2006)), regularization of derivative expectation operator (RODEO, Lafferty and Wasserman (2008)), and the gradient based variable ranking (Hermes and Buhmann, 2000; Mukherjee and Zhou, 2006; Mukherjee and Wu, 2006).

The gradient measures the change of a function along the coordinate variables. It is natural to rank the variables using the gradients information. This is the idea of feature selection in Hermes and Buhmann (2000); Mukherjee and Zhou (2006); Mukherjee and Wu (2006). Moreover, in case of an additive noise model (2), these methods do not include irrelevant variable nor loss relevant variables.

Theorem 3. *Suppose the model* (2) *holds and f_* is differentiable. Then $j \in R$ if and only if $\frac{\partial f_*}{\partial x_j}$ is not identically* 0.

This motivates variable ranking by certain norm of the partial derivatives. Natural choices include the L_1 and L_2 norms.

2.2 Gradient outer product and dimension reduction

We consider the following heuristic model: assume the functional dependence between the response variable y and the input variable $\mathbf{x}$ is given by

$$y = f_*(\mathbf{x}) + \epsilon = g_*(\beta_1^T \mathbf{x}, \ldots, \beta_d^T \mathbf{x}) + \epsilon \tag{5}$$

where $\beta_1, \ldots, \beta_d$ are are unknown vectors in $\mathbb{R}^p$. Let $\mathcal{S}$ denote the subspace spanned by these β_i's. Then $P_{\mathcal{S}}\mathbf{x}$, where $P_{\mathcal{S}}$ denotes the projection operator onto the subspace space $\mathcal{S}$, provides a sufficient summary of the information in x relevant to y. Estimating $\mathcal{S}$ or β_i's becomes the central problem in supervised dimension reduction.

Though we define $\mathcal{S}$ here via a model assumption (5), formal definition based on conditional independence is available. In the statistical literatures of dimension reduction two central concepts are usually used. The central subspace (Cook, 1998) is defined to be the intersection of all subspaces $\mathcal{S}$ such that $y \perp\!\!\!\perp \mathbf{x} | P_{\mathcal{S}}\mathbf{x}$. The

central mean subspace (Cook and Li, 2002) is defined to be the intersection of all subspaces $\mathcal{S}$ such that $\mathbb{E}(y|\mathbf{x}) \perp\!\!\!\perp \mathbf{x}|P_{\mathcal{S}}\mathbf{x}$. Both exists under very mild conditions. For the additive models (5) two concepts coincide. In the sequel we will assume without loss of generality that the central (mean) subspace exists and β_i's form a basis. Also, we follow Cook and Yin (2001) and refer to $\mathcal{S}$ as the dimension reduction (d.r.) subspace and β_i's the d.r. directions.

The gradient vector $\nabla f_*(x) = (\frac{\partial}{\partial x_1} f_*(\mathbf{x}), \ldots, \frac{\partial}{\partial x_p} f_*(\mathbf{x}))^T$ measures how the function f changes at the point $\mathbf{x}$ with respect to its coordinates. A natural idea is to retrieve the d.r. subspace using gradient information. It turns out the gradient outer product matrix defined as

$$\mathbf{G} = \mathbb{E}[(\nabla f_*(x))(\nabla f_*(x))^T]$$

plays a central role. The following lemma is immediate.

Theorem 4. *Assume the model* (5) *and the differentiability of* f_*. *Then the d.r. subspace* $\mathcal{S}$ *is exactly the subspace spanned by the eigenvectors of* $\mathbf{G}$ *with nonzero eigenvalues.*

This theorem suggests to retrieve the d.r. subspace by eigen-decomposition of gradient outer product matrix $\mathbf{G}$. In addition, it guarantees no predictive direction is lost.

Several dimension reduction methods follow this idea implicitly or explicitly (Xia et al., 2002; Mukherjee et al., 2006b). Recall many classical dimension reduction methods such as SIR and PHD suffer the loss of d.r. directions in certain situations. Compared to those methods, gradient outer product based method is advantageous in preventing degeneracy. Some theoretical relations and comparisons between gradient based method and SIR or PHD were discussed in Wu et al. (2007); Mukherjee et al. (2006a).

To close this section we remark that most variable selection approaches do not estimate how the relevant variables covary and does not provide information of

d.r. subspace. Gradients relates the variable selection and dimension reduction together. So estimating the gradient can be used for both tasks and are of great interest.

3 Learning gradients by kernel methods

In the literature, several approaches have been proposed to learn the gradient. They include various numerical derivatives methods, local polynomial smoothing (Fan and Gijbels, 1996) and kernel gradient learning (Mukherjee and Zhou, 2006; Mukherjee and Wu, 2006; Mukherjee et al., 2006b). In this section we discuss the idea of kernel gradient learning.

3.1 Learning gradient for regression

In the common regression setting, the variable $(\mathbf{x}, y)$ is assumed to have joint probability distribution $\rho(\mathbf{x}, y) = \rho_{\mathbf{x}}\rho(y|\mathbf{x})$ where $\rho_{\mathbf{x}}$ is the marginal distribution of $\mathbf{x}$ and $\rho(y|\mathbf{x})$ is the conditional distribution of y given $\mathbf{x}$. The objective regression function $f_* = \mathbb{E}(y|\mathbf{x})$ can be obtained by minimizing the variance functional $\text{Var}(f) = \mathbb{E}(y - f(\mathbf{x}))^2$ over $L^2_{\rho_{\mathbf{x}}}$:

$$f_* = \mathbb{E}(y|\mathbf{x}) = \arg \min_{f \in L^2_{\rho_{\mathbf{x}}}} \text{Var}(f).$$

In statistics and machine learning typically ρ is assumed to be unknown. Instead, what we have in hand is a set of random samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ drawn independently and identically distributed from the joint distribution ρ and f_* must be learned from this set of samples.

In gradient learning our target is ∇f_*, not f_*. In Mukherjee and Zhou (2006) the kernel gradient learning for this regression setting is introduced by the following motivation. Suppose that the regression function f_* is smooth. The Taylor series expansion gives us

$$f_*(\mathbf{u}) \approx f_*(\mathbf{x}) + \nabla f_*(\mathbf{x}) \cdot (\mathbf{u} - \mathbf{x}), \text{ if } \mathbf{x} \approx \mathbf{u},$$

which can be evaluated at the data points

$$f_*(\mathbf{x}_i) \approx f_*(\mathbf{x}_j) + \nabla f_*(\mathbf{x}_j) \cdot (\mathbf{x}_i - \mathbf{x}_j) \approx y_j + \nabla f_*(\mathbf{x}_j) \cdot (\mathbf{x}_i - \mathbf{x}_j) \text{ if } \mathbf{x}_i \approx \mathbf{x}_j$$

where $y_j \approx f(\mathbf{x}_j)$. Let $W(\mathbf{x})$ be a weight function that satisfies $W(\mathbf{x}) \to 0$ as $\|\mathbf{x}\| \to \infty$ and $\mathbb{E}W(\mathbf{x}) = 1$. Then $W(\mathbf{x} - \mathbf{u})$ provides a measure of locality. Denote $W_{ij} = W(\mathbf{x}_i, \mathbf{x}_j)$. We expect

$$\text{Var}(f_*) \approx \frac{1}{n} \sum_{i,j=1}^{n} W_{ij} \Big(y_i - y_j - \nabla f_*(\mathbf{x}_j) \cdot (\mathbf{x}_i - \mathbf{x}_j) \Big)^2$$

is small. Therefore, if a vector valued function $\mathbf{f} : \mathbb{R} \to \mathbb{R}^p$ approximates the gradient ∇f_* well, then the weighted square difference

$$\mathcal{E}_n(\mathbf{f}) = \frac{1}{n} \sum_{i,j=1}^{n} W_{ij} \Big(y_i - y_j - \mathbf{f}(x_j) \cdot (x_i - x_j) \Big)^2$$

should also be small. This motivates the idea of estimating the gradient ∇f_* of the regression function by minimizing $\mathcal{E}_n(\mathbf{f})$.

Since minimizing $\mathcal{E}_n(\mathbf{f})$ over all possible p-vector valued functions results in overfitting we restrict the minimization problem on a relatively small function space. In kernel gradient learning the reproducing kernel Hilbert space (RKHS) is used.

A real valued function $K(\mathbf{x}, \mathbf{u})$ on $\mathbb{R}^p \times \mathbb{R}^p$ is called a Mercer kernel if it is continuous, symmetric, and semi-positive definite. The RKHS $\mathscr{H}_K$ associated to a Mercer kernel K is defined to be the closure of the functions spanned by $\{K_{\mathbf{x}} = K(\mathbf{x}, \cdot)\}$ with the norm

$$\|f\|_K = \left(\sum_{i,j=1}^{m} c_i c_j K(\mathbf{u}_i, \mathbf{u}_j) \right)^{1/2} \qquad \text{if } f = \sum_{i=1} c_i K_{\mathbf{u}_i},\ c_i \in \mathbb{R},\ \mathbf{u}_i \in \mathbb{R}^p.$$

The reproducing property is given by

$$f(\mathbf{u}) = \langle f,\ K_{\mathbf{u}} \rangle_K, \qquad \forall f \in \mathscr{H}_K.$$

We refer to Aronszajn (1950) for more properties of RKHS.

We remark that RKHS is very general. For example, the polynomials of degree $\leq m$ form an RKHS with kernel $(1 + \mathbf{x} \cdot \mathbf{u})^m$ and Sobelev spaces are RKHS with spline kernels. The other widely used kernels include radial basis functions and Gaussian kernels $G(\mathbf{x}, \mathbf{u}) = \exp(-\frac{\|\mathbf{x}-\mathbf{u}\|^2}{2\sigma^2})$, $\sigma > 0$. Note also that the RKHS associated to spline kernel and Gaussian kernels are dense in $L^2_{\rho_\mathbf{x}}$ and hence enough to approximate any smooth functions.

In kernel gradient learning, the minimization will be restricted in certain $\mathscr{H}_K^p$ which is a space of p-vector valued functions with each component in $\mathscr{H}_K$. That is

$$\mathscr{H}_K^p = \{\mathbf{f} = (f_1, \ldots, f_p) : \ f_i \in \mathscr{H}_K, \ i = 1, \ldots, p\}$$

and for each $\mathbf{f}$ define $\|\mathbf{f}\|_K^2 = \sum_{i=1}^p \|f_i\|_K^2$. In addition, the Tikhonov regularization technique is introduced to avoid computational instability. This leads to the following kernel gradient learning algorithm (Mukherjee and Zhou, 2006).

Definition 5. *Given the data* $\{(\mathbf{x}_i, y_i), \ i = 1, \ldots, n\}$*, the gradient* ∇f_* *of the regression function is estimated by*

$$\hat{\mathbf{f}}_\lambda := \arg\min_{\mathbf{f} \in \mathscr{H}_K^p} \Big[\mathcal{E}_D(\mathbf{f}) + \lambda \|\mathbf{f}\|_K^2\Big], \tag{6}$$

where $\lambda > 0$ *is a regularization parameter.*

3.2 Optimization

In this subsection we discuss the optimization of the kernel gradient learning algorithm given in Definition 5. Firstly, as a consequence of the reproducing property we have the following representer theorem.

Theorem 6 (Representer theorem)**.** *There exist* $\mathbf{c}_i = (c_{i1}, \ldots, c_{ip})^T \in \mathbb{R}^p, \ i = 1, \ldots, n$ *so that the solution* $\hat{\mathbf{f}}_\lambda$ *of* (6) *is given as*

$$\hat{\mathbf{f}}_\lambda = \sum_{i=1}^n \mathbf{c}_i K_{\mathbf{x}_i}.$$

Theorem 6 is proved in Mukherjee and Zhou (2006). It states that in order to find the solution $\hat{\mathbf{f}}_\lambda$, it is enough to solve these coefficients $\mathbf{c}_i$. This can be realized by solving a linear system of order np (Mukherjee and Zhou, 2006).

Theorem 7. *The vector* $\mathbf{c} = (\mathbf{c}_1^T, \ldots, \mathbf{c}_n^T)^T \in \mathbb{R}^{np}$ *can be solve from the linear system*

$$\Big(\lambda n I_{np} + diag(B_1, \ldots, B_n)(\mathbf{K} \otimes I_p)\Big)\mathbf{c} = Y \tag{7}$$

where I *denotes the identity matrix with the subscript representing the order,* $\mathbf{K}$ *is the* $n \times n$ *kernel matrix on the data* $\mathbf{K}_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$,

$$B_j = \sum_{i=1}^{n} W_{ij}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \in \mathbb{R}^{p\times p}, \qquad j = 1, \ldots, n$$

are $p \times p$ *matrices, and* $Y = (Y_1^T, \ldots, Y_n^T)^T \in \mathbb{R}^{np}$ *with*

$$Y_j = \sum_{i=1}^{n} W_{ij}(y_i - y_j)(\mathbf{x}_i - \mathbf{x}_j) \in \mathbb{R}^p, \qquad j = 1, \ldots, n.$$

Theorem 7 allows the kernel gradient learning algorithm to be solved efficiently when np is not very large. This is usually the case when the dimension p is small.

However, in many applications of high dimensional data analysis, p could be huge and much large than n. In this case the linear system of dimension np may be computationally difficult and techniques to further reduce the computational complexity are necessary.

Let $M = \text{span}\,\{\mathbf{x}_i - \mathbf{x}_j : i, j = 1, \ldots, n\} \subset \mathbb{R}^p$. It is not hard to to check that each $\mathbf{c}_i \in M$ from equation (7). This fact allows to further reduce the computational complexity if the dimension of the subspace M is less than p. Suppose M is of dimension d and it has an orthogonal basis $V = (v_1, \ldots, v_d) \in \mathbb{R}^{p\times d}$. Let $\mathbf{t}_{ij} \in \mathbb{R}^d$ satisfies $\mathbf{x}_i - \mathbf{x}_j = V\mathbf{t}_{ij}$. Then the coefficients $\mathbf{c}_i$ can be solved from a linear system of order nd.

Theorem 8. *Let* $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_n)^T \in \mathbb{R}^{nd}$ *where* $\boldsymbol{\gamma}_i \in \mathbb{R}^d$ *is given by the linear system*

$$\Big(\lambda n I_{nd} + diag(\tilde{B}_1, \ldots, \tilde{B}_n)(\mathbf{K} \otimes I_d)\Big)\boldsymbol{\gamma} = \tilde{Y} \tag{8}$$

where

$$\tilde{B}_j = \sum_{i=1}^{n} W_{ij}\mathbf{t}_{ij}\mathbf{t}_{ij}^T \in \mathbb{R}^{d\times d}, \qquad j = 1, \ldots, n$$

are $d \times d$ matrices, and $\tilde{Y} = (\tilde{Y}_1^T, \ldots, \tilde{Y}_n^T)^T \in \mathbb{R}^{nd}$ with

$$\tilde{Y}_j = \sum_{i=1}^{n} W_{ij}(y_i - y_j)\mathbf{t}_{ij} \in \mathbb{R}^d, \qquad j = 1, \ldots, n.$$

Then we have

$$\mathbf{c}_i = V\boldsymbol{\gamma}_i, \qquad i = 1, \ldots, n.$$

Theorem 8 is proved in Mukherjee and Zhou (2006). Notice that the dimension d of M is no larger than $\min\{n-1, p\}$. In case of $n \ll p$, the linear system (8) is of order no more than n^2.

One may criticize that nd may still be large even for a moderate n. However, we observed that the matrices in (8) are sparse. So it can be efficiently solved by the fast linear system solvers such as biconjugate gradient or LSQR methods.

3.3 Asymptotic convergence

The kernel gradient learning algorithm was shown to be consistent in Mukherjee and Zhou (2006); Mukherjee et al. (2006b).

Theorem 9. *Under certain mild conditions (see Mukherjee and Zhou (2006); Mukherjee et al. (2006b) for details), with large probability*

$$\|\hat{\mathbf{f}}_\lambda - \nabla f_*\|_{L^2_{\rho_\mathbf{x}}} \longrightarrow 0 \qquad as \;\; n \to \infty.$$

The conditions for this asymptotic convergence is mild. They include four parts: (i) regularity condition of the marginal distribution $\rho_\mathbf{x}$ such as the existence and smoothness of the density function, (ii) the smoothness of the target function f_* and its gradient, (iii) the capacity of the RKHS $\mathscr{H}_K$ which guarantees the ∇f_* can be approximated by the functions in $\mathscr{H}_K$, and (iv) correct choices of the weight function W and regularization parameter $\lambda = \lambda(n)$. Though all these conditions are basic and mild they have complicated mathematical representations. We omit the details here but refer to Mukherjee and Zhou (2006); Mukherjee et al. (2006b).

The error bounds and convergence rates is also derived. In case of the support of the marginal distribution is a p dimensional domain in $\mathbb{R}^p$ the convergence rate is of order $O(n^{-1/p})$.

3.4 Kernel gradient learning for binary classification

In binary classification setting we are given the data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ where $y_i \in \{-1, 1\}$ are labels. The target function f_* is the Bayes rule defined as $f_*(\mathbf{x}) = 1$ if $P(y = 1|\mathbf{x}) > P(y = -1|\mathbf{x})$ and $f_*(\mathbf{x}) = -1$ otherwise. Clearly f_* is not smooth and its gradient does not exist. However, we can consider the function

$$f_c(\mathbf{x}) = \log\left[\frac{\rho(y = 1|\mathbf{x})}{\rho(y = -1|\mathbf{x})}\right]$$

whose sign gives the Bayes rule f_*. Under mild conditions f_c is smooth and its gradient exists. They are central quantities in kernel gradient learning for binary classification (Mukherjee and Wu, 2006).

Let $\phi(t) = \log(1 + e^{-t})$ be the logistic loss. Then

$$f_c = \arg\min_{f \in L^2_{\rho_{\mathbf{x}}}} \mathbb{E}[\phi(yf(\mathbf{x}))].$$

Using the first order Taylor expansion of f we have

$$\mathbb{E}[\phi(yf(\mathbf{x}))] \approx \frac{1}{n}\sum_{i,j=1}^{n} W_{ij}\phi\Big(y_i\big(f(\mathbf{x}_j) + \nabla f(\mathbf{x}_j)\cdot(\mathbf{x}_i - \mathbf{x}_j)\big)\Big).$$

Define the empirical risk associated to the loss function ϕ for a real valued function $g : \mathbb{R}^p \to \mathbb{R}$ and a p-vector valued function $\mathbf{f}$ as

$$\mathcal{E}_{\phi,n}(g, \mathbf{f}) = \frac{1}{n}\sum_{i,j=1}^{n} W_{ij}\phi\Big(y_i\big(g(\mathbf{x}_j) + \mathbf{f}(\mathbf{x}_j)\cdot(\mathbf{x}_i - \mathbf{x}_j)\big)\Big).$$

The idea of gradient learning for f_c is as follows: Suppose $(\hat{g},\ \hat{\mathbf{f}})$ is the minimizer of the risk functional $\mathcal{E}_{\phi,n}(g, \mathbf{f})$. Then we expect

$$\mathcal{E}_{\phi,n}(\hat{g},\ \hat{\mathbf{f}}) = \min \mathcal{E}_{\phi,n}(g,\ \mathbf{f}) \approx \mathbb{E}(\phi(yf_c(\mathbf{x}))) \approx \mathcal{E}_{\phi,n}(f_c,\ \nabla f_c)$$

which implies $\hat{g} \approx f_c$ and $\hat{\mathbf{f}} \approx \nabla f_c$.

By incorporating the regularization in RKHS $\mathscr{H}_K$ with this idea the following kernel gradient learning algorithm for binary classification is propose in Mukherjee and Wu (2006).

Definition 10. *The classification function f_c and its gradient ∇f_c is estimated by*

$$(\hat{g}_\lambda, \hat{\mathbf{f}}_\lambda) = \arg \min_{(g,\mathbf{f}) \in \mathscr{H}_K^{p+1}} \left(\mathcal{E}_{\phi,n}(g, \mathbf{f}) + \lambda_1 \|g\|_K^2 + \lambda_2 \|\mathbf{f}\|_K^2 \right),$$

where $\lambda_1, \lambda_2 > 0$ are regularization parameters.

Similar as in the regression setting the representer theorem holds. The optimization problem for the coefficients are convex and can be efficiently solved by Newton's methods where in each iteration a linear system of type (8) is solved. For details see Mukherjee and Wu (2006).

The asymptotic error bounds and convergence rates are also given in Mukherjee and Wu (2006) under mild conditions. The rate has an order of $O(n^{-1/p})$ if the support of $\rho_{\mathbf{x}}$ is a p dimensional domain.

3.5 Manifold setting

A common belief in high dimensional data analysis is that the very high dimensional data are concentrated on a low dimensional manifold and advances in manifold learning literature confirm this belief. In such a manifold setting, the input variable is assumed to come from manifold $\mathcal{M}$ of dimension $d_{\mathcal{M}} \ll p$. We assume the existence of an isometric embedding $\varphi : \mathcal{M} \to \mathbb{R}^p$ and the observed data $(\mathbf{x}_i)_{i=1}^n$ are the image of points $(q_i)_{i=1}^N$ drawn from a distribution on the manifold: $\mathbf{x}_i = \varphi(q_i)$.

In this manifold setting the gradient learning algorithms in Definitions 5 and 10 are still valid. But the interpretation is totally different. The function $\hat{\mathbf{f}}_\lambda$ models the function $\mathrm{d}\varphi(\nabla_{\mathcal{M}} f_c)$ (or $\mathrm{d}\varphi(\nabla_{\mathcal{M}} f_*)$ for binary classification) where $\mathrm{d}\varphi$ is the differential of the map φ and $\nabla_{\mathcal{M}}$ is the gradient operator on the manifold (do Carmo, 1992). However the convergence $\hat{\mathbf{f}}_\lambda \to \mathrm{d}\varphi(\nabla_{\mathcal{M}} f_*)$ usually does not hold. Instead it is proved in Mukherjee et al. (2006b) that

$$(\mathrm{d}\varphi)^* \hat{\mathbf{f}}_\lambda \longrightarrow \nabla_{\mathcal{M}} f_*, \qquad \text{as } n \to \infty.$$

Moreover the convergence rate is of order $O(n^{-1/d_{\mathcal{M}}})$ which depends only the

intrinsic dimension $d_{\mathcal{M}}$ of the manifold $\mathcal{M}$, not on the dimension p of the ambient space. For more details we refer to Mukherjee et al. (2006b).

The manifold setting and the corresponding interpretation explain why the gradient learning algorithms are still efficient in very huge dimensional data analysis even if the sample size is small.

4 Variable selection and dimension reduction

Both variable selection and dimension reduction can be realized using the gradient estimates. Here we briefly discuss the application of kernel gradient estimates to these two tasks.

4.1 Variable selection

The key of variable (feature) selection is a ranking criterion that reflects the importance of each variable. By the discussion in Section 2, the gradient tells how fast a function changes along each coordinate variable. Hence it is natural to rank the variables by certain norms of the partial derivative, i.e., the components of the gradient. The intuition underlying this idea is that if a variable(feature) is important for prediction, then the target function f_* (or f_c in binary classification) changes fast along the corresponding coordinate and the norm of the partial derivative is large.

In the context of kernel gradient learning, the gradient is estimated by $\hat{\mathbf{f}}_\lambda = (\hat{f}_{\lambda,1}, \ldots, \hat{f}_{\lambda,p})$ and the relevance R_i of each variable x_i is measured by the empirical $L^2_{\rho_{\mathbf{x}}}$ norm of $\hat{f}_{\lambda,i}$:

$$R_i = \left(\frac{1}{n} \sum_{j=1}^{n} \left(\hat{f}_{\lambda,i}(\mathbf{x}_j) \right)^2 \right)^{1/2} \approx \|\hat{f}_{\lambda,i}\|_{L^2_{\rho_{\mathbf{x}}}}.$$

Then variables selection using this relevance ranking criterion is called gradient based feature selection (GradFS).

However, such a ranking may be incorrect because the gradient estimate $\hat{\mathbf{f}}_\lambda$ may be very rough when the dimension p is large and n is small. To overcome this difficulty we can adopt the recursive feature elimination (RFE) technique (Guyon et al., 2002). RFE is a greedy backward selection method. It starts with all features and repeatedly removes a feature until all variables have been ranked. The intuition of RFE helping improve the variable ranking lies on that a relevant variable will not be ranked as the least important variable even if the gradient estimate is rough. After the irrelevant variables are removed one by one the gradient estimate becomes better and better and the relevant features is ranked more and more accurate. The feature selection by incorporating the gradient based ranking and RFE techniques is abbreviated as GradRFE method in the sequel. We will show its effectiveness by simulations in the Section 5.

4.2 Dimension reduction

The theoretical foundation for linear dimension reduction using eigen decomposition of gradient outer production matrix has been addressed in Section 2.2. When the gradient is estimated by the kernel methods in Section 3, the gradient outer product matrix is given by

$$\hat{\mathbf{G}} = \frac{1}{n}\sum_{j=1}^{n} \hat{\mathbf{f}}_\lambda(\mathbf{x}_j)\hat{\mathbf{f}}_\lambda(\mathbf{x}_j)^T = \frac{1}{n}\big(\mathbf{c}_1, \ldots, \mathbf{c}_n\big)\mathbf{K}^2\big(\mathbf{c}_1, \ldots, \mathbf{c}_n\big)^T. \tag{9}$$

By the consistency of the gradient estimate there holds $\hat{\mathbf{G}} \to \mathbf{G}$ as n increases. This together with the perturbation theory for matrices implies the consistency of the estimate of the d.r. subspace.

Proposition 11. *Suppose the semiparametric model* (5) *holds. Let* $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ *be the eigenvalues of* $\hat{\mathbf{G}}$ *and* $\hat{\beta}_i$ *be the corresponding eigenvectors. Then* $\lambda_i \to 0$ *if and only if* $i > d$ *and*

$$\text{span}(\hat{\beta}_1, \ldots, \hat{\beta}_d) \longrightarrow \text{span}(\beta_1, \ldots, \beta_d).$$

Though the estimate of d.r. subspace by $\hat{\beta}_i,\ i = 1, \ldots, d$ converges asymptotically it may be rough with limited samples. As we have seen in Section 4.1 that the gradient estimate will be improved after some irrelevant variables are removed. We expect the accuracy of estimate of d.r. subspace is also improved after relevant feature subset selection. In this context gradient learning can be used for simultaneous feature selection and dimension reduction.

5 Simulations

The effectiveness of the variable selection and dimension reduction based on kernel gradient learning has been simulated on various artificial as well as real data sets (Mukherjee and Zhou, 2006; Mukherjee and Wu, 2006; Mukherjee et al., 2006b). Here we provide several more examples. The aim of these examples is to show the power of incorporation of gradient ranking and RFE technique and the simultaneous feature selection and dimension reduction.

In kernel gradient learning algorithms there are three parameters: the weight function, the kernel function, and the regularization parameter. Our experience is that choice of the kernel function does not have big influence on the performance. As for weight function, the Gaussian weight $W(\mathbf{x}) = \exp(-\frac{\|\mathbf{x}\|^2}{2s^2})$ with s being the median of pairwise distance of the sampling points $\mathbf{x}_i$ usually provides acceptable results (though maybe not optimal) when $p > n$ or they are comparable. In case of $n \gg p$ one could choose a slightly smaller s. In our following simulations we do not optimize these two parameters. The regularization parameter will chosen by minimizing the cross validation error defined as

$$CV(\lambda) = \frac{1}{n} \sum_{i,j=1}^{n} W_{ij} \Big(y_i - y_j - \hat{\mathbf{f}}_\lambda^{\setminus i}(\mathbf{x}_j) \cdot (\mathbf{x}_i - \mathbf{x}_j) \Big)^2$$

in regression setting where $\hat{\mathbf{f}}_\lambda^{\setminus i}$ is the the solution of kernel gradient learning with the i-th sample removed. For binary classification a similar cross validation error criterion is used.

variable	x_1	x_2	x_3	x_4	x_5
GradFS	48	100	100	100	100
GradRFE	100	100	100	100	100

Table 1: Frequencies of variables $x_1,\ x_2,\ x_3,\ x_4,\ x_5$ ranked as the top 5 important variables by GradFS and GradRFE in 100 repeats.

5.1 An artificial example

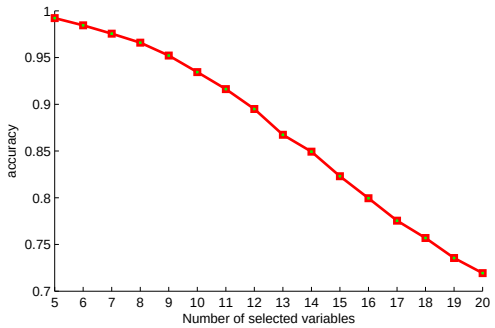
Let $\rho_{\mathbf{x}}$ be uniform on $[0,\ 1]^p$ and

$$y = (x_1 - 0.5)^2 + x_2 + x_3 + x_4 + x_5 + \epsilon$$

where $\epsilon \sim 0.05N(0, 1)$ and N denotes normal distribution. It was pointed out in Turlach (2004) that LARS and LASSO miss the first variable x_1 though it is relevant. This coincides our discussion in Section 2 for $\mathbb{E}[(y - \bar{y})(x_1 - \bar{x}_1) = \mathbb{E}[y(x_1 - 0.5)] = 0$.

In our experiments we take $p = 20$, $n = 100$, and a quadratic kernel $K(\mathbf{x}, \mathbf{u}) = (1 + \mathbf{x} \cdot \mathbf{u})^2$ is used. The experiments are repeated 100 times. We reported the frequencies of $x_1,\ x_2,\ x_3,\ x_4,\ x_5$ selected by the top 5 important variables by gradient feature selection and gradient RFE in Table 5.1. In the experiments we noticed that $x_2,\ x_3,\ x_4,\ x_5$ are always ranked correctly as the top 4 important variables by GradFS showing they are easy to be selected. Theoretically the variable x_1 should be ranked as the fifth important variable. In our simulations it is correctly ranked 48 times in 100 repeats without using RFE technique and all 100 times using RFE technique. This shows that gradient has the ability to identify the nonlinear linear features. Moreover, with limited samples and high dimensions the gradient estimate is rough and the ranking is unstable, while RFE technique can help reduce such instability.

We next study the performance of dimension reduction based on eigen decomposition of the gradient outer product matrix. The d.r. subspace $\mathcal{S}$ is spanned by $\beta_1 = (1,\ 0,\ \ldots,\ 0)^T$ and $\beta_2 = (0,\ 0.5,\ 0.5,\ 0.5,\ 0.5,\ 0,\ \ldots,\ 0)^T$. We measure



the accuracy of the estimated d.r. space $\hat{\mathcal{S}} = \text{span}(\hat{\beta}_1, \hat{\beta}_2)$ by

$$\text{accuracy}(\hat{S}) = \frac{1}{2}\left(\|P_{\mathcal{S}}(\hat{\beta}_1)\|^2 + \|P_{\mathcal{S}}(\hat{\beta}_2)\|^2\right).$$

In Figure 5.1 we report the mean accuracy of 100 repeats for different selected feature subset. The result shows that when there are irrelevant variables, dimension reduction after feature subset selection performs much better than that without feature subset selection.

We applied several state-of-the-art dimension reduction methods in statistical literature to this example for comparison (results not shown). It turns out that SIR misses β_1 and PHD misses β_2. MAVE performs slightly better than eigen decomposition of the gradient outer product matrix without feature selection but still much worse than that after feature subset selection.

5.2 Application to gene expression data

One problem domain where high-dimensions are ubiquitous is the analysis and classification of gene expression data. In Mukherjee and Zhou (2006); Mukherjee and Wu (2006) the feature selection based on gradient estimates are used to the discrimination of acute myeloid leukemia (AML) from acute lymphoblastic leukemia (ALL) using expression data (Golub et al., 1999). Since this data is well linearly separable, GradFS works very well and RFE does not help a lot.

Here we consider another gene expression data: the classification of prostate cancer (Singh et al., 2002). In this data the number of genes is $p = 12600$. The

training data contains $n = 102$ samples, 52 tumor samples and 50 non-tumor samples. The test data is an independent data set and contains 34 samples from a different experiment. We preprocess the data as follows: (i) Since the overall microarray intensities in the training set and the test set have nearly ten fold differences, we multiply a constant to the test data so that the medians of the microarray intensities of both sets are the same. (ii) We restrict the microarray intensity between 1 and 5000 and (iii) take the log transformation. Then we apply GradFS and GradRFE to the training data and rank the features. The prediction is made by linear support vector machine classifier. The performance is measured by both the leave one out (LOO) error over the training set and the classification error for the test set. The result is reported in Tables 5.2. We see that GradRFE is more efficient and stable than GradFS. By GradRFE the top 16 genes provides the classification accuracies 100% and 94% for training data and test data respectively. Recall that if the genes are ranked using the variation of signal-to-noise metric (Golub et al., 1999) the accuracies by using the top 16 genes are 90% and 86% respectively (Singh et al., 2002). This verifies the superiority of gradient based variable ranking and RFE technique for this data set.

6 Conclusions and discussions

We reviewed the kernel gradient learning algorithms proposed in Mukherjee and Zhou (2006); Mukherjee and Wu (2006) and discussed their applications in feature subset selection and dimension reduction. For feature subset selection, this is done in a nonlinear framework. The gradient estimate is showed to be able to retrieve the variables that are missed by linear methods. When only limited samples are available and the gradient estimate is rough, recursive feature elimination helps to improve the accuracy of variable ranking and feature selection. For dimension reduction the gradient method is asymptotically consistent and non-degenerate. Simultaneous dimension reduction and feature selection may provide much better estimates of the d.r. subspace when there are irrelevant variables.

Among these discussions three facts have not been noticed and fully addressed

number of genes	GradRFE		GradFS	
	LOO error	Test error	LOO error	Test error
1	14/ 102	1 / 34	21 / 102	13 / 34
2	11 / 102	1 / 34	15 / 102	8 / 34
4	6 / 102	1 / 34	11 / 102	6 / 34
8	3 / 102	2 / 34	2 / 102	1 / 34
16	0 / 102	2 / 34	8 / 102	2 / 34
32	0 / 102	2 / 34	4 / 102	4 / 34
64	0 / 102	3 / 34	3 / 102	4 / 34
128	1 / 102	4 / 34	2 / 102	5 / 34
256	1 / 102	3 / 34	2 / 102	4 / 34
512	1 / 102	4 / 34	2 / 102	7 / 34
1024	0 / 102	7 / 34	1 / 102	7 / 34
2048	1 / 102	6 / 34	0 / 102	6 / 34
4096	2 / 102	5 / 34	0 / 102	6 / 34
8192	5 / 102	5 / 34	5 / 102	5 / 34
16000	9 / 102	5 / 34	9 / 102	5 / 34

Table 2: Classification error for prostate cancer data by linear SVM classifier after feature subsection using GradFS and GradRFE.

in Mukherjee and Zhou (2006); Mukherjee and Wu (2006). Firstly, feature subset selection by learning gradient has a more solid foundation in the framework of nonlinear models. Secondly, the sparsity allows the gradient algorithm to be solved more efficiently and makes it applicable to problems with moderate sample size. Lastly, the incorporating of RFE technique can greatly improve the accuracy of the feature ranking.

An important issue that has not been well studied for gradient feature selection is the determination of the number of relevant features. In applications when feature subset selection is used as preprocess, this can be done by incorporating the inference step. However, it should be interesting to develop a criterion to determine the number of relevant features independently.

Several extensions have been proposed recently. In Cai and Ye (2008) the lasso type regularization is introduced to the kernel gradient learning that estimate the gradient ∇f_* by

$$\hat{\mathbf{f}}_\lambda = \arg\min_{\mathbf{f}\in\mathscr{H}_K^p} \Big(\mathcal{E}_n(\mathbf{f}) + \lambda \sum_{i=1}^{p} \|f_i\|_K\Big).$$

Its advantage is automatically selection of feature subset. The other extension is given in Ying and Campbell (2008) where the vector valued kernels are introduced to the kernel gradient learning. Recall that the $\hat{\mathbf{f}}_\lambda \neq \nabla \hat{g}_\lambda$ in classification setting. But suitable choice of vector valued kernels ensures $\hat{\mathbf{f}}_\lambda = \nabla \hat{g}_\lambda$. This is of great mathematical interest though it usually does not improve the performance of feature selection and dimension reduction in applications. When a large number of samples are available, the linear system or optimization of the gradient learning algorithms becomes difficult. To overcome this difficulty, online gradient learning algorithms by a gradient descent method were proposed in Dong and Zhou (2008); Cai et al. (2008). They are computationally efficient and has comparable asymptotic convergence rates.

References

N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68(6): 337–404, 1950.

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

A. L. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1-2):245–271, 1997. ISSN 0004-3702.

J. Cai, H. Y. Wang, and D. X. Zhou. Gradient learning in a classification setting by gradient descent. preprint, 2008.

J.-F. Cai and G.-B. Ye. Variable selection and linear feature construction via sparse gradients. preprint, 2008.

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

R. Coifman, S. Lafon, A. Lee, M. Maggioni, B. Nadler, F. Warner, and S. 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. Coifman, S. Lafon, A. Lee, M. Maggioni, B. Nadler, F. Warner, and S. 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. Cook. *Regression Graphics: Ideas for Studying Regressions Through Graphics*. Wiley, 1998.

R. Cook and B. Li. Dimension reduction for conditional mean in regression. *Ann. Stat.*, 30(2):455–474, 2002.

R. Cook and X. Yin. Dimension reduction and visualization in discriminant analysis (with discussion). *Aust. N. Z. J. Stat.*, 43(2):147–199, 2001.

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

X. M. Dong and D. X. Zhou. Learning gradients by a gradient descent algorithm. *J. Math. Anal. Appl.*, 341:1018–1027, 2008.

D. Donoho and C. Grimes. Hessian eigenmaps: new locally linear embedding techniques for highdimensional data. *PNAS*, 100:5591–5596, 2003.

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

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Ann. Statist.*, 32(2):407–499, 2004. ISSN 0090-5364. With discussion, and a rejoinder by the authors.

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

T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, and E. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, 2002.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer-Verlag, New York, 2001. ISBN 0-387-95284-5. Data mining, inference, and prediction.

L. Hermes and J. Buhmann. Feature selection for support vector machines. *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, 2:712–715 vol.2, 2000. doi: 10.1109/ICPR.2000.906174.

R. Kohavi and G. John. Wrappers for feature selection. *Artificial Intelligence*, 97 (1-2):273–324, 1997.

J. Lafferty and L. Wasserman. Rodeo: Sparse, greedy nonparametric regression. *The Annals of Statistics*, 36(1):28–63, 2008.

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

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

Y. Liu and Y. Wu. Variable selection via a combination of l_0 and l_1 penalties. *Journal of Computational and Graphical Statistics*, 16(4):782–798, 2007.

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

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

S. Mukherjee, , Q. Wu, and M. Maggioni. Theoretical comparisons between supervised dimension reduction approaches. Technical report, ISDS, Duke Univ., 2006a.

S. Mukherjee, Q. Wu, and D. Zhou. Learning gradient on manifolds. Technical report, ISDS, Duke University, 2006b.

A. Rakotomamonjy. Variable selection using svm-based criteria. *Journal of Machine Learning Research*, 3:1357–1370, 2003.

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

B. Schölkopf, A. J. Smola, and K. Müller. Kernel principal component analysis. In W. Gerstner, A. Germond, M. Hasler, and J.-D. Nicoud, editors, *Artificial Neural Networks ICANN'97*, volume 1327 of *Springer Lecture Notes in Computer Science*, pages 583–588, Berlinpp, 1997.

D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie, E. S. Lander, M. Loda, P. W. Kantoff, T. R. Golub, and W. R. Sellers. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1:203–209, 2002.

C. Stein. Estimation of the mean of a multivariate normal distribution. *Ann. Stat.*, 9:1135–1151, 1981.

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. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996. ISSN 0035-9246.

B. Turlach. Discussion of "Least angle regression" by efron, hastie, jonstone and tibshirani. *The Annals of Statistics*, 32:494–499, 2004.

V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik. Feature selection for svms. In *Advances in Neural Information Processing Systems*, volume 13, 2001.

Q. Wu, J. Guinney, M. Maggioni, and S. Mukherjee. Learning gradients: predictive models that infer geometry and dependence. Technical report, ISDS, Duke University, 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.

Y. Ying and C. Campbell. Learning coordinate gradients with multi-task kernels. In *COLT*, 2008.

H. H. Zhang. Variable selection for support vector machines via smoothing spline anova. *Statistica Sinica*, 16:659–674, 2006.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(2):301–320, 2005. ISSN 1369-7412.