Biostatistics (2011), **12**, 3, *pp.* 429–444
doi:10.1093/biostatistics/kxq081
Advance Access publication on January 20, 2011

Biological pathway selection through nonlinear dimension reduction

HONGJIE ZHU, LEXIN LI*

Bioinformatics Research Center and Department of Statistics,
North Carolina State University, Box 8203, Raleigh, NC 27695, USA
li@stat.ncsu.edu

SUMMARY

In the analysis of high-throughput biological data, it is often believed that the biological units such as genes behave interactively by groups, that is, pathways in our context. It is conceivable that utilization of priorly available pathway knowledge would greatly facilitate both interpretation and estimation in statistical analysis of such high-dimensional biological data. In this article, we propose a 2-step procedure for the purpose of identifying pathways that are related to and influence the clinical phenotype. In the first step, a nonlinear dimension reduction method is proposed, which permits flexible within-pathway gene interactions as well as nonlinear pathway effects on the response. In the second step, a regularized model-based pathway ranking and selection procedure is developed that is built upon the summary features extracted from the first step. Simulations suggest that the new method performs favorably compared to the existing solutions. An analysis of a glioblastoma microarray data finds 4 pathways that have evidence of support from the biological literature.

Keywords: Dimension reduction; Generalized additive model; Kernel methods; Pathway selection; Sliced inverse regression.

1. INTRODUCTION

In biomedical research, high-throughput technologies are generating massive amounts of high-dimensional data. A representative example is complementary DNA microarray, which simultaneously measures expressions of thousands or tens of thousands of genes in a single experiment. These fast-developing techniques have created a research area called the "omics" studies, which aim at examining the whole spectrum of basic biological units such as genes, proteins, and metabolites. Examples include genomics that studies the whole genome, that is, the sum of all genes of an individual organism, proteomics that studies proteome, that is, the entire complement of proteins produced in an organism, specific type of tissues or other biological system, and metabolomics that studies metabolome, that is, the entire repertoire of metabolites present in cells and/or tissues, which represents the final product of certain cellular processes. An important question in such studies is to identify biological units, that is, genes, proteins, or metabolites, that are related to and influence various clinically relevant phenotypes, for instance, survival outcome from cancer treatment or risk of developing cancers. Given the very high dimensionality

*To whom correspondence should be addressed.

and often the limited number of sample units of such “omics” data, new challenges arise for conventional statistical analysis of the data. As a consequence, it has motivated developments of many new statistical methods; see Bickel *and others* (2009) for an excellent review and the references therein.

Most existing approaches treat genes, proteins, or metabolites individually and as such ignore any prior biological knowledge that characterizes interrelationship among those biological units. It is commonly believed that most biological systems operate not based on just a single unit but instead on their mutual interactions and regulations. Attesting to this belief, recent studies have suggested that clinical outcomes of complex diseases are often associated with multiple genes rather than a single one. This has led researchers to form and define clusters of genes, proteins, and metabolites, referred to as “pathways,” that are believed to function in a coordinated and interactive fashion. Years of intensive biomedical research has accumulated an immense wealth of pathway knowledge, which is primarily available through some well-known pathway databases, for instance, Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000), Reactome (Matthews *and others*, 2008), and BioCyc (Karp *and others*, 2005). Intuitively, employment of such biological pathway knowledge would greatly facilitate our statistical analysis because it offers a means to reduce the potentially enormous modeling space by focusing on biologically meaningful interactions among genes, proteins, or metabolites in related pathways. In addition, the outcomes are expected to be more readily interpretable biologically when taking into account the pathway information.

In the light of this observation, there have been some recent developments of pathway-based statistical analysis that search for pathways that are associated with disease phenotypes. Early efforts have focused on gene set enrichment analysis, that is, to identify pathways that are overrepresented by differentially expressed genes (Subramanian *and others*, 2005), which provides a useful and informative tool (Tian *and others*, 2005). However, such analysis treats each pathway separately without accounting for possible interactions between pathways. Wei and Li (2007) proposed a nonparametric pathway–based regression (NPR) model that considers multiple pathways simultaneously and allows complex interactions among genes within the pathways. To fit their NPR model, they employed a gradient descent boosting algorithm and used classification and regression tree as the base learner. However, the algorithm only produces a ranking based on relative importance of the pathways rather than an explicit selection of relevant pathways. Luan and Li (2008) considered a group additive regression (GAR) model along with a group gradient descent boosting algorithm to identify pathways related to clinical outcomes. But again, their procedure only gives a pathway ranking. Besides, the model only permits linear combinations of genes within the pathways since their choice of the base learner is a linear model (LM). More recently, Ma and Kosorok (2009) used principal components analysis (PCA) as the first step to combine genes in the same pathways into a small number of summary features. They then built statistical models associating the induced summary features with the phenotypes. It is well known, however, that PCA is an unsupervised dimension reduction technique that does not take into account the response information. For this reason, as we will show later, PCA could yield inferior results.

In this article, we propose a method of nonlinear dimension reduction followed by a generalized additive model (GAM) or a generalized linear model (GLM) with L_1 regularization. The goal is to identify informative pathways that are related to the phenotype of interest. In the first step, we couple a reproducing kernel map with a family of sufficient dimension reduction (SDR) estimators to produce summary features of constituent units for each pathway. This step turns the high-dimensional data into its low-dimensional projections. Due to flexibility of kernels and their computational advantages, the new dimension reduction method allows complex interactive relations among biological units and can also handle a very large number of units, even if it far exceeds the number of observations. In addition, the proposed method does not impose any strong parametric model assumptions in the phase of dimension reduction, and as such it grants full flexibility in subsequent model formulation and selection. Our method is motivated by kernel sliced inverse regression (SIR) first proposed by Wu (2008) and Wu *and others* (2008), but we extend their work to a whole family of dimension reduction methods. We also derive a generalized cross-validation (GCV) criterion for the regularized estimation. In the second step, we build a GAM or a GLM based on the induced summary features from all the pathways. We produce both a ranking of pathways by employing an L_1 regularization, as well as an explicit pathway selection by proposing a pseudo pathway selection strategy. Compared with Ma and Kosorok (2009), the main difference lies in the step of dimension reduction. Our method extends PCA in the sense that it takes into account the response information during the reduction, and it permits complex nonlinear associations of genes when producing the summary features. Compared with Wei and Li (2007) and Luan and Li (2008), our method produces both pathway ranking and selection. Our simulation studies confirm the advantages of our proposal over those existing solutions. An analysis of a glioblastoma microarray data identifies 4 informative pathways that have good evidences in the biology literature to support possible associations with the disease.

2. Nonlinear dimension reduction

2.1 *Linear dimension reduction*

For a regression of a response Y given a p-dimensional predictor X, SDR seeks a minimum number of linear combinations, $\eta_1^\top X, \dots, \eta_d^\top X$, such that

$$Y \perp\!\!\!\perp X | (\eta_1^\top X, \dots, \eta_d^\top X). \tag{2.1}$$

The space spanned by $\eta = (\eta_1, \dots, \eta_d)$, called the central subspace and denoted as $\mathcal{S}_{Y|X}$, uniquely exists under minor conditions (Cook, 1996). Given (2.1), we can replace the original p-dimensional X with now d-dimensional $\eta^\top X$. In practice, d is often much smaller than p and thus dimension reduction is achieved. For ease of reference, we call $(\eta_1^\top X, \dots, \eta_d^\top X)$ the sufficient predictors, which will serve as the induced summary features in subsequent modeling.

There have been many methods proposed to estimate $\mathcal{S}_{Y|X}$, most of which can be formulated as a generalized eigen decomposition problem. Specifically, an estimate of a basis of $\mathcal{S}_{Y|X}$ can be obtained by the first d eigenvectors η_j's that correspond to the nonzero eigenvalues λ_j's in a descending order from the decomposition,

$$\Omega_x \eta_j = \lambda_j \Sigma_x \eta_j, \quad j = 1, \dots, d, \tag{2.2}$$

where $\Sigma_x = \mathrm{Cov}(X)$ and Ω_x is a method-specific $p \times p$ semi-positive definite matrix. Some representative estimators include SIR (Li, 1991), where $\Omega_x = \mathrm{Cov}\{E(X|Y) - E(X)\}$; sliced average variance estimation (SAVE) (Cook and Weisberg, 1991), where $\Omega_x = E[\{\Sigma_x - \mathrm{Cov}(X - E(X)|Y)\}^2]$; and directional regression (Li and Wang, 2007), where $\Omega_x = 2E[\{\Sigma_x - \mathrm{Cov}(X - E(X)|Y)\}^2] + 2\mathrm{Cov}^2\{E(X|Y) - E(X)\} + 2E[\{E(X|Y) - E(X)\}^\top\{E(X|Y) - E(X)\}]\mathrm{Cov}\{E(X|Y) - E(X)\}$. All those methods involve the inverse moments $E(X|Y)$ and $\mathrm{Cov}(X|Y)$. To estimate those quantities, we often first partition the sample space of Y into H nonoverlapping intervals, or say slices, and then obtain the sample average or sample covariance within each slice. It is also interesting to note that, most of those SDR methods impose no parametric assumption on $Y|X$. Instead, they require the marginal distribution of X to satisfy that $E(X|\eta^\top X)$ is linear in $\eta^\top X$. This is often viewed as a mild condition since it holds when X is elliptically symmetric and is approximately true when p goes to infinity. In this article, we assume the condition holds approximately since we are dealing with a very large p.

SDR methods following (2.1) and (2.2) yield "linear" dimension reduction because the reduction admits the form of linear combinations of X. This could have some limitations. Consider an illustrative example, where $X = (X_1, \dots, X_6)$ and

$$Y = X_1 + X_2X_3 + X_4^2 + X_5X_6 + \varepsilon, \tag{2.3}$$

with an independent error ε. Then $\mathcal{S}_{Y|X} = \mathbb{R}^6$, with no reduction in dimension possible. Another limitation associated with the SDR methods in (2.2) is that one needs to invert a $p \times p$ covariance matrix Σ_x. When the number of predictors p exceeds the sample size n, one cannot invert the sample estimator of Σ_x. To address this $n < p$ issue, there have been proposals employing the ridge regression idea (Li and Yin, 2008) or the partial least squares (PLSs) idea (Li *and others*, 2007; Cook *and others*, 2007). However, the proposed remedies are very computationally intensive when p becomes very large. Next, we consider a "nonlinear" dimension reduction strategy to address those limitations.

2.2 *Nonlinear dimension reduction using kernel methods*

In the usual kernel-based methods, a set of features are chosen that define a space $\mathcal{F}$, where it is hoped relevant structure will be revealed. The data $(x_1, \ldots, x_n)$ in the input space $\mathcal{X}$ are then mapped to the feature space $\mathcal{F}$ using a mapping $\phi: \mathcal{X} \to \mathcal{F}$, and classification, regression, or clustering is performed in $\mathcal{F}$ using traditional methods. If $\mathcal{F}$ is chosen to be an inner product space and if one defines the kernel function k, with the associated Gram matrix $K \in \mathbb{R}^{n\times n}$ as $K_{ij} = k(x_i, x_j) = \langle \phi(x_i), \phi(x_j)\rangle$, then any algorithm whose operations can be expressed in terms of inner products in the input space can be generalized to an algorithm that operates in the feature space by substituting a kernel function for the inner product. Using the kernel k instead of an inner product in the input space corresponds to mapping the data into a high-dimensional inner product space $\mathcal{F}$ by a usually nonlinear mapping ϕ and taking inner products there, that is, $k(x_i, x_j) = \langle \phi(x_i), \phi(x_j)\rangle$. Common choices of kernels include the polynomial kernel $k(x, x') = (1 + x^\top x')^r$ for some positive integer r, and the Gaussian kernel $k(x, x') = \exp\{-\|x - x'\|^2/(2\sigma^2)\}$.

In our context of nonlinear dimension reduction, the basic idea is to carry out a linear dimension reduction in the space of $\phi(X)$, which in effect results in a nonlinear dimension reduction in the original input space of X. The well-known kernel trick turns the primal problem that depends on the dimension of $\phi(X)$, which is high or even infinite, to a dual problem that only depends on the sample size. As such, the method works for any p regardless of n.

Specifically, in analogy to linear reduction in (2.1), nonlinear dimension reduction seeks

$$Y \perp\!\!\!\perp X | (\langle \beta_1, \phi(X)\rangle, \ldots, \langle \beta_{\tilde{d}}, \phi(X)\rangle). \tag{2.4}$$

Comparing with (2.1), the linear combinations $(\eta_1^\top X, \ldots, \eta_d^\top X)$ are replaced by $\tilde{d}$ inner products $(\langle \beta_1, \phi(X)\rangle, \ldots, \langle \beta_{\tilde{d}}, \phi(X)\rangle)$, and Y depends on X only through those inner products. We again refer them as the sufficient predictors and assume $\tilde{d} \leqslant \min(n, p)$.

In terms of estimation, conceptually, one can estimate β's in a way analogous to (2.2), that is, through the eigen decomposition

$$\Omega_\phi \beta_j = \rho_j \Sigma_\phi \beta_j, \quad j = 1, \ldots, \tilde{d}, \tag{2.5}$$

where $\Sigma_\phi = \text{Cov}\{\phi(X)\}$ and Ω_ϕ is defined similarly as Ω_x except that we replace X with $\phi(X)$. See Wu *and others* (2008) for a theoretical justification of (2.5) for SIR, while similar arguments apply to other SDR estimators. Given $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, estimation of β_j's is obtained by replacing Ω_ϕ and Σ_ϕ in (2.4) with their corresponding sample counterparts $\hat{\Omega}_\phi$ and $\hat{\Sigma}_\phi$, that is,

$$\hat{\Omega}_\phi \beta_j = \rho_j \hat{\Sigma}_\phi \beta_j, \quad j = 1, \ldots, \tilde{d}. \tag{2.6}$$

On the other hand, depending on the choice of the kernel function k, the dimension of the induced mapping $\phi(X)$ can be very high, sometimes even infinite. As such, a direct decomposition through (2.6) is not feasible computationally.

The problem can be solved by noting that the target of nonlinear dimension reduction estimation are the inner products $\langle \beta_j, \phi(X) \rangle$ rather than β_j themselves. These inner products can be obtained by solving a dual problem to (2.6),

$$\tilde{K} J \tilde{K} \alpha_j = \rho_j \tilde{K}^2 \alpha_j, \quad j = 1, \dots, \tilde{d}, \tag{2.7}$$

where $\tilde{K} \in \mathbb{R}^{n \times n}$ is the centered Gram matrix with the (i, j)th element $\tilde{K}_{ij} = K_{ij} - K_{i\cdot} - K_{\cdot j} + K_{\cdot\cdot}$, $K_{\cdot\cdot}$ is the mean of K and $K_{i\cdot}$ and $K_{\cdot j}$ are the means of the ith row and jth column of K. J is a method-specific $n \times n$ matrix that takes the form,

$$J_{\text{SIR}} = \sum_{h=1}^{H} n_h^{-1} b_h b_h^\top,$$

$$J_{\text{SAVE}} = \sum_{h=1}^{H} B_h \tilde{K} B_h, \quad B_h = n^{-1} n_h^{1/2} I_n - n^{-2} n_h^{1/2} \mathbf{1}_n \mathbf{1}_n^\top - n_h^{-1/2} \text{diag}(b_h) + n_h^{-3/2} b_h b_h^\top,$$

$$J_{\text{DIR}} = 2 J_{\text{SAVE}} + 2n^{-1} J_{\text{SIR}} \tilde{K} J_{\text{SIR}} + 2n^{-1} \left(\sum_{h=1}^{H} n_h^{-1} b_h^\top \tilde{K} b_h \right) J_{\text{SIR}},$$

where $\mathbf{1}_n$ is an $n \times 1$ vector with all elements equal to 1, b_h is an $n \times 1$ vector with its ith element equal to 1 if y_i belongs to the hth slice and 0 otherwise, n_h is the number of observations in the hth slice, and I_n is an n-dimensional identity matrix. Then for a new observation $x \in \mathcal{X}$,

$$\langle \beta_j, \phi(x) \rangle = \alpha_j^\top [\tilde{k}(x_1, x), \dots, \tilde{k}(x_n, x)]^\top \tag{2.8}$$

where $\tilde{k}(x_i, x) = k(x_i, x) - n^{-1} \sum_{l=1}^{n} k(x_l, x), i = 1, \dots, n$. So the inner product $\langle \beta_j, \phi(x) \rangle$ can be obtained from the kernel k and α_j's from (2.7). An outline of derivation of the above kernel version of SDR estimators is given in the online appendix available at *Biostatistics* online.

Furthermore, to induce numerical stability to the eigen decomposition (2.7), we follow Wu *and others* (2008) to introduce a ridge regularization,

$$\tilde{K} J \tilde{K} \alpha_j = \rho_j (\tilde{K}^2 + n^2 \tau I_n) \alpha_j, \quad j = 1, \dots, \tilde{d}, \tag{2.9}$$

where $\tau \geqslant 0$ is a ridge parameter. We will discuss tuning of τ in the next section.

Due to (2.7), the proposed nonlinear dimension reduction only involves decomposition of an $n \times n$ matrix, so it can handle $n < p$. Its flexible reduction form beyond the linear combination is also expected to facilitate dimension reduction. As a simple illustration, we reconsider the example (2.3). If one employs a quadratic kernel, then only one linear combination in the mapped feature space is needed to summarize all regression information and thus substantial reduction is achieved. Finally, if a linear kernel is employed, the proposed method reduces to the usual linear dimension reduction.

2.3 *Tuning*

There are 2 tuning parameters in the aforementioned nonlinear dimension reduction procedure: one is the number of slices H, and the other is the ridge parameter τ.

We first consider a variant of dimension reduction estimator of Zhu *and others* (2010), which allows one to avoid the tuning parameter H. The basic idea is to always dichotomize the sample space of Y by the observed y_i, $i = 1, \dots, n$, and then substitute $\mathbf{\Omega}_x$ with the average of $\mathbf{\Omega}_x$ for all such binary partitions in (2.2). In our context of nonlinear dimension reduction, the solution continues to admit the eigen decomposition of (2.9) except that the J matrix now takes the form

$$J = n^{-1} \sum_{i=1}^{n} J_i, \tag{2.10}$$

where J_i is the J matrix derived by performing nonlinear dimension reduction for the ith dichotomization. Also note that (2.10) is applicable for all the aforementioned kernel SDR estimators.

Next, we develop a GCV criterion to choose the ridge parameter τ. Consider the eigen decomposition $n^{-1}\tilde{K} J \tilde{K} = \sum_{j=1}^{m} u_j \upsilon_j \upsilon_j^{\top}$, where $u_1 \geqslant \cdots \geqslant u_m$ are the eigenvalues in descending order, $\upsilon_1, \ldots, \upsilon_m$ are the corresponding eigenvectors, and m is the number of nonzero eigenvalues. Then following Li and Yin (2008), the solution α_j from the eigen decomposition (2.9) can be equivalently obtained by minimizing the objective function,

$$L(A, C) = \sum_{j=1}^{m} u_j \|\upsilon_j - n^{-1}\tilde{K}^2 A c_j\|^2 + n\tau \operatorname{vec}(AC)^{\top}\{U \otimes (n^{-1}\tilde{K}^2)\}\operatorname{vec}(AC),$$

where $A \in \mathbb{R}^{n\times\tilde{d}}$ and $C = (c_1, \ldots, c_m) \in \mathbb{R}^{\tilde{d}\times m}$, $U = \operatorname{diag}(u_1, \ldots, u_m)$, $\otimes$ stands for the kronecker product, and $\operatorname{vec}(\cdot)$ is a matrix operator that stacks all columns of a matrix into a vector. Letting $(\hat{A}, \hat{C}) = \arg\min L(A, C)$, then $\operatorname{span}(\hat{A}) = \operatorname{span}(\alpha_1, \ldots, \alpha_{\tilde{d}})$. Note that minimizing $L(A, C)$ is a least squares type problem. As such we can derive a GCV criterion for choosing τ, following Li and Yin (2008),

$$\frac{\|(I_{nm} - Q)(U^{1/2} \otimes I_n)\operatorname{vec}(V)\|^2}{nm\{1 - \operatorname{trace}(Q)/(nm)\}^2}, \tag{2.11}$$

where $V = (\upsilon_1, \ldots, \upsilon_m) \in \mathbb{R}^{n\times m}$,

$$Q = \{U^{1/2}\hat{A}(\tau)^{\top}(\hat{A}(\tau)U\hat{A}(\tau)^{\top})^{-1}\hat{A}(\tau)U^{1/2}\} \otimes \{\tilde{K}(\tilde{K}^2 + n^2\tau I_n)^{-1}\tilde{K}\},$$

and $\hat{A}(\tau)$ denotes the minimizer of $L(A, C)$ for a given τ. We choose τ such that (2.11) attains its minimum.

3. Pathway ranking and selection

3.1 *Groupwise dimension reduction*

Biological units such as genes, proteins, and metabolites have inherent pathway structures, and units in the same pathway often have coordinated functions in affecting phenotype activities. In this article, we focus on gene pathway, while the methodology applies to other biological pathways as well. We will adopt a simple view of gene pathways by treating each pathway as a static gene cluster. This view has been adopted in many gene pathway studies, for example, Luan and Li (2008), Ma and Kosorok (2009), Pang and Zhao (2008), Shi and Ma (2008), and Wei and Li (2007).

Specifically, our goal is to identify pathways that are relevant to phenotype activities and to model their effects. For that purpose, we first construct gene pathways using information retrieved from public databases; in our study, we use KEGG (http://www.genome.ad.jp/kegg). We then divide genes into groups by the pathways and conduct nonlinear dimension reduction of the phenotype given all the genes in that pathway. Li (2009) noted that some additional conditions are needed to ensure such groupwise dimension reduction to preserve full information. However, the sensitivity analysis in Li (2009) also suggested that, with a reasonable sample size, this groupwise reduction strategy often works satisfactorily. For simplicity, we adopt this strategy in our analysis. For each pathway, we extract one or a few sufficient predictors $\langle \beta_j, \phi(x)\rangle$'s as summary features. Then we fit a GLM or a GAM to connect the phenotype with all the pathways under study. We emphasize that dimension reduction is an important intermediate step in this procedure. It brings down the dimensionality of the data to a much lower and manageable scale, and it grants full flexibility in subsequent model building and selection. Without dimension reduction, it would have been far more difficult, and sometimes even intractable, to apply the conventional methods like GLM and GAM. We also point out that our solution is similar in spirit to Ma and Kosorok (2009), who summarized pathways using the first few principal components of genes. Unlike the PCA-based method, however, our approach takes into account the response information directly in the process of dimension reduction, whereas PCA does not. Moreover, our method permits flexible interactive relations among genes, while PCA is often restricted to main effects only, or at most to second-order interactions, due to its computational feasibility.

In practice, we often select a small number of summary features for each pathway. In our simulation studies in Section 4, we choose only the leading sufficient predictor. It is not our intention to suggest that one summary feature would always be sufficient for all data analysis. However, based on our limited experiences, we do find that it is often the case that a small number of sufficient predictors are good enough, especially for a nonlinear dimension reduction method. Moreover, a small number of summary features are also easier for the interpretation purpose. A similar view can be found in the schematic model of Chatterjee *and others* (2006) in the genetics studies.

3.2 *Model-based pathway selection after dimension reduction*

Suppose there are G pathways. We denote the summary features from the gth pathway as $Z_{g1}, \ldots, Z_{gd_g}$, for $g = 1, \ldots, G$. We then fit a GLM (McCullagh and Nelder, 1989) or a GAM (Hastie and Tibshirani, 1990) of Y on Z's,

$$g\{E(Y|Z)\} = \theta_0 + \theta_{11} Z_{11} + \cdots + \theta_{1d_1} Z_{1d_1} + \cdots + \theta_{G1} Z_{G1} + \cdots + \theta_{Gd_G} Z_{Gd_G},$$

$$g\{E(Y|Z)\} = \theta_0 + f_1(Z_{11}, \ldots, Z_{1d_1}; \theta_1) + \cdots + f_G(Z_{G1}, \ldots, Z_{Gd_G}; \theta_G),$$

where $g(\cdot)$ is a known link function and f_g's are unspecified smooth functions.

For the purpose of both pathway ranking and selection, we employ L_1 regularization in GLM and GAM. For GLM, we adopt the group Lasso penalty of Yuan and Lin (2006) since the presence or absence of a pathway in the model depends on the entire group of parameters $(\theta_{g1}, \ldots, \theta_{gd_g})$, so they should be shrunk to zero simultaneously. For GAM, we adopt the 2-step procedure of nonnegative garrote (Breiman, 1995; Yuan and Lin, 2007). That is, we first obtain the GAM estimates $f_g(Z_{g1}, \ldots, Z_{gd_1}; \hat{\theta}_g)$ of f_g, $g = 1, \ldots, G$, with no penalty. We then introduce a shrinkage coefficient w_g for each f_g and penalize on those shrinkage coefficients w's. As an example, for the Gaussian data, we consider the minimization of

$$\{Y - w_1 f_1(Z_{11}, \ldots, Z_{1d_1}; \hat{\theta}_1) - \cdots - w_G f_G(Z_{G1}, \ldots, Z_{Gd_G}; \hat{\theta}_G)\}^2$$

over $(w_1, \ldots, w_G)$ subject to the constraints that $\sum_{g=1}^{G} w_g < \zeta$ for some nonnegative shrinkage parameter ζ, and $w_g \geqslant 0$, $g = 1, \ldots, G$. A decreasing penalty parameter ζ would shrink some w_g's to be exactly zero, which in effect screening out those irrelevant pathways. Computationally, the minimization can be solved by calling readily available standard Lasso algorithm such as LARS (Efron *and others*, 2004) and is straightforward and fast.

We note that the entire solution path can be obtained for the above procedure. As such, the order of those pathways entering the model provides a natural way to rank the pathways. Moreover, an explicit pathway selection is possible, provided that an appropriate criterion is available to tune ζ. However, our numerical experiences have found that the criterion like Bayesian information criterion that is often recommended in the penalized estimation literature does not work satisfactorily in our context. For this reason, we propose here a heuristic pseudo pathway selection strategy. The idea is similar in spirit to, though not the same as, Wu *and others* (2007), who achieved variable selection in a conventional LM setup by introducing pseudo variables into the data. Our strategy calls to add a group of pseudo variables that are generated from a standard normal distribution independent of the existing pathways. We treat these pseudo variables as if they were from a single pathway and apply the nonlinear dimension reduction technique to obtain summary features of this pseudo pathway. We then amend the pseudo pathway to the original G pathways and conduct pathway ranking. We repeat this procedure for B times (say, $B = 100$). For each of those G pathways in the original data, we record the frequency of times that pathway shows up earlier than the known pseudo pathway on the solution path. We declare a pathway relevant to the response if its associated frequency is above a prespecified thresholding rate r (say, $r = 0.9$).

4. SIMULATIONS

4.1 *Simulation setup*

We generate totally G pathways, each of which contains p_G variables. Each variable follows a uniform distribution between 0 and 1, and variables within the pathway admit a compound symmetric correlation structure with correlation 0.2. The phenotype Y is related to 4 pathways, $P_1, \ldots, P_4$, through the function

$$Y = 9P_1 + 1.5\exp(3P_2) + 75(P_3 - 0.5)^2 + 3.75\sin\{2\pi(P_4 - 1)/3\} + \varepsilon,$$

where ε is a standard normal error independent of all pathways. The coefficients in front of each pathway function are to make their effects comparable in magnitude. The 4 relevant pathways are composed of individual variables as

$$\begin{aligned} P_1 &= X_{11} - X_{12}, \\ P_2 &= X_{21} + X_{22} + X_{23} - X_{21} \times X_{22} - X_{22} \times X_{23} - X_{21} \times X_{23}, \\ P_3 &= X_{31}, \\ P_4 &= X_{41} + X_{42} + X_{43} + X_{44} + X_{45}, \end{aligned}$$

where X_{gj} denotes the jth variable in the gth pathway. So the first, third, and fourth pathways depend on their individual variables in a linear fashion, whereas the second pathway involves both linear and interaction terms. Meanwhile, the first pathway affects the response through a linear function, whereas the other 3 through nonlinear functions. The sample size is set as $n = 100$. We consider both a relatively small number of pathways, $G = 10$, which is encountered in metabolomics studies, and a large number of pathways, $G = 50$, which is closer to gene pathway analysis. For each pathway, we first set a small number of variables, $p_G = 5$, which is mainly because some methods we are to compare can only work when $n > p$. Later, we will also consider the $p_G = 150$ case. In the real data analysis, we will further complement our numerical study with a scenario where the number of predictors in groups is large and some are larger than the sample size.

4.2 *Pathway ranking*

We compare our method of nonlinear dimension reduction followed by the regularized GLM or GAM, with GAR of Luan and Li (2008), and NPR of Wei and Li (2007). The comparison is in terms of pathway ranking for 2 reasons, first, ranking is of interest itself in genomics studies (Choi *and others*, 2009), and second, both GAR and NPR only produce a rank of pathways according to their relevance to the response, rather than pathway selection. In our setup, we repeat the data replications 100 times and report the number of times that each of those 4 truly relevant pathways is ranked among the top 4 pathways as our evaluation criterion. We also compare some variants at the steps of dimension reduction and post reduction model selection. Specifically, at the dimension reduction step, we compare PCA (Ma and Kosorok, 2009), PLS, the conventional SIR and SAVE, and SIR using a Gaussian kernel (KSIR). We choose the kernel version of SIR only due to the simplicity and the popularity of SIR in the dimension reduction literature. We employ the ridge regularized KSIR in (2.9), which is tuned by the GCV criterion given in (2.11). At the modeling step, we compare the GLM, which reduces to a LM with a normally distributed response in our setup, and the GAM when $G = 10$. We use LM only when $G = 50$ due to the limited sample size ($n = 100$).

Table 1 reports the results. Comparing 5 dimension reduction methods first under $G = 10$, we see that PCA followed by GAM works fine for pathways P_2, P_3, and P_4 but fails for pathway P_1. This is not surprising though, since by design the leading principal component direction should be close to $(1, 1, 1, 1, 1)^\top$, which is perpendicular to the actual direction $(1, -1, 0, 0, 0)^\top$. This simple example reveals a key drawback of PCA, which does dimension reduction in an unsupervised fashion that does not take into account the response information. In our example, the response is affected by the first pathway that is aligned along the direction $(1, -1, 0, 0, 0)^\top$, whereas PCA totally ignores that information. PLS does take into account the response information, but it still fails for P_2 and P_3. Similar behavior is observed for SIR. This is partly due to that pathway P_2 consists of both main effects and interaction terms of the individual variables, and for this reason, one linear combination cannot capture all regression information. For pathway P_3, although it only consists of a single variable, its effect in the response model is dominated by a quadratic trend to which PLS and SIR are insensitive. SAVE performs unsatisfactorily for this example, mainly due to the limited sample size. Finally, KSIR achieves the best performance across all pathways, thanks to its flexibility to accommodate various structures. Moreover, it is interesting to note that, after KSIR, GAM, and LM yield very similar results, indicating that KSIR has captured most

Table 1. *Pathway ranking and pathway selection for simulation studies. Under "pathway ranking" are numbers of times of the* 4 *truly relevant pathways being ranked as the top* 4 *most important ones among all pathways out of* 100 *data replications. Under "pathway selection" are numbers of times of each pathway being selected by the pseudo pathway selection strategy. Column* P_{rest} *reports the average selected times of all the rest of irrelevant pathways*

Method		Pathway ranking				Pathway selection				
Reduction	Model	P_1	P_2	P_3	P_4	P_1	P_2	P_3	P_4	P_{rest}
PCA	GAM	16	94	74	100	9	89	61	98	12
	LM	41	35	37	100	16	21	22	99	16
PLS	GAM	98	48	52	99	98	39	46	96	12
	LM	100	32	25	99	97	20	13	92	14
SIR	GAM	99	56	48	97	98	40	40	93	11
	LM	100	33	25	97	98	20	14	92	14
SAVE	GAM	51	89	99	44	33	82	99	33	11
	LM	65	37	42	66	52	23	20	48	17
KSIR	GAM	95	90	92	95	98	90	94	95	12
	LM	94	90	92	94	97	90	92	95	12
GAR (Luan and Li, 2008)		78	42	40	0	—	—	—	—	—
NPR (Wei and Li, 2007)		87	37	91	84	—	—	—	—	—

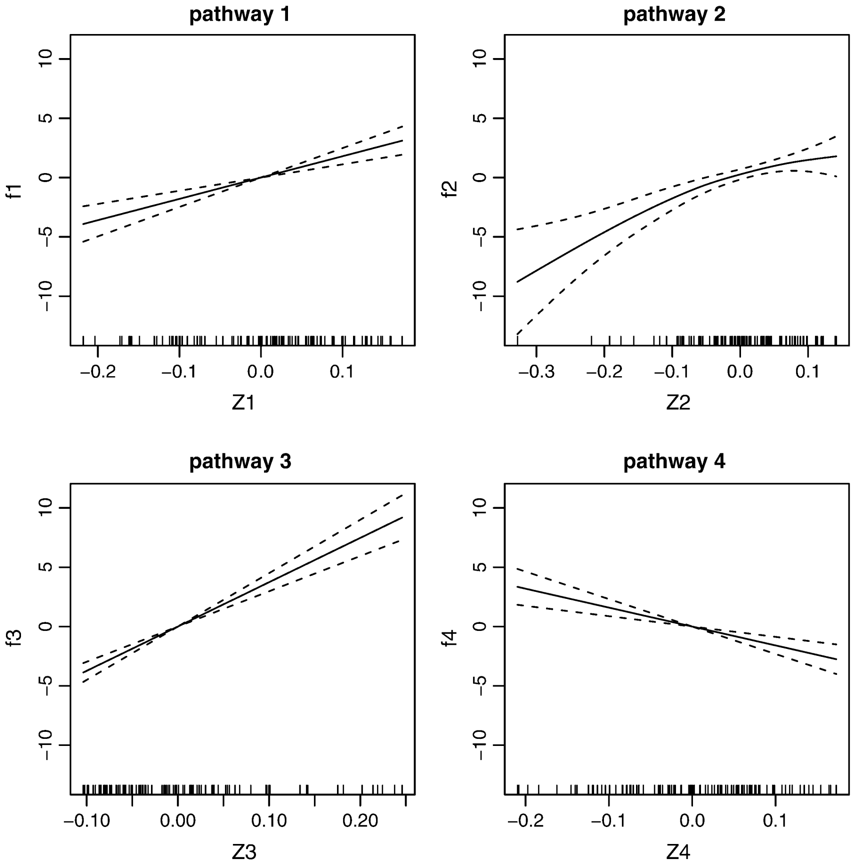

Fig. 1. Component smooth functions of the fitted GAM based on the KSIR summary features of the 4 relevant pathways. Upper and lower dashed lines are 2 standard errors above and below the estimate of the smooth function (solid line).

nonlinear effects in the model. This observation is further reinforced by Figure 1, which shows the GAM fit for each pathway based on the summary feature produced by KSIR. It is seen from the plot that, although some functions show a little nonlinear trend, overall the functions are all close to being linear, which partly explains why LM works about the same as GAM in this example.

Next, we compare our method with GAR of Luan and Li (2008) and NPR of Wei and Li (2007). It is clearly seen that GAR performs poorly for pathways P_2, P_3, and P_4. For P_2, this is because the base learner in GAR only models the main effects, whereas P_2 involves the interaction terms; and for P_3 and P_4, GAR is not designed to handle nonlinear effects, whereas P_3 and P_4 both have a dominating nonlinear association with the response. NPR is designed to tackle nonlinear effects, which explains why it works well with P_3 and P_4. On the other hand, its performance is less satisfactory for P_2, mainly due to the limited maximum depth of regression tree (which is set as 2) as its base learner. By slightly increasing the maximum depths, NPR is seen (results not shown here) to identify P_2 more frequently, but at the cost of decreasing frequency of finding the other 3 pathways.

We have also conducted simulations with $G = 50$ pathways, which exhibits a similar qualitative pattern as that for 10 pathways. Our general finding is that the sensitivity of all methods slightly decreases with an increasing number of total pathways. However, our method maintains the most competitive performance. For example, the number of times that $P_1, \ldots, P_4$ are ranked as the top 4 pathways among 100 replications are 12, 10, 10, and 96 for PCA, 16, 11, 12, and 0 for GAR, 40, 14, 36, and 46 for NPR, and 78, 77, 90, and 81 for KSIR. For brevity, the full tabular output of the results of 50 pathways is omitted.

We also briefly consider a case where the number of variables in pathway is larger than the sample size. Specifically, we generate 10 pathways, each of which consists of 150 variables. Only the first pathway P_1 is relevant to the response through the model $Y = \exp(P_1) + \varepsilon$, and $P_1 = X_1 + \cdots + X_6 - X_7 - \cdots - X_{10} + X_1^i + X_2^i - X_3^i - X_4^i$, where the last 4 terms are interactions that are randomly selected from all possible 2-way interactions of the first 10 predictors without replacement. For this small-n-large-p setup, the usual SIR, SAVE, and GAR are not applicable. The number of times that P_1 is ranked as the top pathway is recorded for 100 data replications. With the LM fitting, the results for PCA, PLS, and KSIR are 73, 0, and 85, respectively, whereas with the GAM fitting, the numbers are 47, 0, and 81. The number for NPR is 72. Again, our proposed method achieves the best performance in this setup. We also note that PLS fails in this example, partly due to the sparse and weak signals in P_1. As we increase the number of informative variables in P_1, which in effect increases the signal strength of P_1, PLS starts to identify the pathway. For example, with 50 informative variables in P_1, PLS finds this pathway 21 times out of 100. However, its performance is still inferior to other methods in this example. Overall, our proposed method achieves the best performance and is seen to be capable of handling both complicated variable structures within a pathway as well as nonlinear pathway effects on the response.

4.3 *Pathway selection*

In addition to pathway ranking, we next examine the performance of our method in terms of pathway selection using the pseudo pathway strategy proposed in Section 3.2. Specifically, to match our simulation example, we generate 5 pseudo variables for a pseudo pathway, obtain the leading sufficient predictor by a dimension reduction method, and then amend it with the original G pathways to produce the solution path. We repeat this procedure for $B = 100$ times and declare one of those G pathways being selected if its frequency of showing up earlier on the solution path than the pseudo pathway out of 100 times is above the prespecified threshold value $r = 0.9$. Table 1 reports the number of times that each of the truly relevant pathways being selected out of 100 data replications, plus the average number of all the rest of irrelevant pathways, with $G = 10$. The results for $G = 50$ are similar and thus are omitted. We first observe that the proposed method works pretty well in the sense that it achieves a high rate of identifying the truly relevant pathways, whereas maintaining a low rate of selecting those irrelevant pathways. In addition, it echoes our observations in Section 4.2 that KSIR outperforms all other dimension reduction solutions. Moreover, GAM and LM perform similarly if KSIR is employed in the reduction step.

5. A REAL DATA ANALYSIS

We analyze a microarray gene expression data of Horvath *and others* (2006) that studies glioblastoma. Glioblastoma is the most common primary malignant brain tumor and one of the most lethal one among all cancers. A data set of gene expressions of 120 patients were collected and normalized. Among those patients, only 9 were alive at the end of the study. To avoid complication of the response censoring, we focus on those $n = 111$ patients who died and use their (uncensored) time to death as the response variable. All the genes were mapped to the 33 regulatory pathways recorded in the KEGG database. Of the total 1668 nodes of the 33 pathways, 1498 were found in our data set. The list of those 33 pathways are given in Table 2 (column 1), along with the number of genes in that pathway from the KEGG database (column 2) and the number found in our data set (column 3). Note that for some pathways, the number

Table 2. *Pathway ranking and pathway selection for the analysis of the glioblastoma microarray data. The top 4 pathways are marked in bold face*

Pathway name	Number of genes		Selection		Ranking	
	Database	Data set	GAM	LM	GAM	LM
MAPK signaling pathway	269	249	**97**	**100**	**2**	**3**
Calcium signaling pathway	154	142	15	5	12	19
Cytokine–cytokine receptor interaction	142	126	**99**	**100**	**4**	**4**
Phosphatidylinositol signaling system	70	65	7	8	29	13
Neuroactive ligand–receptor interaction	133	116	**100**	**100**	**1**	**1**
Cell cycle	48	44	17	55	13	5
Ubiquitin mediated proteolysis	21	19	7	5	19	22
Apoptosis	75	72	44	0	14	24
Wnt signaling pathway	143	124	35	0	17	33
Transforming growth factor-beta signaling pathway	66	61	1	0	32	31
Axon guidance	111	100	5	5	25	17
Focal adhesion	121	115	50	7	5	14
Extracellular matrix-receptor interaction	53	50	6	2	24	25
Cell adhesion molecules	91	79	28	53	9	6
Adherens junction	70	68	10	5	15	23
Tight junction	105	91	12	1	10	27
Gap junction	92	88	3	0	30	29
Complement and coagulation cascades	53	51	**95**	**100**	**3**	**2**
Toll-like receptor signaling pathway	73	71	25	21	8	9
Jak-STAT signaling pathway	98	91	28	0	7	30
Natural killer cell mediated cytotoxicity	121	109	4	7	27	15
Circadian rhythm	7	6	10	22	16	8
Regulation of actin cytoskeleton	169	154	9	5	31	20
Insulin signaling pathway	134	129	5	9	23	12
Adipocytokine signaling pathway	64	60	1	5	33	21
Type II diabetes mellitus	42	40	8	0	18	32
Type I diabetes mellitus	5	5	4	1	28	28
Alzheimer’s disease	12	11	37	35	6	7
Prion diseases	9	7	12	16	11	10
Unknown	9	9	7	6	21	16
Unknown	22	21	19	5	20	18
Unknown	11	11	9	2	22	26
Unknown	10	8	4	14	26	11

of genes are larger than the sample size. The goal of this analysis is then to identify pathways that are strongly associated with the patient’s survival time from the brain cancer.

We analyze this data using the ridge regularized kernel SIR with a Gaussian kernel for dimension reduction, followed by fitting a GAM or an LM with L_1 regularization. We report the results for both pathway ranking, based on the order of appearance along the entire solution path, and pathway selection, based on the pseudo pathway strategy. Table 2 reports the number of times out of 100 pseudo pathway generations that each original pathway appears before the known pseudo one on the solution path (column 4 for the results based on GAM, and column 5 for LM). The table also reports the ranking of all the pathways (column 6 for GAM, and column 7 for LM). It is seen that 4 pathways stand out clearly: complement and coagulation cascades, mitogen-activated protein kinase (MAPK) signaling pathway, cytokine–cytokine receptor interaction, and neuroactive ligand–receptor interaction. The 4 pathways are ranked as top 4 by both methods, and the number of times they appear ahead of a pseudo pathway are all above 95. Besides, the results based on GAM and LM agree well.

There seem to exist recurring biological evidences to support our findings of those 4 important pathways. For the complement and coagulation cascades pathway, it was reported in Liu *and others* (2008) that the glioma cell invasiveness depends on proteases of the coagulation and complement cascades. In addition, suppression of the tissue factor–dependent coagulation cascade is found to be a contributing factor for the development of intratumoral hemorrhage ire glioblastoma (Takeshima *and others*, 2000). Coagulation cascade activation was also found to be related to glioma cell proliferation (Ogiichi *and others*, 2000). The MAPK signaling pathway is involved in various cellular functions, including cell proliferation, differentiation, and migration. It has been linked to being responsible for the malignant phenotype, including increased proliferation, defects in apoptosis, ability to induce neovascularization, and invasiveness (Liu *and others*, 2008), which are important prerequisites for the infiltrative and destructive growth patterns of malignant gliomas. Moreover, Mawrin *and others* (2003) and Pelloski *and others* (2006) reported the prognostic relevance of MAPK expression in glioblastoma multiforme. It is worth to mention that when Li and Li (2008) analyzed the same data set using a network-constrained regularization and variable selection method, they identified a well-connected subnetwork of genes, while genes from the MAPK signaling pathway constitute the majority of that subnetwork. The goal of Li and Li (2008) was not to select individual pathways, but to enhance gene selection by using the network information. But both studies point to the same pathway of interest, that is, the MAPK signaling pathway. Cytokine–cytokine receptor interactions and neuroactive ligand–receptor interactions are also highly likely to be associated with glioblastoma. Cytokines are important regulators and mobilizers of cells involved in cell growth, differentiation and death, angiogenesis, and development and repair processes aimed at the restoration of homeostasis. Cytokine–cytokine receptor interaction is the way Cytokines induces responses to an activating signal. Neuroactive ligand–receptor interactions also involve signaling molecules and interactions that play critical roles in a variety of important cellular processes, such as apoptosis, cytolysis, and cell proliferation.

6. Discussion

The focus of this article is to propose a 2-step statistical method for analyzing high-throughput biological data while utilizing pathway information. An important intermediate step is nonlinear dimension reduction, which turns high-dimensional data into a much lower dimensional and manageable summary features and permits flexible predictor interactions within the groups as well as flexible pathway effects on the phenotype response. Although there is no guarantee that nonlinear dimension reduction would always be superior than linear reduction, its flexible reduction form may facilitate both dimension reduction itself, as partly illustrated by example (2.3) and modeling after dimension reduction, as demonstrated by our simulation studies. Moreover, depending on the choice of kernels, the kernel-based nonlinear dimension reduction includes the usual linear reduction as a special case. The new method comes with a price of a more complicated but tractable implementation, and the computer code (in R) is available upon request.

There are a number of possible avenues for future extensions. First, we have been focusing on ranking and selecting the entire pathway as a whole, while it could also be of interest to select individual genes within the pathway that are relevant to the phenotype. Both Wei and Li (2007) and Luan and Li (2008) can target individual genes. To accomplish this task, it requires an effective variable selection technique in the kernel setting, which, to our knowledge, has only limited success (Zhang, 2006). Second, the pathway information is now only used to divide the genes into groups, while how to incorporate further pathway information, for example, the known network structure of a given pathway, is important. Finally, kernel selection remains a critical yet open question for all kernel-based methods. All these problems are warranted for future research.

Supplementary material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

Acknowledgment

The authors thank the editor, an associate editor, and 2 referees for their constructive comments, which have helped greatly improve the paper. *Conflict of Interest:* None declared.

Funding

National Science Foundation (DMS-0706919) to L.L.

References

Bickel, P., Brown, J., Huang, H. and Li, Q. (2009). An overview of recent developments in genomics and the statistical methods that bear on them. *Technical Report*. Berkeley, CA: University of California.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* **37**, 373–384.

Chatterjee, N., Kalaylioglu, Z., Moslehi, R., Peters, U. and Wacholder, S. (2006). Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. *American Journal of Human Genetics* **79**, 1002–1016.

Choi, N. H., Shedden, K., Sun, Y. and Zhu, J. (2009). Penalized regression methods for ranking multiple genes by their strength of unique association with a quantitative trait. *Technical Report*. Ann Arbor, MI: University of Michigan.

Cook, R. D. (1996). Graphics for regressions with a binary response. *Journal of the American Statistical Association* **91**, 983–992.

Cook, R. D., Li, B. and Chiaromonte, F. (2007). Dimension reduction in regression without matrix inversion. *Biometrika* **94**, 569–584.

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

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics* **32**, 407–451.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. London, UK: Chapman and Hall.

Horvath, S., Zhang, B., Carlson, M., Lu, K.V., Zhu, S., Felciano, R.M., Laurance, M.F., Zhao, W., Qi, S., Chen, Z. *and others* (2006). Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a novel molecular target. *Proceedings of National Academy of Sciences of the United States of America* **103**, 17402–17407.

Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28**, 27–30.

Karp, P. D., Ouzounis, C. A., Moore-Kochlacs, C., Goldovsky, L., Kaipa, P., Ahren, D., Tsoka, S., Darzentas, N.,Kunin, V. and Lopez-Bigas, N. (2005). Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Research* **19**, 6083–6089.

Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association* **102**, 997–1008.

LI, C. AND LI, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics* **24**, 1175–1182.

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

LI, L. (2009). Exploiting predictor domain information in sufficient dimension reduction. *Computational Statistics and Data Analysis* **53**, 2665–2672.

LI, L., COOK, R. D. AND TSAI, C. L. (2007). Partial inverse regression. *Biometrika* **94**, 615–625.

LI, L. AND YIN, X. (2008). Sliced inverse regression with regularizations. *Biometrics* **64**, 124–131.

LIU, Z., GARTENHAUS, R. B., TAO, M. AND JIANG, F. (2008). Gene and pathway identification with Lp penalized Bayesian logistic regression. *BMC Bioinformatics* **9**, 412.

LUAN, Y. AND LI, H. (2008). Group additive regression models for analysis of genomic data. *Biostatistics* **9**, 100–113.

MA, S. AND KOSOROK, M. R. (2009). Identification of differential gene pathways with principal component analysis. *Bioinformatics* **25**, 882–889.

MATTHEWS, L., GOPINATH, G., GILLESPIE, M. *and others* (2008). Reactome knowledgebase of biological pathways and processes. *Nucleic Acids Research* **37**, 619–622.

MAWRIN, C., DIETE, S., TREUHEIT, T., KROPF, S., VORWERK, C. K., BOLTZE, C., KIRCHES, E., FIRSCHING, R. AND DIETZMANN, K. (2003). Prognostic relevance of MAPK expression in glioblastoma multiforme. *International Journal of Oncology* **33**, 641–648.

MCCULLAGH, P. AND NELDER, J. A. (1989). *Generalized Linear Models*, 2nd edition. London: Chapman and Hall.

OGIICHI, T., HIRASHIMA, Y., NAKAMURA, S., ENDO, S., KURIMOTO, M. AND TAKAKU, A. (2000). Tissue factor and cancer procoagulant expressed by glioma cells participate in their thrombin-mediated proliferation. *Journal of Neuro-Oncology* **46**, 1–9.

PANG, H. AND ZHAO, H. (2008). Building pathway clusters from random forests classification using class votes. *BMC Bioinformatics* **9**, 87.

PELLOSKI, C. E., LIN, E., ZHANG, L., YUNG, W. K. A., COLMAN, H., LIU, J., WOO, S. Y., HEIMBERGER, A. B., SUKI, D., PRADOS, M. *and others* (2006). Prognostic associations of activated mitogen-activated protein kinase and akt pathways in glioblastoma. *Clinical Cancer Research* **12**, 3935–3941.

SHI, M. AND MA, S. (2008). Identifying subset of genes that have influential impacts on cancer progression: a new approach to analyze cancer microarray data. *Functional Integrative Genomics* **8**, 361–373.

SUBRAMANIAN, A, TAMAYO, P, MOOTHA, V. K., MUKHERJEE, S., EBERT, B. L., GILLETTE, M. A., PAULOVICH, A., POMEROY, S. L., GOLUB, T. R., LANDER, E. S. *and others* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15545–15550.

TAKESHIMA, H., NISHI, T., KURATSU, J., KAMIKUBO, Y., KOCHI, M. AND USHIO, Y. (2000). Suppression of the tissue factor-dependent coagulation cascade: a contributing factor for the development of intratumoral hemorrhage in glioblastoma. *Internaltional Journal of Molecular Medicine* **6**, 271–276.

TIAN, L, GREENBERG, S. A., KONG, S. W., ALTSCHULER, J., KOHANE, I. S. AND PARK, P. J. (2005). Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 13544–13549.

WEI, Z. AND LI, H. (2007). Nonparametric pathways-based regression models for analysis of genomic data. *Biostatistics* **8**, 265–284.

WU, H. M. (2008). Kernel sliced inverse regression with applications on classification. *Journal of Computational and Graphical Statistics* **17**, 590–610.

WU, Q., LIANG, F. AND MUKHERJEE, S. (2008). Regularized sliced inverse regression for kernel models. *Technical Report*. Durham, NC: Duke University.

WU, Y., BOOS, D. D. AND STEFANSKI, L. A. (2007). Controlling variable selection by the addition of pseudovariables. *Journal of the American Statistical Association* **477**, 235–243.

YUAN, M. AND LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B* **68**, 49–67.

YUAN, M. AND LIN, Y. (2007). On the non-negative garrotte estimator. *Journal of the Royal Statistical Society, Series B* **69**, 143–161.

ZHANG, H. H. (2006). Variable selection for support vector machines via smoothing spline ANOVA. *Statistica Sinica* **16**, 659–674.

ZHU, L. P., WANG, T., ZHU, L. X. AND FERRÉ, L. (2010). Sufficient dimension reduction through discretization-expectation estimation. *Biometrika* **97**, 295–304.

[*Received April 13, 2010; revised December 12, 2010; accepted for publication December 12, 2010*]