VARIABLE SELECTION AND REGRESSION ANALYSIS FOR GRAPH-STRUCTURED COVARIATES WITH AN APPLICATION TO GENOMICS

Caiyan Li[*] and **Hongzhe Li**[*]
University of Pennsylvania School of Medicine

Abstract

Graphs and networks are common ways of depicting information. In biology, many different biological processes are represented by graphs, such as regulatory networks, metabolic pathways and protein-protein interaction networks. This kind of a priori use of graphs is a useful supplement to the standard numerical data such as microarray gene expression data. In this paper, we consider the problem of regression analysis and variable selection when the covariates are linked on a graph. We study a graph-constrained regularization procedure and its theoretical properties for regression analysis to take into account the neighborhood information of the variables measured on a graph, where a smoothness penalty on the coefficients is defined as a quadratic form of the Laplacian matrix associated with the graph. We establish estimation and model selection consistency results and provide estimation bounds for both fixed and diverging numbers of parameters in regression models. We demonstrate by simulations and a real dataset that the proposed procedure can lead to better variable selection and prediction than existing methods that ignore the graph information associated with the covariates.

Keywords and phrases

Regularization; Sign consistency; Network; Laplacian Matrix; High dimensional data

1. Introduction

There has been a growing interest in penalized least squares problems via L_1 or other types of regularization, especially in high-dimensional settings. Important penalty functions that can lead to sparse variable selection in regression include Lasso (Tibshirani, 1996) and SCAD (Fan and Li, 2001). In particular, Lasso has the crucial advantage of being a convex problem, which leads to efficient computational algorithms by coordinate descent (Efron *et al.*, 2004; Friedman *et al.*, 2007; Wu and Lange, 2008) and sparse solutions. Zou (2006) proposed a novel adaptive Lasso procedure and presented results on model selection consistency and oracle properties of the parameter estimates. Zhao and Yu (2006) presented the irrepresentable condition for model selection consistency of Lasso. Zhang and Huang (2006) studied the sparsity and bias of the Lasso selection in high-dimensional linear regression. Fan and Li (2001) and Huang and Xie (2007) established the asymptotic oracle properties of the SCAD-penalized least squares estimators when the number of covariates is fixed or increases with the sample sizes. These novel penalized estimation methods are quite

effective in selecting relevant variables and in predicting future outcomes, especially in high-dimensional settings.

New estimation procedures have also been developed in recent years to account for certain structures of the explanatory variables. These include the group Lasso procedure (Yuan and Lin, 2006) when the explanatory variables are grouped or organized in a hierarchical manner, the elastic net (Enet) procedure (Zou and Hastie, 2005) that deals with groups of highly correlated variables, and the fused Lasso (Tibshirani *et al.*, 2005) that imposes the L_1 penalty on the absolute differences of the regression coefficients in order to account for some smoothness of the coefficients. Nardi and Rinaldo (2008) established the asymptotic properties of the group Lasso estimator for linear models. Jia and Yu (2008) provided conditions for model selection consistency of the elastic net when $p \gg n$. Zou and Zhang (2009) proposed an adaptive elastic net with a diverging number of parameters and established its oracle property. Among these procedures, the Enet regularization and the fused Lasso are particularly appropriate for the analysis of genomic data, where the former encourages a grouping effect and the latter often leads to smoothness of the coefficient profiles for ordered covariates.

Motivated by a genomic application in order to account for network information in the analysis of genomic data, Li and Li (2008) proposed a network-constrained regularization procedure for fitting linear regression models and for variable selection, where the predictors in the regression model are genomic data that are measured on the genetic networks, which we call the graph-structured covariates. In particular, we assume that the covariates in the regression model are values of the nodes on a graph, where a link between two nodes may indicate a functional relationship between two genes in genetic network or physical neighbor between two voxels on brain images. Since many biological networks are constructed using data from high-throughput experiments, often there is a probability associated with an edge to indicate the certainty of a link. Such an edge probability can be used as a weight in undirected graph, in which case we have a weighted graph. This graph-constrained regularization procedure is similar in spirit to the fused Lasso (Tibshirani *et al.*, 2005), both of which try to smooth the regression coefficients in certain ways. However, the fused Lasso does not utilize prior graph information. Second, instead of using the L_2 norm on the differences of the coefficients of the linked variables, the fused Lasso uses the L_1 norm on the differences, which tends to lead to the same regression coefficients for linked variables. In this paper, we consider the general problem of regression analysis when the explanatory variables are nodes on a graph and present a cyclical coordinate descent algorithm (Friedman *et al.*, 2008) to implement the network-constrained regularization procedure of Li and Li (2008). This algorithm provides new insight on how neighboring variables affect the coefficient estimate of a node. We also extend the procedure of Li and Li (2008) to account for the possibility of different signs of the regression coefficients for neighboring variables. In addition, we provide theoretical results of the estimates, including sign consistency and error bounds of the estimator and L_2 consistency.

This paper is organized as follows. In Section 2, we describe the problem of regression analysis with covariates measured on graphs. We then present a graph-constrained estimation (Grace) procedure in order to account for the graph structures in Section 2.1 and an efficient coordinate descent algorithm to implement the proposed regularization methods in Section 2.3. We present the estimation and model selection consistency results in Section 3. We provide Monte Carlo simulation results in Section 4 and results from the application to the analysis of a dataset on the gene expression of brain aging in Section 5. Finally, we give a brief discussion of the methods and results.

2. Regression Analysis for Covariates Measured on a Graph

Consider a weighted graph $G = (V, E, W)$, where $V = \{1, \cdots, p\}$ is the set of vertices that correspond to the p predictors, $E = \{u \sim v\}$ is the set of edges indicating that the predictors u and v are linked on the graph and there is an edge between u and v, and W is the set of weights of the edges, where $w(u, v)$ denotes the weight of edge $e = (u \sim v)$. In genomic studies, biological networks are often represented as graphs, an edge between u and v on the graph can indicate some functional relationship between them and the weight can be used to measure the uncertainty of the edge between two vertices, for example, indicating the probability of having an edge between two variables when the graph is constructed from data. For each given sample, we assume that we have numerical measurements on each of the vertices and these measurements are treated as explanatory variables in a regression analysis framework. For the uth node, let x_{iu} be the numerical measurement of the uth vertex on the graph for the ith individual. Further, let $\mathrm{x}_u = (x_{1u}, \cdots, x_{nu})^T$ be the measured values at the uth vertex for n i.i.d. samples. Consider the problem of variable selection and estimation where we have design matrix $\mathrm{X} = (\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_p) \in \mathcal{R}^{n \times p}$ and response vector $\mathrm{y} = (y_1, y_2, \cdots, y_n)^T \in \mathcal{R}^n$, and they follow a linear model

$$\mathrm{y} = \mathrm{X}\beta + \varepsilon, \tag{2.1}$$

where $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T \sim N(0, \sigma^2 I_n)$ and $\beta = (\beta_1, \ldots, \beta_p)^T$. Throughout this paper we assume that the predictors and the response are centered so that

$$\sum_{i=1}^{n} y_i = 0, \ \sum_{i=1}^{n} x_{ij} = 0, \quad \text{and} \ \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1 \quad \text{for } j = 1, \ldots, p.$$

In this paper, we consider that the design matrix X is a deterministic matrix in the fixed design settings.

When p is large, we assume that model (2.1) is "sparse", i.e., most of the true regression coefficients β are exactly zero. Without loss of generality we assume the first q elements of vector β are non-zeroes. Denote $\beta_{(1)} = (\beta_1, \ldots, \beta_q)^T$ and $\beta_{(2)} = (\beta_{q+1}, \ldots, \beta_p)^T$, then element-wise $\beta_{(1)} \neq 0$ and $\beta_{(2)} = 0$. Now write $\mathrm{X}_{(1)}$ and $\mathrm{X}_{(2)}$ as the first q and last $p - q$ columns of X, respectively, and let $\mathrm{C} = \frac{1}{n}\mathrm{X}^T\mathrm{X}$, which can then be expressed in the following block-wise form:

$$\mathrm{C} = \begin{pmatrix} \mathrm{C}_{11} & \mathrm{C}_{12} \\ \mathrm{C}_{21} & \mathrm{C}_{22} \end{pmatrix}.$$

The goal of this paper is to develop a regularization procedure for selecting the true relevant variables. Different from the existing approaches, we particularly account for the fact that the explanatory variables are related on a graph. We make this more precise in the next Section.

2.1. Graph-constrained regularization and variable Selection

In order to account for the fact that the p explanatory variables are measured on a graph, we first introduce the Laplacian matrix (Chung, 1997) associated with a graph. Let the degree of the vertex v be $d_v = \Sigma_{u \sim v} \, w(u, v)$. We say u is an isolated vertex if $d_u = 0$. Following Chung (1997), we define the Laplacian matrix L for graph G with the uvth element defined by

$$L(u,v) = \begin{cases} 1 - w(u,u)/d_u & \text{if } u = v \text{ and } d_u \neq 0 \\ -w(u,v)/\sqrt{d_u d_v} & \text{if } u \text{ and } v \text{ are adjacent} \\ 0 & \text{otherwise.} \end{cases} \qquad (2.2)$$

It is easy to verify that this matrix is positive semi-definite with 0 as the smallest eigenvalue and 2 as the largest eigenvalue when all the weights are 1 (Chung, 1997). To allow the matrix to change with n, we further express this matrix in block-wise form,

$$L = \begin{pmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{pmatrix},$$

where L_{11} corresponds to the q nodes that are relevant to the response and L_{22} corresponds to the $p - q$ nodes that are not relevant.

The Laplacian matrix has the following interpretations. For a given vector β, the edge derivative of β along the edge $e(u, v)$ at u is defined as

$$\left. \frac{\partial \beta}{\partial e} \right|_u = \sqrt{w(u,v)} \left(\frac{\beta_u}{\sqrt{d_u}} - \frac{\beta_v}{\sqrt{d_v}} \right),$$

and therefore the local variation of β at u can be measured by

$$\sqrt{\sum \left(\left. \frac{\partial \beta}{\partial e} \right|_u \right)^2}.$$

The smoothness of vector β with respect to the graph structure can be expressed as

$$\beta^T L \beta = \sum_{u \sim v} \left(\frac{\beta_u}{\sqrt{d_u}} - \frac{\beta_v}{\sqrt{d_v}} \right)^2 w(u,v).$$

This variation functional for vectors β penalizes vectors that differ too muchmover nodes that are linked. It contains a scaling by $\sqrt{d_u}$. One intuitive reason for such a scaling is to allow a small number of nodes with large d_u to have more extreme values of β_u while the usually much greater number of nodes with small d_u should not ordinarily allow to have very large β_u. This variation functional has been widely used in semi-supervised learning on graphs (Zhu, 2005; Zhou *et al.*, 2004).

For many problems with covariates measured on a graph, we would expect that the neighboring variables are correlated and therefore the regression coefficients would show some smoothness. One way to account for such a dependence of the regression coefficients is to impose a Markov random field (MRF) prior to the collection of β vectors. The MRF decomposes the joint prior distribution of the β_u's into lower-dimensional distributions based on the graph-neighborhood structures. A common MRF model is the Gaussian MRF model that assumes that the joint distribution of β is given by

$$f(\beta) \propto \exp\left\{ -\frac{1}{2\sigma^2}\beta^T L\beta \right\},$$

which is an improper density. Based on this Gaussian MRF prior assumption, Li and Li (2008) introduced the following graph-constrained estimation of the regression coefficients, denoted by $\hat{\beta}$,

$$\widehat{\beta} = \operatorname{argmin}_\beta Q(\beta, \lambda_1, \lambda_2) \tag{2.3}$$

where

$$
\begin{aligned}
Q(\beta, \lambda_1, \lambda_2) &= \left\| y - X\beta \right\|_2^2 + \lambda_1 \left\| \beta \right\|_1 + \lambda_2 \beta^T L\beta, \\
&= (y - X\beta)^T (y - X\beta) + \lambda_1 \sum_u |\beta_u| + \lambda_2 \sum_{u \sim v} \left(\frac{\beta_u}{\sqrt{d_u}} - \frac{\beta_v}{\sqrt{d_v}} \right)^2 w(u, v),
\end{aligned}
$$

where L is the Laplacian as defined in (2.2) and the tuning parameters λ_1, λ_2 control the amount of regularization for sparsity and smoothness. For the special case when $\lambda_2 = 0$, the estimate reduces to the Lasso, and when L is the identity matrix, the estimate reduces to the elastic net estimates.

2.2. An adaptive graph-constrained regularization

The Grace procedure may not perform well when two variables that are linked on the graph have different signs in their regression coefficients, in which case the coefficients are not expected to be locally smooth. For example, for genetic networks, two genes might be negatively correlated with the phenotypes and are therefore expected to have different signs in their regression coefficients. In order to account for the sign differences, we can first perform a standard least square regression when $p < n$ or the elastic net regression when p n and denote the estimate as $\tilde{\beta}$. We can then modify the above objective function as

$$
\begin{aligned}
Q^*(\lambda_1, \lambda_2, \beta) &= \left\| y - X\beta \right\|_2^2 + \lambda_1 \left\| \beta \right\|_1 + \lambda_2 \beta^T L^*\beta, \\
&= \left\| y - X\beta \right\|_2^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{u \sim v} \left(\frac{\operatorname{sign}(\tilde{\beta}_u)\beta_u}{\sqrt{d_u}} - \frac{\operatorname{sign}(\tilde{\beta}_v)\beta_v}{\sqrt{d_v}} \right)^2 w(u, v),
\end{aligned}
$$

where

$$
L^*(u, v) = \begin{cases}
1 - w(u, u)/d_u & \text{if } u = v \text{ and } d_u \neq 0 \\
-\operatorname{sign}(\tilde{\beta}_u)\operatorname{sign}(\tilde{\beta}_v)w(u, v)/\sqrt{d_u d_v} & \text{if } u \text{ and } v \text{ are adjacent} \\
0 & \text{otherwise.}
\end{cases}
$$

Note that the L* matrix is still positive semi-definite. We call the β defined by

$$\widehat{\beta} = \operatorname{argmin}_\beta Q^*(\beta, \lambda_1, \lambda_2) \tag{2.4}$$

the adaptive Grace (aGrace).

2.3. A Coordinate Descent Algorithm

Friedman *et al.* (2008) presented a coordinate descent algorithm for solving the Lasso and the Enet regularization. In this section, we develop a similar algorithm for the proposed graph-constrained regularization. We only present the detailed algorithm for the optimization problem defined by equation (2.3). Similar algorithms can be developed by the aGrace defined by (2.4). If we let $\lambda = (\lambda_1 + 2\lambda_2)/2n$ and $\alpha = \lambda_1/(\lambda_1 + 2\lambda_2)$, the Grace can be written as

$$\widehat{\beta}(\lambda, \alpha) = \operatorname*{argmin}_{\beta} \left\{ R(\beta) := \frac{1}{2n} \left\| y - X\beta \right\|_2^2 + \lambda P_\alpha(\beta) \right\},$$ (2.5)

where

$$P_\alpha(\beta) := (1-\alpha)\frac{1}{2}\beta^T L\beta + \alpha \left\| \beta \right\|_1 = (1-\alpha)\frac{1}{2}\sum_{u \sim v} \left(\frac{\beta_u}{\sqrt{d_u}} - \frac{\beta_v}{\sqrt{d_v}} \right)^2 + \alpha \sum_{u=1}^{p} |\beta_u|$$

is the graph-constrained penalty function.

Given a covariate x_u, suppose we have estimated $\tilde{\beta}_v$ for $v \neq u$, and we want to partially minimize the objective function with respect to β_u. We can rewrite the objective function in (2.5) as

$$R(\beta) = \frac{1}{2n}\sum_{i=1}(y_i - \sum_{v \neq u}x_{iv}\tilde{\beta}_v - x_{iu}\beta_u)^2 + \lambda(1-\alpha)\frac{1}{2}\sum_{v \sim u}\left(\frac{\beta_u}{\sqrt{d_u}} - \frac{\tilde{\beta}_v}{\sqrt{d_v}}\right)^2 + \lambda\alpha|\beta_u| + \lambda(1-\alpha)\frac{1}{2}\sum_{\substack{w \sim v \\ w, v \neq u}}\left(\frac{\tilde{\beta}_w}{\sqrt{d_w}} - \frac{\tilde{\beta}_v}{\sqrt{d_v}}\right)^2 + \lambda\alpha\sum_{w \neq u}|\tilde{\beta}_w|.$$

We would like to compute the gradient at β_u, which only exists when $\beta_u \neq 0$. We first consider the case that the covariate u is connected to some other nodes (variables) on the network. If $\beta_u > 0$, due to the standardization of the covariates, we can differentiate the objective function $R(\beta)$ with respective to β_u as

$$\frac{\partial R}{\partial \beta_u} = -\left[\frac{1}{n}\sum_{i=1}x_{iu}(y_i - \sum_{v \neq u}x_{iv}\tilde{\beta}_v) + \lambda(1-\alpha)\sum_{v \sim u}\frac{\tilde{\beta}_v}{\sqrt{d_u d_v}}\right] + \lambda\alpha + [1 + \lambda(1-\alpha)]\beta_u.$$

Similarly, we can get the corresponding expression when $\beta_u < 0$. Following the calculus by Donoho and Johnstone (1994) and Friedman *et al.* (2007), we obtain the coordinate-wise update form for β_u as

$$\tilde{\beta}_u \leftarrow \frac{S\left(\frac{1}{n}\sum_{i=1}x_{iu}(y_i - \tilde{y}_i^{(u)}) + \lambda(1-\alpha)\sum_{v \sim u}\frac{\tilde{\beta}_v}{\sqrt{d_u d_v}}, \lambda\alpha\right)}{1 + \lambda(1-\alpha)},$$ (2.6)

where

- $\tilde{y}_i^{(u)} = \sum_{v \neq u} x_{iv} \tilde{\beta}_v$ is the partial residual for fitting β_u, i.e., the fitted value excluding the contribution from x_{iu}. Since the covariates are standardized, $\frac{1}{n}\sum_{i=1} x_{iu}(y_i - \sum_{v \neq u} x_{iv} \tilde{\beta}_v)$ is the simple least-squares coefficient while fitting the partial residual to x_{iu}, $i = 1, \cdots, n$.

- $S(z, \gamma)$ is the soft-thresholding operator with value

$$\text{sign}(z)(|z| - \gamma)_+ = \begin{cases} z - \gamma & \text{if } z > 0 \text{ and } \gamma < |z| \\ z + \gamma & \text{if } z < 0 \text{ and } \gamma < |z| \\ 0 & \text{otherwise.} \end{cases}$$

When covariate u is not connected to other nodes on the network, i.e., when it has no neighbors, the corresponding coordinate-wise updating formula becomes the Lasso updating formula, that is

$$\tilde{\beta}_u \leftarrow S\left(\frac{1}{n}\sum_{i=1} x_{iu}(y_i - \tilde{y}_i^{(u)}), \lambda\alpha \right). \tag{2.7}$$

Comparing the two updated forms of (2.6) and (2.7), an intuitive explanation can be drawn to help understand the effect of the graph-constraint penalty on the coefficients. For an isolated predictor, the graph penalty is vanished and thus we only apply a soft-thresholding for the Lasso penalty. While for a connected predictor, form (2.6) takes into account the graph-constraint to the penalty by adding the scaled summation of the coefficients of the neighboring covariates to the simple least-squares coefficient and applying a proportional shrinkage for the graph penalty.

Given α, we can compute the solution path for a decreasing sequence of values for λ, starting from the smallest value λ_{max} for which there is no covariate selected, i.e., $\hat{\beta} = 0$. Similarly to Friedman *et al.* (2007), we can set $\lambda_{max} = \max_l | < x_l, y > |/n\alpha$, $\lambda_{min} = \varepsilon\lambda_{max}$ and construct a sequence of K values of λ decreasing from λ_{max} to λ_{min} on the log scale. Typical values are $\varepsilon = 0.001$ and $K = 100$. Cross-validation (CV) can be used to select the two tuning parameters α and λ.

3. Error Bound and Model Selection Consistency for Fixed and Diverging *p*

In this section, we provide some theoretical results on the proposed Grace procedure, including the error bounds, L_2 consistency of Grace and the model selection consistency for both fixed and diverging p when p tends to infinity with the sample size n. Our theoretical development follows that of Zhao and Yu (2006), Jia and Yu (2008) and Zou and Zhang (2009) on sign consistency of Lasso and adaptive elastic net estimates. In our theoretical analysis, we assume the following regularity conditions throughout:

A1 We use $\Lambda_{min}(M)$ and $\Lambda_{max}(M)$ to denote the minimum and maximum eigenvalues of a positive definite matrix M, respectively. We further assume that $C = \frac{1}{n}X^T X$ is positive definite and

$$b \leq \Lambda_{min}(C) \leq \Lambda_{max}(C) \leq B$$

where b and B are two positive constants that do not depend on n.

A2 $\frac{1}{n} max_{1 \leq i \leq n} \sum_{j=1}^{p} x_{ij}^2 \to 0$, as $n \to \infty$.

These two conditions assume that the predictor matrix has a reasonably good behavior and were also assumed in Zhao and Yu (2006) and in Zou and Zhang (2009). Condition (A1) is also the condition (F) in Fan and Peng (2004) and condition (A2) ensures that the rows of the matrix X behave like a sample from a probability distribution in $\mathcal{R}^p$ (Portnoy, 1984). These two conditions hold naturally if one assumes that x_i are *i.i.d* with finite second moments.

3.1. Error bound and L_2-consistency of Grace

We first provide the following non-asymptotic risk bound for the Grace of the regression coefficients defined by (2.1) for any p and n:

Theorem 3.1—Given the data (y, X), define the Grace as

$$\widehat{\beta}(\lambda_1, \lambda_2) = \operatorname{argmin}_\beta \{ \|y - X\beta\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \beta^T L\beta \},$$

for non-negative tuning parameters λ_1 and λ_2. Then under the regularity condition (A1), we have

$$E(\|\widehat{\beta}(\lambda_1, \lambda_2) - \beta\|_2^2) \leq \frac{4\lambda_2^2 \Lambda_{\max}^2(L) \|\beta_{(1)}\|_2^2 + 4pnB\sigma^2 + 2\lambda_1^2 p}{n^2 \Lambda_{\min}^2(C + \frac{\lambda_2}{n}L)}. \tag{3.1}$$

The proof of this theorem is given in the Supplemental Materials. Note that this result is not asymptotic and holds for any p and $q < p$. From this theorem, under the regularity assumption (A1) and the following further assumptions on p and the tuning parameters λ_1, λ_2,

A3 $\lim_{n \to \infty} \frac{p}{n} = 0$,

A4 $\lim_{n \to \infty} \frac{\lambda_1 \sqrt{p}}{n} = 0$,

A5

$\lim_{n \to \infty} \frac{\lambda_2}{n} = 0$ and $\lim_{n \to \infty} \frac{\lambda_2 \|\beta_{(1)}\|_2}{n} = 0$,

we have

$$\|\widehat{\beta}(\lambda_1, \lambda_2) - \beta\|_2^2 \xrightarrow{P} 0,$$

which implies that the Grace of β is L_2 consistent. This result implies that the Grace procedure chooses the important variables with high probability and that falsely chosen variables by Grace have very small coefficients. The L_2 consistency result suggests that we may use some hard-thresholding procedure to further eliminate the variables with very small Grace coefficients. Alternatively, an interesting randomized selection procedure proposed by Bickel *et al.* (2008) can be used to further eliminate the variables with small estimated Grace coefficients. Note that under the classical setting where p, q and β_i are all fixed as $n \to \infty$, the assumptions A3–A5 hold when $\lambda_i/n \to \infty$, $i = 1, 2$.

3.2. Model selection consistency when *p* is fixed

We next establish the results on model selection consistency for the standard case where p and q are fixed when $n \to \infty$. Following Zhao and Yu (2006), we define the Grace of β to be sign consistent if there exists λ_1 and λ_2 as functions of n such that

$$\lim_{n \to \infty} \Pr(\text{sign}(\widehat{\beta}(\lambda_1, \lambda_2)) = \text{sign}(\beta)) = 1.$$

To establish the sign consistency of the Grace, we first provide the following graph-constrained irrepresentable condition (GC-IC): there exists $\eta > 0$ and $\lambda_1 > 0$, $\lambda_2 > 0$, such that

$$\left| \left(C_{21} + \tfrac{\lambda_2}{n} L_{21}\right) \left(C_{11} + \tfrac{\lambda_2}{n} L_{11}\right)^{-1} \left[\text{sign}(\beta_{(1)}) + \tfrac{2\lambda_2}{\lambda_1} L_{11} \beta_{(1)} \right] - \tfrac{2\lambda_2}{\lambda_1} L_{21} \beta_{(1)} \right| \leq \mathbf{1} - \eta, \tag{3.2}$$

where $\mathbf{1}$ is a vector of 1's with length $p - q$ and the inequality holds element-wise. Further, we assume that $C \to C_0$, where C_0 is positive definite. The GC-IC is a consequence of the Karush-Kuhn-Tucker (KKT) conditions for the following constrained optimization problem that corresponds to the penalized optimization problem of equation (2.5),

$$\widehat{\beta}(\lambda, \alpha) = \text{argmin}_\beta \left\{ \frac{1}{2n} \left\| y - X\beta \right\|_2^2 : P_\alpha(\beta) \leq \lambda \right\}.$$

Theorem 3.2—For fixed p, q and β, if $C \to C_0$, where C_0 is positive definite and condition (A.2) holds, the graph-constrained estimate is sign consistent if and only if GC-IC (3.2) holds for λ_1, λ_2 that satisfy $\lambda_1 / \sqrt{n} \to \infty$ and $\lambda_i / n \to 0$, for $i = 1, 2$.

This theorem is a special case of Theorem 3.3 and its proof is similar to that of Zhao and Yu (2006) for Lasso estimates. We therefore omit its proof in this paper. Note that the required conditions on the sparsity tuning parameter λ_1 are the same as those for the Lasso (Zhao and Yu, 2006), for example, $\lambda_1 = \sqrt{n} \log n$ is a suitable choice. This theorem indicates that under some restrictive conditions of the design matrix and the Laplacian matrix of the network, the sign consistency property holds for the graph-constrained regularization. To gain further insight into GC-IC, consider the special cases when λ_2 is preselected and fixed and when λ_1 goes to infinity, the GC-IC reverses back to the irrepresentable condition for the Lasso given in Zhao and Yu (2006) and the graph-constrained penalty function

$\lambda_1 \left\| \beta \right\|_1 + \lambda_2 \beta^T L \beta = \lambda_1 (\left\| \beta \right\|_1 + \tfrac{\lambda_2}{\lambda_1} \beta^T L \beta)$ is reduced to the Lasso penalty.

3.3. Model selection consistency when *p* diverges

We now consider the model selection consistency of the graph-constrained regularization procedure under the settings when the number of covariates p also goes to infinity as $n \to \infty$, in which case, the assumptions and the regularity conditions for Theorems 3.1 and 3.2 become inappropriate as C does not converge and β may change as n grows. The following theorem shows that for the general scalings when p, q and n all go to infinity, under some additional conditions between p, q and n, GC-IC also guarantees that the Grace is sign consistent in selecting the true model.

Theorem 3.3—Suppose each column of X is normalized to L_2-norm of n and GC-IC (3.2) holds. Define $\rho := \min |(C_{11} + \frac{\lambda_2}{n} L_{11})^{-1} (C_{11} \beta_{(1)})|$ and $C_{\min} = \Lambda_{\min}(C_{11})$, where $\Lambda_{\min}(.)$ denotes the minimal eigenvalue. Let $W_{\max} = \max_{u,v} \{ w(u, v) \}$. If λ_1 and λ_2 are chosen such that

a. If $L_{12} = 0$,

$$\frac{\lambda_1^2}{n \log(p-q)} \to \infty,$$

or if $L_{12} \neq 0$,

$$\frac{\lambda_1^2}{\log(p-q) \left(n + \frac{\lambda_2^2 W_{\max}^2}{n C_{\min}} \right)} \to \infty;$$

b.

$$\frac{1}{\rho} \left\{ \sqrt{\frac{\log q}{n C_{\min}}} + \frac{\lambda_1}{n} \| (C_{11} + \frac{\lambda_2}{n} L_{11})^{-1} \operatorname{sign}(\beta_{(1)}) \|_\infty \right\} \to 0,$$

then the Grace $\hat\beta(\lambda_1, \lambda_2)$ is sign consistent as $n \to \infty$.

A proof analogous to Jia and Yu (2008) can be found in Supplemental Materials. Theorem 3.3 gives a general sign consistency result for the Grace for general scalings of p, q and n. If $C_{\min} \geq a$ for some $a > 0$ and $\rho \geq \rho_0$ for some $\rho_0 > 0$, it is easy to check that the conditions $\log q / n \to \infty$ and $\lambda_1 \sqrt{q}/n \to 0$ guarantee that condition (b) in Theorem 3.3 holds. In the settings when p and q are fixed, if C_{11} converges to a non-negative definite matrix C_{110}, ρ converges to a non-negative number. In addition, it is easy to verify that the conditions in Theorem 3.2 guarantee that the conditions (a) and (b) in Theorem 3.3 hold.

4. Monte Carlo Simulations

We conducted Monte Carlo simulations to evaluate the proposed Grace and aGrace procedures and to compare the performance of this procedure with Lasso and Enet in terms of prediction errors and identification of relevant variables. We simulated the graph to mimic gene regulation modules. We use genes and variables interchangeably in this section. We assumed that the graph consisted of 200 unconnected regulatory modules with 200 transcription factors (TFs) and each regulates 10 different genes for a total of 2,200 variables. Among these modules and genes, we further assumed that four TFs and their 10 regulated genes (for a total of 44 variables) were associated with the response based on the following model

$$Y = \sum_{u=1}^{44} \beta_u X_u + \varepsilon. \tag{4.1}$$

We considered two different models. For the first model, we assume that the coefficients in model (4.1) were specified as

$$\beta = (2, \underbrace{\frac{2}{\sqrt{10}}, \ldots, \frac{2}{\sqrt{10}}}_{10}, -2, \underbrace{\frac{-2}{\sqrt{10}}, \ldots, \frac{-2}{\sqrt{10}}}_{10}, 4, \underbrace{\frac{4}{\sqrt{10}}, \ldots, \frac{4}{\sqrt{10}}}_{10}, -4, \underbrace{\frac{-4}{\sqrt{10}}, \ldots, \frac{-4}{\sqrt{10}}}_{10}, 0, \ldots, 0),$$

and the ε was random mean-zero normal error with variance $\sigma^2 = \sum_u \beta_u^2/4$. For each TF, the X value was simulated from a $N(0, 1)$ distribution, and conditional on the value of the TF, we simulated the expression levels of the genes that they regulated from a conditional normal distribution with correlations of 0.2, 0.5 and 0.9, respectively. We therefore had a total of 2200 variables and 44 of them were relevant. For the second model, we considered the case when the regulated genes had different signs in regression coefficients, where the regression coefficients in model (4.1) have the same absolute values as in Model 1, but for each simulated module, three out of the 10 genes regulated by the TF had different signs from the other 7 genes. The X values were generated in the same way as in previous simulations. In this model, genes that are regulated by the same transcription factor are assumed to have different regression coefficients.

For each model, our simulated data consisted of a training set, an independent validation set and an independent test set with a sample size of 200 for all three datasets. Models were fitted on training data only, and the validation data were used to select the tuning parameters. We computed the prediction mean-squared errors on the test dataset. For each model, we repeated the simulations 100 times. Table 1 shows the prediction mean-square errors for several different procedures. For Model 1 when the neighboring genes have the same signs in regression coefficients, we observed that the Grace gave the smallest prediction errors for all four models with different correlations among the predictors. Both Grace and Enet performed better than Lasso in prediction. When the correlation is very high, the prediction errors of these procedures were comparable, however, Grace still gave the smallest prediction error among the procedures compared. When the signs of the regression coefficients were the same, aGrace was reduced to Grace and gave the same prediction results. For Model 2 when the neighboring variables have different sings of coefficients, aGrace adjusting for the signs of the regression coefficients gave the smallest prediction errors, further indicating the importance of adjusting for the signs in the regularization. In general, Grace gave similar prediction results as the Enet, except when the correlation between the transcription factors and their regulated genes was very high, in which case Enet resulted in slightly smaller prediction error.

To compare the performance on variable selection, Figure 1 shows the receiver operating characteristic (ROC) curves of several different procedures in selecting the relevant variables for the models with correlation of 0.2 and 0.9 between the TF and their regulated genes. For Grace, aGrace and Enet, the ROC curves were obtained as a function of the sparsity parameter λ_1 with tuning parameter λ_2 selected based on 5-fold cross-validation among the values of 0.1, 1, 10, 100 and 1000. For Model 1 when the neighboring genes have the same signs in regression coefficients (Figure 1 plots (a) and (b)), Grace gave much larger areas under the ROC cruves than Enet and Lasso, indicating better performance in variable selection for Grace. In addition, five-fold cross validation always chose the largest $\lambda_2 = 1000$ for Grace and $\lambda_2 = 0.1$ for Enet in all 100 replications. For Model 2 when the neighboring variables have different sings of coefficients, aGrace adjusting for the signs of the regression coefficients performed better than the other three procedures compared, resulting in larger areas under the curves, and Grace still performed better than Lasso and Enet on variable selections in both low and high correlation scenarios. When the correlation among the relevant variables is low, the 5-fold CV selected $\lambda_2 = 1000$ for aGrace and $\lambda_2 = 0.1$ for Enet in all 100 replications and selected $\lambda_2 = 100$ for Grace in most of the replications. When the correlation among the relevant variables was high, the 5-fold CV selected $\lambda_2 = 1000$ for aGrace and $\lambda_2 = 100$ for Grace in most of the 100 replications and selected $\lambda_2 = 0.1$ for Enet in all the replications.

5. Application to Network-based Analysis of Gene Expression Data

To demonstrate the proposed method, we consider the problem of identifying age-dependent molecular modules based on the gene expression data measured in human brains of individuals of different ages published in Lu *et al.*(2004). In this study, the gene expression levels in the postmortem human frontal cortex were measured using the Affymetrix arrays for 30 individuals ranging from 26 to 106 years of age. To identify the aging-regulated genes, Lu *et al.* (2006) performed simple linear regression analysis for each gene with age as a covariate. We analyzed this dataset by combining the KEGG regulatory network information with the gene expression data (Kanehisa and Goto, 2002). In particular, we limited our analysis to the genes that can be mapped to the KEGG regulatory work and focused on the problem of identifying the subnetworks of the KEGG regulatory network that are associated with human brain aging. By merging the gene expression data with the KEGG regulatory pathways, the final KEGG network includes 1305 genes and 5288 edges.

We treated the logarithm of the individual age as the response variable and the expression levels (after $\log_{10}$ transformation) of 1305 genes on the KEGG network as the explanatory variables in our analysis. Table 1 shows the results of several different procedures where the tuning parameters were selected by five-fold cross validations. Overall, we observed that the Lasso selected the fewest number of genes with relatively large cross-validation errors and Grace and Enet selected roughly the same number of genes with similar CV errors. However, the adaptive Grace resulted in more identified genes with similar CV errors than the other two procedures. Figure 2 shows the subnetworks identified by four different estimation procedures. As a comparison, we also included the genes selected by Lasso although it did not select any linked pairs of genes on the KEGG network. It is interesting to note that as we impose more constraints on the regression coefficients, more linked genes are identified. Both Enet and Grace identified some common subnetworks that were associated with brain aging. These included fibroblast growth factors (FGF) and its receptors. It has been demonstrated that FGFs are associated with many developmental processes including neural induction (Bottcher and Niehrs, 2005) and are involved in multiple functions including cell proliferation, differentiation, survival and aging (Yeoh and de Haan, 2007). It is also interesting to observe that mitogen-activated protein kinase (MAPK) (MAPK1 and MAPK9) and the specific MAPK kinase (MAP2K) were also identified by Enet and Grace. The MAPKs play important roles in induction of apoptosis (Hayesmoore *et al.*, 2008). Other interesting genes include RAS protein-specific guanine nucleotide-releasing factor 1 (RASGRF1), the functionality of which is highly significant in various contexts of the central nervous system. In the hippocampus, RASGRF2 has been shown to interact with the NR2A subunits of NMDARs, triggering Ras-ERK activation and induction of long-term potentiation, a form of neuronal plasticity that contributes to memory storage in the brain (Tian *et al.*, 2004; Lu *et al.*, 2004). Finally, the insulin receptor gene (INSR) is also identified. INSR binds insulin (INS) and regulates energy metabolism. Evidence from model organisms, including results from fruit flies (Tatar *et al.*, 2001) and roundworms (Kimura *et al.*, 1997), relates INSR homologues to aging, most likely as part of the GH1/IGF1 axis. These results indicated that our method can indeed recover some biologically interesting molecular modules or KEGG subnetworks that are related to brain aging in humans.

It is important to point out that the adaptive Grace identified several small sets of connected genes that were missed by Enet or the standard Grace. One of the subnetworks included EPHRIN and Eph receptor, both of which were found to be related to neural development and entohino-hippocampal axon targeting (Flanagan and Vaderhaeghen, 1998; Stein *et al.*, 1999). Another subnetwork was part of the Jak-State signaling, which is important in both mature and aging brains (De-Frajaa *et al.*, 2000). Aging was also found to be associated with

increased human T cell CC chemokine receptor gene expression (Yung *et al.*, 2003). Other interesting subnetworks included PVRL3-PVRL1 that are associated with cell adhesion.

6. Discussion

We have introduced and studied the theoretical properties of a graph-constrained regularized estimation procedure for linear regressions in order to incorporate information coded in graphs. Such a regularization procedure can also be regarded as a penalized least squared estimation where the penalty is defined as a combination of the L_1 penalty and L_2 penalty on degree-scaled differences of coefficients between variables linked on the graphs. This penalty function induces both sparsity and smoothness with respect to the graph structure of the regression coefficients. Simulation studies indicated that when the coefficients are similar for variables that are neighbors on the graph, the proposed procedure has better prediction and identification performance than other commonly used regularization procedures such as Lasso and elastic net regressions. Such improvement is resulted from effectively utilizing the neighboring information in estimating the regression coefficients. If the smoothness assumption on the coefficients does not hold, we expect that the cross-validation selects a very small value of λ_2 and therefore the proposed procedure would perform similarly as the Lasso. In analysis of the brain ageing gene expression data, different from Lasso, the new procedure tends to identify set of linked genes on the networks, which often leads to better biological interpretation of the genes identified. Although the methods presented are largely motivated by applications in genomic data, they can be applied to other settings when the covariates are nodes on general graphs, such as in image analysis.

Although the methods presented in this paper were developed mainly for linear models, similar methods can be developed for the generalized linear models and the censored survival data regression models, where we can use the negative of the logarithm of the likelihood or partial likelihood as the loss function. Similar to the techniques presented in Friedman *et al.* (2007) and Wu and Lange (2008), we can use the coordinate descent procedure together with the iterative reweighted least square to obtain the solution path. Such models have great applications in genomic data analysis in identifying the genes or subnetworks that are associated with binary or censored survival data outcomes. Other extensions include replacing the L_1 part of the Grace penalty with other sparse penalty functions such as SCAD or bridge penalty (Huang *et al.*, 2008). Important future research also includes how to handle covariates that are linked on directed graphs. Finally, to incorporate the fact that the linked nodes might be negatively correlated and the corresponding regression coefficients may have different signs, we introduced an adaptive sign-adjusted graph-constrained regularization procedure and showed that such a procedure can perform better then the original graph-constrained regularization. The theoretical property of such an adaptive procedure is unknown and is an area for future research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References

1. Friedman J, Hastie T, Hoefling H, Tibshirani R. Path-wise coordinate optimization. Annals of Applied Statistics. 2007; 1:302–332.

2. Bickel, PL.; Ritov, Y.; Tsybakov, AB. Technical report, Statistics. UC Berkeley; 2008. Hierarchical selection of variables in sparse high-dimensional regression.

3. Bottcher RT, Niehrs C. Fibroblast growth factor signaling during early vertebrate development. Endocrine Reviews. 2005; 26:6377.

4. Chung, F. CBMS Reginal Conferences Series. Vol. 92. American Mathematical Society; Providence: 1997. Spectral Graph Theory.

5. De-Fraja C, Conti L, Govoni S, Battaini F, Cattaneo E. STAT signalling in the mature and aging brain. International Journal of Developmental Neuroscience. 2000; 18:439–446. [PubMed: 10817928]

6. Donoho D, Johnstone I. Ideal spatial adaptation via wavelet shrinkage. Biometrika. 1994; 81:425–455.

7. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. Annals of Statistics. 2004; 32:407499.

8. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of American Statistical Association. 2001; 96:1348–1360.

9. Fan J, Peng H. Nonconcave penalized likelihood with a diverging number of parameters. Annals of Statistics. 2004; 32:928961.

10. Flanagan JG, Vanderhaeghen P. The ephrins and Eph receptors in neural development. Annual Review Neuroscience. 1998; 21:30945.

11. Hayesmoore JB, Bray NJ, Cross WC, Owen MJ, O'Donovan MC, Morris HR. The effect of age and the H1c MAPT haplotype on MAPT expression in human brain. Neurobiology of Aging. 2008 in press.

12. Huang J, Xie H. Asymptotic oracle properties of SCAD-penalized least squares estimators. IMS Lecture Notes - Monograph Series, Asymptotics: Particles, Processes and Inverse Problems. 2007; 55:149–166.

13. Jia, J.; Yu, B. Technical Report 756, Statistics. UC Berkeley; 2008. On model selection consistency of elastic net when $p \gg n$.

14. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Research. 2002; 28:27–30. [PubMed: 10592173]

15. Kimura KD, Tissenbaum HA, Liu Y, Ruvkun G. daf-2, an insulin receptor-like gene that regulates longevity and diapause in Caenorhabditis elegans. Science. 1997; 277:942–946. [PubMed: 9252323]

16. Li C, Li H. Network-constrained regularization and variable selection for analysis of genomic data. Bioinformatics. 2008; 24:1175–1182. [PubMed: 18310618]

17. Lu T, Pan Y, Kao SY, Li C, Kohane I, Chan J, Yankner BA. Gene regulation and DNA damage in the aging human brain. Nature. 2004; 429:883–891. [PubMed: 15190254]

18. Nardi Y, Rinado A. On the asymptotic properties of the group lasso estimator for linear models. Electronic Journal of Statistics. 2008; 2:605–633.

19. Portnoy S. Asymptotic behavior of M-estimators of p regression parameters when p/n is large. I. Consistency. Annals of Statistics. 1984; 12:1298–1309.

20. Stein E, Savaskan NE, Ninnemann O, Nitsch R, Zhou R, Skutella T. A role for the Eph ligand ephrin-A3 in entorhino-hippocampal axon targeting. Journal of Neuroscience. 1999; 19:8885–8893. [PubMed: 10516308]

21. Tatar M, Kopelman A, Epstein D, Tu MP, Yin CM, Garofalo RS. A mutant Drosophila insulin receptor homolog that extends life-span and impairs neuroendocrine function. Science. 2001; 292:107–110. [PubMed: 11292875]

22. Tian X, Gotoh T, Tsuji K, Lo EH, Huang S, Feig LA. Developmentally regulated role for Ras-GRFs in coupling NMDA glutamate receptors to Ras, Erk and CREB. EMBO J. 2004; 23(7):1567–75. [PubMed: 15029245]

23. Tibshirani RJ. Regression shrinkage and selection via the lasso. Journal of Royal Statistical Society B. 1996; 58:267–288.

24. Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K. Sparsity and smoothness via the fused lasso. Journal of Royal Statistical Society SER B. 2005; 67:91–108.

25. Wu TT, Lange K. Coordinate descent algorithms for lasso penalized regression. Annals of Applied Statistics. 2008; 2:224–244.

26. Yeoh JS, de Haan G. Fibroblast growth factors as regulators of stem cell self-renewal and aging. Mechanisms of Ageing and Development. 2007; 128:17–24. [PubMed: 17118424]

27. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. Journal of Royal Statistical Society B. 2006; 68:49–67.

28. Yung RL, Mo R. Aging Is Associated with Increased Human T Cell CC Chemokine Receptor Gene Expression. Journal of Interferon & Cytokine Research. 2003; 23:575–582. [PubMed: 14585197]

29. Zhang C, Huang J. The sparsity and bias of the Lasso selection in high-dimensional linear regression. Annals of Statistics. 2006; 36:1567–1594.

30. Zhao P, Yu B. On Model Selection Consistency of Lasso. Journal of Machine Learning Research. 2006; 7:2541–2567.

31. Zhou, D.; Bousquet, O.; Lal, T.; Weston, J.; Scholkopf, B. NIPS. Vol. 16. Cambridge, MA, USA: MIT press; 2004. Learning with local and global consistency; p. 321-328.

32. Zhu, X. Technical report 1530, Computer Sciences. University of Wisconsin-Madison; 2005. Semi-supervised learning literature survey.

33. Zou H. The adaptive lasso and its oracle properties. Journal of the American Statistical Association. 2006; 101:1418–1429.

34. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of Royal Statistical Society SER B. 2005; 67:301–320.

35. Zou H, Zhang HH. On the adaptive elastic net with a diverging number of parameters. Annals of Statistics. 2009; 37:1733–1751. [PubMed: 20445770]
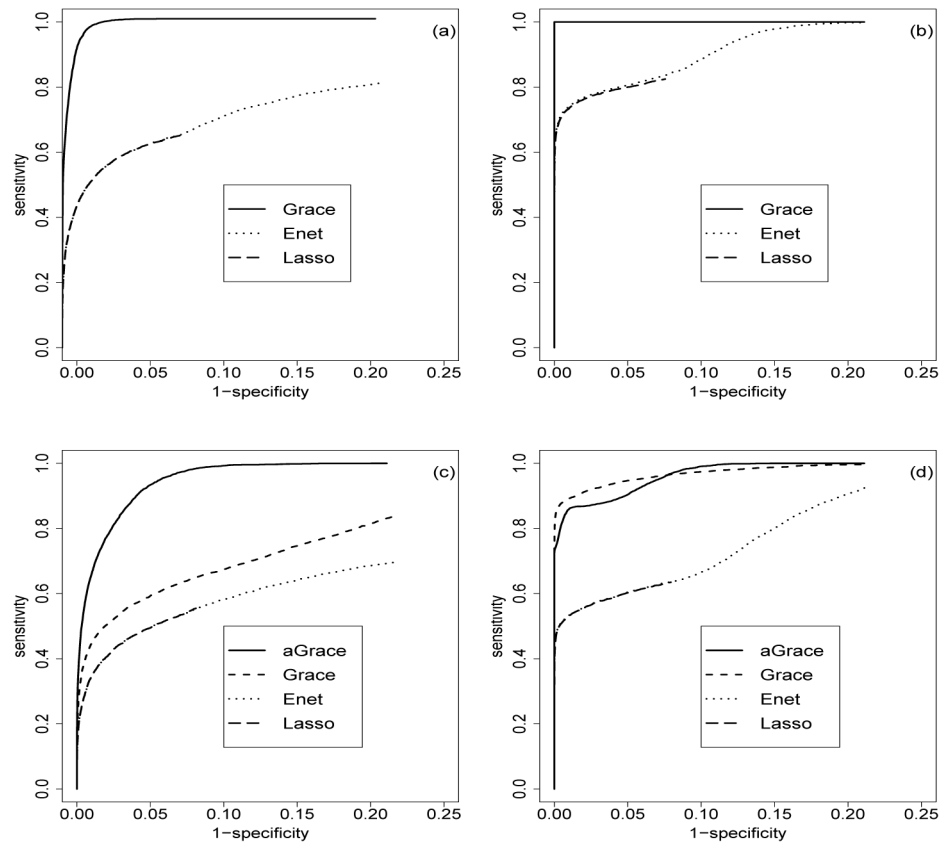
Fig 1.
Comparison of ROCs for Grace, aGrace, Enet and Lasso for Model 1 (plots (a) and (b)) and Model 2 (plots (c) and (d)) and for correlations of 0.2 (plots (a) and (c)) and 0.9 (plots (b) and (d)). The ROCs were calculated as a function of the sparsity parameter λ_1. For Grace, aGrace and Enet, the tuning parameter λ_2 was selected based on 5-fold CV.
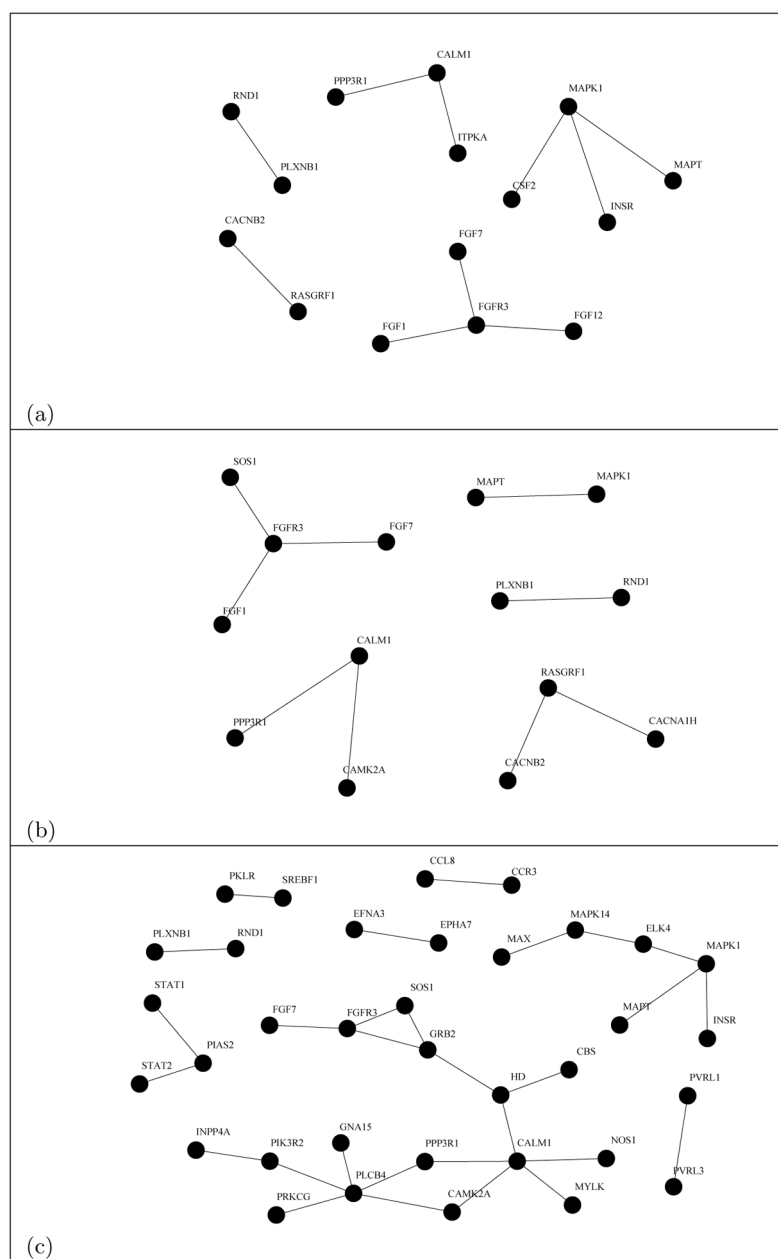
Fig 2.
Subnetworks identified by (a) Elastic net (Enet), (b) Grace and (c) aGrace for brain aging gene expression data (only those genes that are linked on the KEGG network are plotted).

Table 1

Comparison of prediction mean-square errors (SE) using Grace, aGrace, Enet and Lasso for three different correlation structures of 0.2, 0.5 and 0.9 between the transcription factors and their regulated genes for each of the two models considered. The results are based on 100 replications.

Method	Model 1 (Cor)			Model 2 (Cor)		
	0.2	0.5	0.9	0.2	0.5	0.9
Grace	24.93 (2.97)	23.22 (2.41)	22.56 (2.20)	53.08 (6.45)	42.07 (5.03)	28.20 (2.87)
aGrace	24.93 (2.97)	23.22 (2.41)	22.56 (2.20)	27.70 (3.66)	26.23 (3.03)	25.55 (2.76)
Enet	51.33 (6.65)	37.37 (4.69)	25.82 (2.67)	56.18 (7.22)	45.65 (5.81)	27.33 (2.72)
Lasso	53.41 (6.68)	40.30 (4.94)	27.82 (2.98)	57.62 (7.01)	47.65 (5.53)	29.23 (2.78)

Table 2

Results of analysis of brain aging gene expression data by four different procedures, including the number of genes selected (No. Genes), the number of linked KEGG edges (No. Edges), five-fold cross-validation error (CV error) and the values of the tuning parameters selected (λ_2 for Grace, aGrace and Enet and $s_1 = \Sigma_v |\beta_v|$.)

	No. Genes	No. Edges	CV error	Tuning parameters
Grace	45	9	0.079	λ_2=0.1, s_1= 4.72
aGrace	73	28	0.077	λ_2=0.01, s_1=6.97
Enet	41	10	0.077	λ_2=1.0, s_1=5.64
Lasso	18	0	0.098	s_1=5.65