Published in final edited form as:

Biometrics. 2016 June ; 72(2): 484–493. doi:10.1111/biom.12418.

A Significance Test for Graph-Constrained Estimation

Sen Zhao and

Department of Biostatistics, University of Washington, Seattle, Washington, U.S.A

Ali Shojaie

Department of Biostatistics, University of Washington, Seattle, Washington, U.S.A

Sen Zhao: senz@u.washington.edu; Ali Shojaie: ashojaie@u.washington.edu

Summary

Graph-constrained estimation methods encourage similarities among neighboring covariates presented as nodes of a graph, and can result in more accurate estimates, especially in high-dimensional settings. Variable selection approaches can then be utilized to select a subset of variables that are associated with the response. However, existing procedures do not provide measures of uncertainty of the estimates. Further, the vast majority of existing approaches assume that available graph accurately captures the association among covariates; violations to this assumption could severely hurt the reliability of the resulting estimates. In this paper, we present a new inference framework, called the Grace test, which produces coefficient estimates and corresponding *p*-values by incorporating the external graph information. We show, both theoretically and via numerical studies, that the proposed method asymptotically controls the type-I error rate regardless of the choice of the graph. We also show that when the underlying graph is informative, the Grace test is asymptotically more powerful than similar tests that ignore the external information. We study the power properties of the proposed test when the graph is not fully informative and develop a more powerful Grace-ridge test for such settings. Our numerical studies show that as long as the graph is reasonably informative, the proposed inference procedures deliver improved statistical power over existing methods that ignore external information.

Keywords

Biological networks; Graph-constrained estimation; High-dimensional data; Significance test; Variable selection

1. Introduction

Interactions among genes, proteins and metabolites shed light into underlying biological mechanisms, and clarify their roles in carrying out cellular functions (Zhu et al., 2007; Michailidis, 2012). This has motivated the development of many statistical methods to

8. Supplementary Materials

The Web Appendix referenced in Section 2, Section 3, Section 5 and Section 6 are available with this paper at the *Biometrics* website on Wiley Online Library. The R codes for the simulation study and real data example are also available at the *Biometrics* website.

incorporate existing knowledge of biological networks into data analysis (see e.g. Kong et al., 2006; Wei and Pan, 2008; Shojaie and Michailidis, 2009, 2010b). Such methods can lead to identification of novel biological mechanisms associated with the onset and progression of complex diseases (see e.g. Khatri et al., 2012).

External network information may be summarized using an undirected weighted graph $G = (V, E, W)$, whose node set $V = \{1, \ldots, p\}$ corresponds to p covariates. The edge set E of the graph encodes similarities among covariates, in the sense that two vertices $u, v \in V$ are connected with an edge $e = (u \sim v) \in E$ if covariates u and v are "similar" to each other. The similarity between neighboring nodes $(u \sim v)$ is captured by weights $w(u, v)$. Such similarities can for instance correspond to interactions between genes or phylogenetic proximities of species.

A popular approach for incorporating network information is to encourage smoothness in coefficient estimates corresponding to neighboring nodes in the network using a *network smoothing penalty* (Li and Li, 2008; Slawski et al., 2010; Pan et al., 2010; Li and Li, 2010; Huang et al., 2011; Shen et al., 2012). This approach can also be generalized to induce smoothness among similar covariates defined based on a distance matrix or "kernel" (Randolph et al., 2012) which, for instance, capture similarities among microbial communities according to lineages of a phylogenetic tree (Fukuyama et al., 2012).

The smoothness induced by the network smoothing penalty can result in more accurate parameter estimations, particularly when the sample size n is small compared to the number of covariates p. Sparsity-inducing penalties, like the ℓ_1 penalty (Li and Li, 2008, 2010) or the minimum convex penalty (MCP) (Huang et al., 2011), can then be used to select a subset of covariates $\boldsymbol{X}$ associated with the response $\boldsymbol{y}$ for improved interpretability and reduced variability. It has been shown that, under appropriate assumptions, the combination of network smoothing and sparsity-inducing penalties can consistently select the subset of covariates associated with the response (Huang et al., 2011). However, such procedures do not account for the uncertainty of the estimator, and in particular, do not provide p-values.

A number of new approaches have recently been proposed for formal hypothesis testing in penalized regression, including resampling and subsampling approaches (Meinshausen and Bühlmann, 2010), ridge test with deterministic design matrices (Bühlmann, 2013), and the low-dimensional projection estimator (LDPE) for ℓ_1-penalized regression (Zhang and Zhang, 2014; van de Geer et al., 2014). However, there are currently no inference procedures available for methods that incorporate external information using smoothing penalties. Inference procedures for kernel machine learning methods (Liu et al., 2007), on the other hand, test the global association of covariates and are hence not appropriate for testing the association of individual covariates.

Another limitation of existing approaches that incorporate external network information, including those using network smoothing penalties, is their implicit assumption that the network is accurate and informative. However, existing networks may be incomplete or inaccurate (Hart et al., 2006). As shown in Shojaie and Michailidis (2010a), such inaccuracies can severely impact the performance of network-based methods. Moreover,

even if the network is accurate and complete, it is often unclear whether network connectivities correspond to similarities among corresponding coefficients, which is necessary for methods based on network smoothing penalties.

To address the above shortcomings, we propose a testing framework, the *Grace test*, which incorporates external network information into high-dimensional regression and corresponding inferences. The proposed framework builds upon the graph-constrained estimation (Grace) procedure of Li and Li (2008), Slawski et al. (2010) and Li and Li (2010), and utilizes recent theoretical developments for the ridge test by Bühlmann (2013). As part of our theoretical development, we generalize the ridge test with fixed design to the setting with random design matrices $\boldsymbol{X}$. This generalization was suggested in the discussion of Bühlmann (2013) as a possible extension of the ridge test, and results in improved power compared to the original proposal.

Our theoretical analysis shows that the proposed testing framework controls the type-I error rate, regardless of the informativeness or accuracy of the incorporated network. We also show, both theoretically and using simulation experiments, that if the network is accurate and informative, the Grace test offers improved power over existing approaches that ignore such information. Finally, We propose an extension of the Grace test, called the Grace-ridge or *GraceR* test, for settings where the network may be inaccurate or uninformative.

The rest of the paper is organized as follows. In Section 2, we introduce the Grace estimation procedure and the Grace test. We also formally define the "informativeness" of the network. Section 3 investigates the power of the Grace test, in comparison to its competitors. In Section 4, we propose the Grace-ridge (GraceR) test for robust estimation and inference with potentially uninformative networks. We apply our methods to simulated data in Section 5 and to data from The Cancer Genome Atlas (TCGA) in Section 6. We end with a discussion in Section 7. Due to space limitations, proofs of theoretical results and additional details of simulated and real-data analyses are gathered in the online Supplementary Material.

Throughout this paper, we use normal lowercase letters to denote scalars, bold lowercase letters to denote vectors and bold uppercase letters to denote matrices. We denote columns of an $n \times p$ matrix $\boldsymbol{X}$ by $\boldsymbol{x}_j$, $j = 1, \ldots, p$ and its rows by $\boldsymbol{x}^i$, $i = 1, \ldots, n$. For any two symmetric matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, we denote $\boldsymbol{A} \preceq \boldsymbol{B}$ if $\boldsymbol{B} - \boldsymbol{A}$ is positive semi-definite, or $\lambda_0(\boldsymbol{B} - \boldsymbol{A}) \geq 0$, where λ_0 denotes the smallest eigenvalue of a symmetric matrix. For an index set J, we denote by $\boldsymbol{A}_{(J,J)}$ the $|J| \times |J|$ sub-matrix corresponding to the rows and columns indexed by J.

Finally, for a p-vector $\boldsymbol{\beta}$, we let $\|\boldsymbol{\beta}\|_k \triangleq \left(\sum_{i=1}^{p} |\beta_i|^k\right)^{1/k}$ for $k \in \mathbb{Z}^+$ and $\|\boldsymbol{\beta}\|_\infty \triangleq \max_i \beta_i$.

2. The Grace Estimation Procedure and the Grace Test

2.1 The Grace Estimation Procedure

Let $\boldsymbol{L}$ be the matrix encoding the external information in an undirected weighted graph $G = (V, E, W)$. In general, $\boldsymbol{L}$ can be any positive semi-definite matrix, or kernel, capturing the "similarity" between covariates. In this paper, however, we focus on the case where $\boldsymbol{L}$ is the graph Laplacian matrix,

$$\boldsymbol{L}_{(u,v)} \triangleq \begin{cases} d_u & \text{if } u=v \\ -w(u,v) & \text{if } u \text{ and } v \text{ are connected} \\ 0 & \text{otherwise} \end{cases},$$

with $d_u = \Sigma_{v\sim u}\, w(u, v)$ denoting the degree of node u. We also assume that weights $w(u, v)$ are nonnegative. However, the definition of Laplacian and the analysis in this paper can be generalized to also accommodate negative weights (Chung, 1997).

Let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) \in \mathbb{R}^{n\times p}$ be the $n\times p$ design matrix and $\boldsymbol{y} \in \mathbb{R}^n$ be the response vector in the linear model

$$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}^*+\boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon}\sim N_n(\boldsymbol{0},\sigma_\varepsilon^2\boldsymbol{I}_n), \quad \boldsymbol{x}^i\sim^{iid} N_p(\boldsymbol{0},\boldsymbol{\Sigma}) \text{ for } i=1,\ldots,n. \quad (1)$$

Multivariate normality of covariates is commonly assumed in analysis of biological networks, particularly, when estimating interactions among genes or proteins using Gaussian graphical models (see e.g. de la Fuente et al., 2004). Interestingly, the underlying assumption of network smoothing penalties – that connected covariates after scaling have similar associations with the response – is also related to the assumption of multivariate normality (Shojaie and Michailidis, 2010b). In this paper, we assume $\boldsymbol{y}$ is centered and columns of $\boldsymbol{X}$ are centered and scaled, i.e. $\sum_{i=1}^n y_i=0$ and $\sum_{i=1}^n X_{(i,j)}=0,\ \boldsymbol{x}_j^\top\boldsymbol{x}_j=n$ for $j = 1, \ldots, p$. We denote the scaled Gram matrix by $\hat{\boldsymbol{\Sigma}} \triangleq \boldsymbol{X}^\top\boldsymbol{X}/n$.

For a non-negative tuning parameter h, Grace solves the following optimization problem:

$$\hat{\boldsymbol{\beta}}(h)=\arg\min_{\boldsymbol{\beta}}\left\{\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+h\boldsymbol{\beta}^\top\boldsymbol{L}\boldsymbol{\beta}\right\}=\left(n\hat{\boldsymbol{\Sigma}}+h\boldsymbol{L}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{y}. \quad (2)$$

When $\boldsymbol{L}$ is the Laplacian matrix, $\boldsymbol{\beta}^\top\boldsymbol{L}\boldsymbol{\beta} = \Sigma_{u\sim v}(\boldsymbol{\beta}_u - \boldsymbol{\beta}_v)^2 w(u, v)$ (Huang et al., 2011). Hence, the Grace penalty $\boldsymbol{\beta}^\top\boldsymbol{L}\boldsymbol{\beta}$ encourages smoothness in coefficients of connected covariates, according to weights of edges. Henceforth, we call $\boldsymbol{L}$ the penalty weight matrix.

For any tuning parameter $h > 0$, Equation (2) will have a unique solution if $(n\hat{\boldsymbol{\Sigma}}+h\boldsymbol{L})$ is invertible. However, if $p > n$ and $rank(\boldsymbol{L}) < p$ this condition may not hold. With a Gaussian design $\boldsymbol{x}^i \sim^{iid} N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$, it follows from Bai (1999) that if $\liminf_{n\to\infty} \lambda_0(\boldsymbol{\Sigma}) > 0$, and if there exists a sequence of index sets $C_n \subset \{1, \ldots, p\}$, $\lim_{n\to\infty} |C_n|/n < 1$, such that $\liminf_{n\to\infty} \lambda_0(\boldsymbol{L}_{(V\setminus C_n, V\setminus C_n)}) > 0$, then $(n\hat{\boldsymbol{\Sigma}}+h\boldsymbol{L})$ is almost surely invertible. In this section we hence assume that $(n\hat{\boldsymbol{\Sigma}}+h\boldsymbol{L})$ is invertible. This condition is relaxed in Section 4, when we propose the more general Grace-ridge (GraceR) test.

As mentioned in the Introduction, several methods have been proposed to select the subset of relevant covariates for Grace. For example, Li and Li (2008, 2010) added an ℓ_1 penalty to the Grace objective function,

$$\hat{\boldsymbol{\beta}}_{\ell_1}(h, h_1) = \arg\min_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + h\boldsymbol{\beta}^\top \boldsymbol{L}\boldsymbol{\beta} + h_1\|\boldsymbol{\beta}\|_1 \right\}. \quad (3)$$

Huang et al. (2011) instead added the MCP and proposed the sparse Laplacian shrinkage (SLS) estimator. While these methods perform automatic variable selection, they do not provide measures of uncertainty, i.e. confidence intervals or *p*-values. In this paper, we instead propose an inference procedure that provides *p*-values for estimated coefficients from Equation (2). The resulting *p*-values can then be used to assess the significance of individual covariates, and select a subset of relevant variables.

2.2 The Grace Test

Before introducing the Grace test, we present a lemma that characterizes the bias of the Grace estimation procedure.

Lemma 1—*For any* $h > 0$*, assume* $(n\hat{\boldsymbol{\Sigma}} + h\boldsymbol{L})$ *is invertible. Then, given* $\boldsymbol{X}$*,* $\hat{\boldsymbol{\beta}}(h)$ *as formulated in* (2) *is an unbiased estimator of* $\boldsymbol{\beta}^*$ *if and only if* $\boldsymbol{L}\boldsymbol{\beta}^* = \boldsymbol{0}$*. Moreover,*

$$\|\mathbf{Bias}(\hat{\boldsymbol{\beta}}(h)|\boldsymbol{X})\|_2 \leq \frac{h\|\boldsymbol{L}\boldsymbol{\beta}^*\|_2}{\lambda_0(n\hat{\boldsymbol{\Sigma}} + h\boldsymbol{L})}. \quad (4)$$

Because the bias of the Grace estimator depends directly on the magnitude of $\boldsymbol{L}\boldsymbol{\beta}^*$, we consider $\boldsymbol{L}$ to be informative if $\boldsymbol{L}\boldsymbol{\beta}^*$ is small. According to Lemma 1, the Grace estimator will be unbiased only if $\boldsymbol{\beta}^*$ lies in the space spanned by the eigenvectors of $\boldsymbol{L}$ with 0 eigenvalues. In reality, however, this condition cannot be checked from data. Thus, to control the type-I error rate, we must adjust for this potential estimation bias.

Our testing procedure is motivated by the ridge test proposed in Bühlmann (2013), which we briey discuss next. First, note that ridge is also a biased estimator of $\boldsymbol{\beta}^*$, and its *estimation bias* is negligible only if the ridge tuning parameter is close to zero. In addition to the estimation bias, Bühlmann (2013) also accounted for the *projection bias* of ridge regression for a *fixed* design matrix $\boldsymbol{X}$. This is because for fixed design matrices with $p > n$, $\boldsymbol{\beta}^*$ is not uniquely identifiable, as there are infinitely many $\boldsymbol{\beta}$'s such that $\mathrm{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}$. Using ridge regression, $\boldsymbol{\beta}^*$ is only estimable if it lies in the row space of $\boldsymbol{X}$, $\mathscr{R}(\boldsymbol{X})$, which is a proper subspace of $\mathbb{R}^p$ when $p > n$. If $\boldsymbol{\beta}^*$ does not lie in this subspace, the ridge estimated regression coefficient is indeed the projection of $\boldsymbol{\beta}^*$ onto $\mathscr{R}(\boldsymbol{X})$, which is not identical to $\boldsymbol{\beta}^*$. This gives rise to the projection bias.

To account for these two types of biases, Bühlmann (2013) proposed to shrink the ridge estimation bias to zero by shrinking the ridge tuning parameter to zero, while controlling the

projection bias using a stochastic bias bound derived from a lasso initial estimator. A side effect of shrinking the ridge tuning parameter to zero is that the variance of covariates with high multi-collinearity could become large; this would hurt the statistical power of the ridge test. In addition, the stochastic bound for the projection bias is rather loose. This double-correction of bias further compromises the power of the ridge test.

In this paper, we develop a test for random design matrices, which was suggested in the discussion of Bühlmann (2013) as a potential extension. With random design matrices, we do not incur any projection bias. This is because the regression coefficients in this case are uniquely identifiable as $\boldsymbol{\Sigma}^{-1}\text{Cov}(\boldsymbol{X}, \boldsymbol{y})$ under the joint distribution of $(\boldsymbol{X}, \boldsymbol{y})$. Here, $\boldsymbol{\Sigma}$ denotes the population covariance matrix of covariates and $\text{Cov}(\boldsymbol{X}, \boldsymbol{y})$ is the population covariance between the covariates and the response; see Shao and Deng (2012) for a more elaborate discussion of identifiability for fixed and random design matrices.

To control the type-I error rate of the Grace test, we adjust for the potential estimation bias using a stochastic bound derived from an initial estimator. By adjusting for the estimation bias using a stochastic upper bound, the Grace tuning parameter needs not be very small. Thus, the variance of Grace estimator is less likely to be unreasonably large; this results in improved power for the Grace test. Power properties of the Grace test are more formally investigated in Section 3. Next, we formally introduce our testing procedure.

Consider the null hypothesis $H_0 : \beta_j^*=0$ for some $j \in \{1, \ldots, p\}$. Let $\tilde{\boldsymbol{\beta}}$ be an initial estimator with asymptotic ℓ_1 estimation accuracy, i.e. $\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = o_p(1)$. The Grace test statistic is defined as

$$\hat{\boldsymbol{z}}^G=\hat{\boldsymbol{\beta}}(h)+h(n\hat{\boldsymbol{\Sigma}}+h\boldsymbol{L})^{-1}\boldsymbol{L}\tilde{\boldsymbol{\beta}}, \quad (5)$$

where $\hat{\boldsymbol{\beta}}(h)$ is the Grace estimator from (2) with tuning parameter h. Plugging in (2) and adding and subtracting $h(n\hat{\boldsymbol{\Sigma}}+h\boldsymbol{L})^{-1}\boldsymbol{L}\boldsymbol{\beta}^*$, we can write

$$\hat{z}_j^G=\beta_j^*+Z_j^G+\gamma_j^G, \quad j=1,\ldots,p, \quad (6)$$

where

$$\begin{aligned} Z_j^G|\boldsymbol{X} \quad &\sim N\left(0, n\sigma_\varepsilon^2\left[(n\hat{\boldsymbol{\Sigma}}+h\boldsymbol{L})^{-1}\hat{\boldsymbol{\Sigma}}(n\hat{\boldsymbol{\Sigma}}+h\boldsymbol{L})^{-1}\right]_{(j,j)}\right), \\ \boldsymbol{\gamma}^G \quad &\triangleq h(n\hat{\boldsymbol{\Sigma}}+h\boldsymbol{L})^{-1}\boldsymbol{L}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*). \end{aligned}$$

Next, we derive an asymptotic stochastic bound for γ_j^G such that under the null hypothesis

$$|\gamma_j^G| \lesssim^{asy.} \Gamma_j^G \text{ or equivalently, } \lim_{n\to\infty} Pr\left(|\gamma_j^G| \leq \Gamma_j^G\right)=1. \quad (7)$$

Then, under the null hypothesis, $|\hat{z}_j^G| \lesssim^{asy} \cdot |Z_j^G| + \Gamma_j^G$, which allows us to asymptotically control the type-I error rate.

To complete our testing framework, we use the fact under suitable conditions and with proper tuning parameter h_{Lasso}, described in Theorem 1, the ℓ_1 estimation error of the lasso,

$$\tilde{\boldsymbol{\beta}}(h_{Lasso}) = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + h_{Lasso} \|\boldsymbol{\beta}\|_1 \right\}, \quad (8)$$

is asymptotically controlled (Bühlmann and van de Geer, 2011). We thus use the lasso estimator as the initial estimator for the Grace test, i.e. $\boldsymbol{\beta} \triangleq \tilde{\boldsymbol{\beta}}(h_{Lasso})$. Theorem 1 then constructs a Γ_j^G that satisfies Condition (7). First, we present required conditions.

- **A0**: $(n\hat{\boldsymbol{\Sigma}} + h\boldsymbol{L})$ is invertible.
- **A1**: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$ where $\boldsymbol{x}^i \sim^{iid} N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ for $i = 1, \ldots, n$ and $\boldsymbol{\varepsilon} \sim N_n(\boldsymbol{0}, \sigma_{\boldsymbol{\varepsilon}}\boldsymbol{I})$.
- **A2**: Let $S_0 \triangleq \{j : \beta_j^* \neq 0\}$ be the active set of $\boldsymbol{\beta}^*$ with cardinality $s_0 \triangleq |S_0|$. We have $s_0 = o([n/\log p]^{\xi})$ for some $0 < \xi < 1/2$.
- **A3**: The $\boldsymbol{\Sigma}$-compatibility condition (Bühlmann and van de Geer, 2011) in Definition 1 is met for the set S_0 with compatibility constant $\liminf_{n\to\infty} \phi_{\boldsymbol{\Sigma}}^2 = d > 0$, where d is a constant.
- **A4**: h and $\boldsymbol{L}$ are such that

$$\left[(n\hat{\boldsymbol{\Sigma}} + h\boldsymbol{L})^{-1} h\boldsymbol{L}\right]_{(j,j)} = \mathcal{O}_p\left(\left[\frac{n}{\log p}\right]^{\frac{1}{2}-\xi}\right).$$

Definition 1 ($\boldsymbol{\Sigma}$-Compatibility Condition)—For an index set $S \subset \{1, \ldots, p\}$ with cardinality s, define $\boldsymbol{\beta}^S$ and $\boldsymbol{\beta}^{S^c}$ such that $\beta_j^S \triangleq \beta_j 1_{\{j \in S\}}$, $\beta_j^{S^c} \triangleq \beta_j 1_{\{j \notin S\}}$. We say that the $\boldsymbol{\Sigma}$-compatibility condition is met for the set S with compatibility constant $\varphi_{\boldsymbol{\Sigma}} > 0$ if for all $\boldsymbol{\beta} \in \mathbb{R}^p$ in the cone $\|\boldsymbol{\beta}^{S^c}\|_1 \leq 3\|\boldsymbol{\beta}^S\|_1$, we have

$$\|\boldsymbol{\beta}^S\|_1^2 \leq \boldsymbol{\beta}^\top \boldsymbol{\Sigma} \boldsymbol{\beta} \frac{s}{\phi_{\boldsymbol{\Sigma}}^2}. \quad (9)$$

As discussed in Section 2.1, **A0** is required for uniqueness of the Grace estimator, and is shown to hold with probability tending to one under the Gaussian design (Bai, 1999). **A2** is a

standard assumption, and requires the number of relevant covariates to not grow too fast, so that the signal is not substantially diluted among those relevant covariates. Note that with $p = \mathcal{O}(\exp(n^{\nu}))$ for some $\nu < 1$, s_0 can grow to infinity as $n \to \infty$. The $\boldsymbol{\Sigma}$-compatibility condition in **A3** is closely related to the restricted eigenvalue assumption introduced in Bickel et al. (2009). Assumption **A4** is made for improved control of type-I error, and can be relaxed at a cost of potential loss of power with finite samples; see Remark 1. On the other hand, given $\boldsymbol{X}$ and $\boldsymbol{L}$, when $h/n \to \infty$, the eigenvectors and eigenvalues of $(n/h)\hat{\boldsymbol{\Sigma}}+\boldsymbol{L}$ converge to the eigenvectors and eigenvalues of $\boldsymbol{L}$. This indicates that $(n\hat{\boldsymbol{\Sigma}}+h\boldsymbol{L})^{-1}h\boldsymbol{L}$ converges to a diagonal matrix with diagonal entries equal to 0 or 1, and **A4** is satisfied.

Theorem 1—*Suppose Assumptions* **A0** – **A4** *are satisfied, and let* $\tilde{\boldsymbol{\beta}} \triangleq \tilde{\boldsymbol{\beta}}(\tilde{h}_{Lasso})$ *with the tuning parameter* $h_{Lasso} \asymp \sqrt{\log p/n}$. *Let*

$$\Gamma_j^G \triangleq h\|[(n\hat{\boldsymbol{\Sigma}}+h\boldsymbol{L})^{-1}\boldsymbol{L}]_{(j,-j)}\|_{\infty}\left(\frac{\log p}{n}\right)^{\frac{1}{2}-\xi}, \quad (10)$$

where $\|[(n\hat{\boldsymbol{\Sigma}}+h\boldsymbol{L})^{-1}\boldsymbol{L}]_{(j,-j)}\|_{\infty} \triangleq \max_{i:i\neq j}|(n\hat{\boldsymbol{\Sigma}}+h\boldsymbol{L})^{-1}\boldsymbol{L}|_{(j,i)}$ *is the maximum in absolute value of entries in row j without the diagonal entry. Then* Γ_j^G *satisfies condition* (7).

Under the null hypothesis $H_0 : \beta_j = 0$, *for any* $\alpha > 0$ *we have*

$$\lim_{n\to\infty} Pr\,(|\hat{z}_j^G| > \alpha) \leq \lim_{n\to\infty} \Pr\,(|Z_j^G| + \Gamma_j^G > \alpha). \quad (11)$$

Remark 1—If we instead consider

$$\Gamma_j^G = h\|[(n\hat{\boldsymbol{\Sigma}}+h\boldsymbol{L})^{-1}\boldsymbol{L}]_{(j,.)}\|_{\infty}\left(\frac{\log p}{n}\right)^{\frac{1}{2}-\xi},$$

we can relax Assumption **A4** and still control the asymptotic type-I error rate. Theorem 1 can then be similarly proved without **A4**. However, as $h/n \to \infty$, $(n\hat{\boldsymbol{\Sigma}}+h\boldsymbol{L})^{-1}h\boldsymbol{L}$ converges to a diagonal matrix, in which case $\|[(n\hat{\boldsymbol{\Sigma}}+h\boldsymbol{L})^{-1}h\boldsymbol{L}]_{(j,.)}\|_{\infty} \gg \|[(n\hat{\boldsymbol{\Sigma}}+h\boldsymbol{L})^{-1}h\boldsymbol{L}]_{(j,-j)}\|_{\infty}$. This looser stochastic bound may result in lower power in finite samples.

Theorem 1 shows that regardless of the choice of $\boldsymbol{L}$, the type-I error rate of the Grace test is asymptotically controlled. The stochastic bound Γ_j^G relies on the unknown sparsity parameter ξ. Following Bühlmann (2013) we suggest a small value of ξ, and use $\xi = 0.05$ in the simulation experiments in Section 5 and real data example in Section 6.

Using (11), we can test H_0 using the asymptotically valid two-sided p-value

$$P_j^G = 2\left(1-\Phi\left[\frac{(|\hat{z}_j^G|-\Gamma_j^G)_+}{\sqrt{\mathrm{Var}(Z_j^G|\boldsymbol{X})}}\right]\right), \quad (12)$$

where Φ is the standard normal c.d.f., and $a_+ = \max(a, 0)$. Calculating p-values requires estimating σ_ε^2 and choosing a suitable tuning parameter h. We can estimate σ_ε^2 using any consistent estimator, such as the scaled lasso (Sun and Zhang, 2012). In the simulation experiments and real data example, we choose h using 10-fold cross-validation (CV).

Note that, when simultaneously testing multiple hypotheses: $H_0 : \beta_j^* = 0$ for any $j \in J \subseteq \{1, \ldots, p\}$ versus $H_a : \beta_j^* \neq 0$ for some $j \in J$, we may wish to control the false discovery rate (FDR). Because covariates in the data could be correlated, test statistics on multiple covariates may show arbitrary dependency structure. We thus suggest controlling the FDR using the procedure of Benjamini and Yekutieli (2001). Alternatively, we can control the family-wise error rate (FWER) using, e.g. the method of Holm.

3. Power of the Grace Test

In this section, we investigate power properties of the Grace test. Our first result describes sufficient conditions for detection of nonzero coefficients.

Theorem 2—*Assume Assumptions* **A0** – **A4** *are met. If for some h, some* $0 < \alpha < 1$, $0 < \psi < 1$*, conditional on* $\boldsymbol{X}$*, we have*

$$|\beta_j^*| > 2\Gamma_j^G + q_{(1-\alpha/2)}\sqrt{\mathrm{Var}(Z_j^G|\boldsymbol{X})} + q_{(1-\psi/2)}, \quad (13)$$

where $\Phi(q_{(1-\alpha/2)}) = 1-\alpha/2$*. Then using the same tuning parameter h in the Grace test, we get* $\lim_{n\to\infty} Pr\,(P_j^G \leq \alpha|\boldsymbol{X}) \geq \psi$.

Having established the sufficient conditions for detection of non-null hypotheses in Theorem 2, we next turn to comparing the power of the Grace test with its competitors: the Grace test, the ridge test with small tuning parameters $h_2 = \mathcal{O}(1)$ and no bias correction, and the GraceI test, which is the Grace test with identity penalty weight matrix $\boldsymbol{I}$. The ridge test may be considered as a variant of the test proposed in Bühlmann (2013) without the adjustment of the projection bias – because we assume the design matrix is random, we incur no projection bias in the estimation procedure.

As indicated in Lemma 1, the estimation bias of the Grace procedure depends on the informativeness of the penalty weight matrix $\boldsymbol{L}$. When $\boldsymbol{L}$ is informative, we are able to increase the size of the tuning parameter, which shrinks the estimation variance without

inducing a large estimation bias. Thus, with an informative $\boldsymbol{L}$, we are able to obtain a better prediction performance, as shown empirically in Li and Li (2008); Slawski et al. (2010); Li and Li (2010). In such setting, the larger value of the tuning parameter, e.g. as chosen by CV, also results in improved testing power, as discussed next.

Theorem 3 compares the power of the Grace test to its competitors in a simple setting of $p = 2$ predictors, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. In particular, this result identifies sufficient conditions under which the Grace test has asymptotically superior power. It also gives conditions for the GraceI test to have higher power than the ridge test. The setting of $p = 2$ predictors is considered mainly for ease of calculations, as in this case, we can directly derive closed form expressions of the corresponding test statistics. Similar results are expected to hold for $p > 2$ predictors, but require additional derivations and notations.

Assume $\boldsymbol{y}=\boldsymbol{x}_1\beta_1^*+\boldsymbol{x}_2\beta_2^*+\boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}\sim N_2(\boldsymbol{0},\sigma_\varepsilon^2\boldsymbol{I})$, and $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ are scaled. Denote

$$\boldsymbol{L}\triangleq\begin{pmatrix}1 & l\\ l & 1\end{pmatrix},\quad \hat{\boldsymbol{\Sigma}}\triangleq\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}=\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}.$$

Theorem 3 considers the power for testing the null hypothesis $H_0 : \beta_1^*=0$, in settings where $\beta_1^* \neq 0$, without any constraints on β_2^*.

Theorem 3—*Suppose Assumptions* **A0** – **A4** *are met. Let* $P_j^G(h_n^G)$, $P_j^{GI}(h_n^{GI})$ *and* P_j^R *be the Grace, GraceI and ridge p-values, respectively, with tuning parameters* h_n^G *for Grace and* h_n^{GI} *for GraceI. Define*

$$\Upsilon_{p,n}(h,l,p,|\beta_1|)\triangleq\frac{[(h/n+1)^2-(\rho+lh/n)^2]\cdot|\beta_1|-[\log p/n]^{1/2-\xi}\cdot|(l-\rho)h/n|}{\sqrt{(1+2h/n)(1-\rho^2)+(h/n)^2(1+l^2-2l\rho)}}. \quad (14)$$

Then, conditional on the design matrix $\boldsymbol{X}$*, under the alternative hypothesis* $\beta_1^*=b\neq 0$*, the following statements hold with probability tending to 1, as* $n\to\infty$.

a. *If* $\lim_{n\to\infty}\Upsilon_{p,n}(h_n^G,l,\rho,|b|)\geq\lim_{n\to\infty}\Upsilon_{p,n}(h_n^{GI},0,\rho,|b|)$*, then* $\lim_{n\to\infty}[P_1^G(h_n^G)/P_1^{GI}(h_n^{GI})]\leq 1$.

b. *If* $\lim_{n\to\infty}\Upsilon_{p,n}(h_n^G,l,\rho,|b|)\geq\sqrt{1-\rho^2}|b|$*, then* $\lim_{n\to\infty}[P_1^G(h_n^G)/P_1^R]\leq 1$.

c. *If* $\lim_{n\to\infty}\Upsilon_{p,n}(h_n^{GI},0,\rho,|b|)\geq\sqrt{1-\rho^2}|b|$*, then* $\lim_{n\to\infty}[P_1^{GI}(h_n^{GI})/P_1^R]\leq 1$.

Theorem 3 indicates that, as h_n^G/n and h_n^{GI}/n diverge to infinity, both $\Upsilon_{p,n}(h_n^G,l,\rho,|\beta_1^*|)$ and $\Upsilon_{p,n}(h_n^{GI},0,\rho,|\beta_1^*|)$ approach infinity. This implies, on one hand, that for h_n^G and h_n^{GI} sufficiently large, both the Grace and GraceI tests are asymptotically more powerful than the

ridge test. On the other hand, we can only compare the powers of the Grace and GraceI tests under some constraints on their tuning parameters. With equal tuning parameters for Grace and GraceI, $h_n^G = h_n^{GI}$, we can show, after some algebra, that as $h_n^G/n = h_n^{GI}/n \to \infty$, we have $\lim_{n\to\infty} \Upsilon_{p,n}(h_n^G, l, \rho, |\beta_1^*|) \geq \lim_{n\to\infty} \Upsilon_{p,n}(h_n^{GI}, 0, \rho, |\beta_1^*|)$ if $(1-l^2) \geq \sqrt{(1+l^2-2l\rho)}$. In this case, the Grace test is more powerful than the GraceI test if l is between 0 and l^*, where l^* is the unique root in [−1, 1] of the cubic equation $l^3 - 3l + 2\rho = 0$. Figure 1(a) compares the powers of the Grace and GraceI tests with equal tuning parameters $h_n^G/n = h_n^{GI}/n = 10$ and $\beta_1^* = 1$. It can be seen that, the Grace test asymptotically outperforms the GraceI test when l is close to ρ with equally large tuning parameters. However, when l is far from ρ, the GraceI test could be more powerful. This observation, and the empirical results in Section 5 motivate the development of the GraceR test, introduced in Section 4.

A similar comparison for powers of the Grace and the ridge test, with $h_n^G/n = 10$ and $\beta_1^* = 1$, is provided in Figure 1(b). These results suggest that, with large Grace tuning parameters, Grace substantially outperforms the ridge test in almost all scenarios. The result for the Grace and ridge comparison is similar with $h_n^G/n = 1$.

4. The Grace-Ridge (GraceR) Test

As discussed in Section 2, an informative $\boldsymbol{L}$ results in reduced bias of the Grace procedure, by choosing a larger tuning parameter h. The result in Theorem 3 goes beyond just the bias of the Grace procedure. It shows that for certain choices of $\boldsymbol{L}$, i.e. when l is close to the true correlation parameter ρ, the Grace test can have asymptotically superior power. This additional insight is obtained by accounting for, not just the bias of the Grace procedure, but also its variance, when investigating the power.

However, in practice, there is no guarantee that existing network information truly corresponds to similarities among coefficients, or is complete and accurate. To address this issue, we introduce the Grace-ridge (GraceR) test. The estimator used in GraceR incorporates two Grace-type penalties induced by $\boldsymbol{L}$ and $\boldsymbol{I}$:

$$\hat{\boldsymbol{\beta}}(h_G, h_2) = \arg\min_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + h_G \boldsymbol{\beta}^\top \boldsymbol{L}\boldsymbol{\beta} + h_2 \boldsymbol{\beta}^\top \boldsymbol{\beta} \right\} = (n\hat{\boldsymbol{\Sigma}} + h_G \boldsymbol{L} + h_2 \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{y}. \quad (15)$$

Using data-adaptive choices of tuning parameters h_G and h_2, we expect this test to be as powerful as the Grace test if $\boldsymbol{L}$ is informative, and as powerful as the GraceI test, otherwise.

Another advantage of the GraceR over the Grace test is improved bias-variance tradeoff. If $\boldsymbol{L}$ is (almost) singular, the variance of the Grace test statistic, which depends on the eigenvalues of $(n\hat{\boldsymbol{\Sigma}} + h\boldsymbol{L})$, could be large even for reasonably large h. Thus, even though our discussion in Section 2.1 shows that $(n\hat{\boldsymbol{\Sigma}} + h\boldsymbol{L})$ is almost surely invertible, with finite samples, its smallest eigenvalue could be very small, if not zero. If $\boldsymbol{L}$ is informative, $\boldsymbol{L}\boldsymbol{\beta}$ and hence the bias in (4) are small. Thus, the rank-deficiency of $(n\hat{\boldsymbol{\Sigma}} + h\boldsymbol{L})$ can be alleviated by

choosing a large value of h. However, if $\boldsymbol{L}\boldsymbol{\beta}$ is non-negligible, choosing a large value of h may result in a large bias, even larger than the ridge estimate. to the extent which may offset the benefit from the variance reduction. The finite sample type-I error rate of the Grace test may thus be controlled poorly. By incorporating an additional ℓ_2 penalty, we can better control the eigenvalues and achieve a better bias-variance trade-off.

The GraceR optimization problem leads to the following test statistic:

$$\hat{\boldsymbol{z}}^{GR}=\hat{\boldsymbol{\beta}}(h_G,h_2)+(n\hat{\boldsymbol{\Sigma}}+h_G\boldsymbol{L}+h_2\boldsymbol{I})^{-1}(h_G\boldsymbol{L}+h_2\boldsymbol{I})\tilde{\boldsymbol{\beta}}. \quad (16)$$

Similar to Section 2.2, we can write

$$\hat{z}_j^{GR}=\beta_j^*+Z_j^{GR}+\gamma_j^{GR}, \quad j=1,\ldots,p, \quad (17)$$

where

$$\begin{aligned} Z_j^{GR}|\boldsymbol{X} \quad &\sim N\left(0, n\sigma_\varepsilon^2\left[(n\hat{\boldsymbol{\Sigma}}+h_G\boldsymbol{L}+h_2\boldsymbol{I})^{-1}\hat{\boldsymbol{\Sigma}}(n\hat{\boldsymbol{\Sigma}}+h_G\boldsymbol{L}+h_2\boldsymbol{I})^{-1}\right]_{(j,j)}\right), \\ \boldsymbol{\gamma}^{GR} \quad &\triangleq (n\hat{\boldsymbol{\Sigma}}+h_G\boldsymbol{L}+h_2\boldsymbol{I})^{-1}(h_G\boldsymbol{L}+h_2\boldsymbol{I})(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*). \end{aligned}$$

Similar to the Grace test in in Section 2.2, we choose $\tilde{\boldsymbol{\beta}}$ to be an initial lasso estimator, and derive an asymptotic stochastic bound for γ_j^{GR} such that $|\gamma_j^{GR}|\lesssim^{asy.}\Gamma_j^{GR}$. Equation (12) is again used to obtain two-sided p-values for H_0. Theorems 4 and 5 parallel the previous results for the Grace test, and establish GraceR's asymptotic control of type-I error rate, and conditions for detection of non-null hypotheses. Proofs of these results are similar to Theorems 1 and 2, and are hence omitted. We first state an alternative to Assumption **A4**. This assumption can be justified using an argument similar to that for Assumption **A4**, and can also be relaxed with the cost of reduced power for the GraceR test.

- **A4′**: h_G, h_2 and $\boldsymbol{L}$ are such that

$$\left[(n\hat{\boldsymbol{\Sigma}}+h_G\boldsymbol{L}+h_2\boldsymbol{I})^{-1}(h_G\boldsymbol{L}+h_2\boldsymbol{I})\right]_{(j,j)}=\mathcal{O}_p\left(\left[\frac{n}{\log p}\right]^{\frac{1}{2}-\xi}\right).$$

Theorem 4

Assume Assumptions **A1** – **A3** *and* **A4′** *are met. The following* Γ_j^{GR} *satisfies the stochastic bound for GraceR.*

$$\Gamma_j^{GR} \triangleq \left\| \left[(n\hat{\boldsymbol{\Sigma}} + h_G \boldsymbol{L} + h_2 \boldsymbol{I})^{-1} (h_G \boldsymbol{L} + h_2 \boldsymbol{I}) \right]_{(j,-j)} \right\|_\infty \left(\frac{\log p}{n} \right)^{\frac{1}{2}-\xi}. \quad (18)$$

Then, under the null hypothesis, for any $a > 0$,

$$\lim_{n\to\infty} Pr\left(|\hat{\boldsymbol{z}}_j^{GR}| > \alpha\right) \leq \lim_{n\to\infty} Pr\left(|Z_j^{GR}| + \Gamma_j^{GR} > \alpha\right). \quad (19)$$

Theorem 5

Assume Assumptions **A1** – **A3** *and* **A4′** *are met. If for some* $h_G > 0$ *and* $h_2 > 0$, *conditional on* $\boldsymbol{X}$, *we have*

$$|\beta_j^*| > 2\Gamma_j^{GR} + q_{(1-\alpha/2)} \sqrt{\mathrm{Var}(Z_j^{GR}|\boldsymbol{X})} + q_{(1-\psi/2)} \quad (20)$$

for some $0 < \alpha < 1$ *and* $0 < \psi < 1$. *Then using the same* h_G *and* h_2 *in the GraceR test, we get* $\lim_{n\to\infty} Pr\left(P_j^{GR} \leq \alpha | \boldsymbol{X}\right) \geq \psi$.

5. Simulation Experiments

In this section, we compare the Grace and GraceR tests with the ridge test (Bühlmann, 2013) with small tuning parameters, low-dimensional projection estimator (LDPE) for inference (Zhang and Zhang, 2014; van de Geer et al., 2014) and the GraceI test. To this end, we consider a graph similar to Li and Li (2008), with 50 hub covariates (genes), each connected to 9 other satellite covariates (genes). The 9 satellite covariates are not connected with each other, nor are covariates in different hub-satellite clusters. In total the graph includes $p = 500$ covariates and 450 edges; see Figure S1 in the online Supplementary Material for an illustration with 5 hub-satellite clusters. We build the underlying true Laplacian matrix $\boldsymbol{L}^*$ according to the graph with all edge weights equal 1.

To assess the effect of inaccurate or incomplete network information, we also consider variants of the Grace and GraceR tests with incorrectly specified graphs, where a number of randomly selected edges are added or removed. The number of removed or added (perturbed) edges relative to the true graph is NPE $\in \{-165, -70, -10, 0, 15, 135, 350\}$, with negative and positive numbers indicating removals and additions of edges, respectively. For example, NPE = −165 indicates 165 of the 450 edges in the true graph represented by $\boldsymbol{L}^*$ are randomly removed in the perturbed graph with corresponding perturbed Laplacian matrix $\boldsymbol{L}$. This represents the case with incomplete network information. On the other hands, NPE = 350 indicates that in addition to the 450 true edges in $\boldsymbol{L}^*$, we also randomly add 350 wrong edges to $\boldsymbol{L}$. The NPE values considered correspond to similar normalized spectral differences for settings where edges are removed or added, i.e. $\|\boldsymbol{L} - \boldsymbol{L}^*\|_2 / \|\boldsymbol{L}^*\|_2 \approx (0.75,$

0.50, 0.25, 0, 0.25, 0.50, 0.75). Thus, the size of perturbation to the graph is roughly the same with NPE = −165 and 350. The perturbed penalty weight matrix $\boldsymbol{L}$ is then used in the Grace and GraceR tests. Since ($\boldsymbol{X}^\top \boldsymbol{X} + h\boldsymbol{L}$) may not be invertible, for Grace, we add a value of 0.01 to the diagonal entries of $\boldsymbol{L}$ to make it positive definite. No such correction is needed for GraceR and GraceI because of the ℓ_2 penalty.

In each simulation replicate, we generate $n = 100$ independent samples, where for the 50 hub covariates in each sample, $x_k^{hub} \sim^{iid} N(0,1)$, $k = 1, \ldots, 50$, and for the 9 satellite covariates in the k-th hub-satellite cluster, $x_l^{hub_k} \sim^{iid} N(0.9 \times x_k^{hub}, 0.9)$, $l = 1, \ldots, 9$, $k = 1, \ldots, 50$. This is equivalent to simulating $\boldsymbol{x}^i \sim^{iid} \boldsymbol{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ for $i = 1, \ldots, 100$ with $\boldsymbol{\Sigma} = (\boldsymbol{L}^* + 0.11 \times \boldsymbol{I})^{-1}$, where $\boldsymbol{L}^*$ corresponds to the partial covariance structure of the covariates.

We consider a sparse model in which covariates in the first hub-satellite cluster are equally associated with the outcome, and those in the other 49 clusters are not. Specifically, we let

$$\boldsymbol{\beta}^* \triangleq \frac{1}{\sqrt{10}}(\underbrace{1, \ldots, 1}_{10}, \underbrace{0, \ldots, 0}_{p-10})^\top.$$

We then simulate $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$; with $\boldsymbol{\varepsilon} \sim N_n(\boldsymbol{0}, \sigma_\varepsilon^2 \boldsymbol{I}_n)$, and consider $\sigma_{\boldsymbol{\varepsilon}} \in \{9.5, 6.3, 4.8\}$ to produce expected $R^2 = 1 - \sigma_\varepsilon^2 / \mathrm{Var}(\boldsymbol{y}) \in \{0.1, 0.2, 0.3\}$.

Throughout the simulation iterations, $\boldsymbol{L}^*$ and $\boldsymbol{\beta}^*$ are kept fixed, and $\boldsymbol{L}$, $\boldsymbol{X}$ and $\boldsymbol{\varepsilon}$ are randomly generated in each repetition. We set the sparsity parameter $\xi = 0.05$, and $h_{Lasso} = 4\hat{\sigma}_\varepsilon \sqrt{3 \log p / n}$, where $\hat{\sigma_{\boldsymbol{\varepsilon}}}$ is calculated using the scaled lasso (Sun and Zhang, 2012). As suggested in Bühlmann (2013), the tuning parameter for the ridge test is set to 1. Tuning parameters for LDPE, Grace, GraceR and GraceI are chosen by 10-fold CV. We use two-sided significance level $\alpha = 0.05$ and calculate the average and standard error of powers from 10 non-zero coefficients and the type-I error rates of each test from 490 zero coefficients. Figure 2 summarizes the mean powers and type-I error rates of tests across $B = 100$ simulated data sets, along with the corresponding 95% confidence intervals. Detail values of powers and type-I error rates, as well as an expanded simulation with a larger range of NPE, are available in Table S2, S3 and Figure S4 in the online Supplementary Material.

Comparing the power of the tests, it can be seen that the Grace test with correct choices of $\boldsymbol{L}$ (NPE = 0) results in highest power. The performance of the Grace test, however, deteriorates as $\boldsymbol{L}$ becomes less accurate. The performance of the GraceR test is, on the other hand, more stable. It is close to the Grace test when the observed $\boldsymbol{L}$ is close to the truth, and is roughly as good as the GraceI test when $\boldsymbol{L}$ is significantly inaccurate. As expected, our testing procedures asymptotically control the type-I error rate, in that observed type-I error rates are not significantly different from $\alpha = 0.05$.

6. Analysis of TCGA Prostate Cancer Data

We examine the Grace and GraceR tests on a prostate adenocarcinoma dataset from The Cancer Genome Atlas (TCGA) collected from prostate tumor biopsies. After removing samples with missing measurements, we obtain a dataset with $n = 321$ samples. For each sample, the prostate-specific antigen (PSA) level and the RNA sequences of 4739 genes are available. Genetic network information for these genes is obtained from the Kyoto Encyclopedia of Genes and Genomes (KEGG), resulting in a dataset with $p = 3450$ genes and $|E| = 38541$ edges.

We center the outcome and center and scale the covariates. For the Grace and GraceR tests, we set the sparsity parameter $\xi = 0.05$ and $h_{Lasso} = 4\hat{\sigma}_{\varepsilon}\sqrt{3\log p/n}$, where $\hat{\sigma_{\varepsilon}}$ is calculated using the scaled lasso (Sun and Zhang, 2012). We control the false discovery rate at FDR = 0.05 level using the method of Benjamini and Yekutieli (2001).

To increase the chance of selecting "hub" genes, we use the normalized Laplacian matrix $\boldsymbol{L}^{(norm)} = \mathbf{D}^{-1/2}\boldsymbol{L}\mathbf{D}^{-1/2}$, where $\mathbf{D}$ is the diagonal degree matrix for the KEGG network with edge weights set to 1. The Grace penalty induced by the normalized Laplacian matrix encourages smoothness of coefficient estimates based on the degrees of respective nodes, $\boldsymbol{\beta}^{\top}\boldsymbol{L}^{(norm)}\boldsymbol{\beta} = \Sigma_{u\sim v}(\boldsymbol{\beta}_u/\sqrt{d_u} - \boldsymbol{\beta}_v/\sqrt{d_v})^2 w(u,v)$ (Li and Li, 2008). We add 0.001 to the diagonal entries of $\boldsymbol{L}^{(norm)}$ to induce positive definitiveness in the Grace test.

As shown in Figure 3(a), the Grace test with tuning parameter selected by 10-fold CV identifies 54 genes that are associated with PSA level. They consist of 42 histone genes, 11 histone deacetylase (HDAC) genes and the paired box gene 8 (PAX8). Histone and HDAC genes are densely connected in the KEGG network. With the network smoothing penalty, the Grace regression coefficients of histone and HDAC genes are all positive with a similar magnitude. Existing literature indicates that the histone and HDAC genes are associated with the occurrence, progression, clinical outcomes or recurrence of prostate cancer. Figure 3(b) shows the result for the GraceR test. GraceR identifies 5 histone genes, which are also identified by the Grace test. In addition, GraceR identifies 11 genes that are not identified by Grace. Prior work has identified 9 of those 11 genes to be associated with PSA level or the severity and stage of cancer. Additional details about existing evidence in support of genes identified using Grace and GraceR tests, as well as extended results on prediction performance and stability of the Grace test are provided in Section S6 in the online Supplementary Material.

As a comparison, the GraceI test with 10-fold CV identifies 16 disconnected genes, 11 of them are also identified by the GraceR test. Ridge test (Bühlmann, 2013) with tuning parameter $h_2 = 1$ identifies 4 disconnected genes, which are also identified by the GraceR test. The low-dimensional projection estimator (LDPE) with tuning parameters chosen by 10-fold CV identifies 10 disconnected genes. Seven of these genes are identified by GraceR and two by Grace.

7. Discussion

In this paper, we proposed the Grace and GraceR tests that incorporate external graphical information regarding the similarity between covariates. Such external information is presented in the form of a penalty weight matrix $\boldsymbol{L}$, which is considered to be the (normalized) graph Laplacian matrix in this paper. However, any positive semi-definite matrix can be used as $\boldsymbol{L}$. The proposed inference framework thus allows researchers in different fields to incorporate relevant external information through $\boldsymbol{L}$. For example, we can use various distance and kernel metrics that measure the (dis)similarity between species in phylogenetic studies. We can also use the adaptive graph Laplacian matrix (Li and Li, 2010) so that coefficients of negatively correlated covariates are penalized to have the opposite signs. Regardless of the choice of $\boldsymbol{L}$, our proposed procedures asymptotically control the type-I error rate; the power of the Grace test, however, depends on the informativeness of $\boldsymbol{L}$. The power of the GraceR test is on the other hand less dependent on the choice of $\boldsymbol{L}$.

The Grace test introduced in this paper is not scale invariant. That is, the Grace test with the same tuning parameter could produce different p-values with data ($\boldsymbol{X}$, $\boldsymbol{y}$) and ($\boldsymbol{X}$, $k\boldsymbol{y}$), where $k \neq 1$ is a constant. This is clear as the test statistic $\hat{z}_j$ depends on $\boldsymbol{y}$ whereas the stochastic bound Γ_j^G does not. To make the Grace and GraceR tests scale invariant, we can simply choose the tuning parameter for our lasso initial estimator to be $h_{Lasso} = C\sigma_\varepsilon \sqrt{\log p/n}$ with a constant $C > 2\sqrt{2}$. Sun and Zhang (2012) show that the lasso is scale invariant in this case. We would also need to use scaled invariant stochastic bounds $\tilde{\Gamma}_j^G \triangleq \sigma_\varepsilon \Gamma_j^G$ and $\tilde{\Gamma}_j^{GR} \triangleq \sigma_\varepsilon \Gamma_j^{GR}$ in our Grace and GraceR tests. Note that multiplying any constant in Γ_j^G and Γ_j^{GR} does not change our asymptotic control of the type-I error rate.

In this paper, cross validation (CV) is used to choose tuning parameters of the Grace and GraceR tests. However, CV does not directly maximize the power of these tests. Selection of tuning parameters for optimal testing performance can be a fruitful direction of future research. Another useful extension of the proposed framework is its adaptation to generalized linear models (GLM).

The Grace and GraceR tests are implemented in the R package **Grace**, available on the Comprehensive R Archive Network (CRAN).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank Dr. Ruben Dezeure of the Seminar for Statistics of the Department of Mathematics at ETH Zürich for providing the code for LDPE. We thank the Associated Editor and two referees for helpful comments that lead to substantial improvements on our manuscript. This work was partially supported by grants from the U.S. National Institute of Health and National Science Foundation.

References

Bai Z. Methodologies in spectral analysis of large dimensional random matrices: A review. Statistica Sinica. 1999; 9:611–677.

Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. The Annals of Statistics. 2001; 29:1165–1188.

Bickel P, Ritov Y, Tsybakov A. Simultaneous analysis of Lasso and Dantzig selector. The Annals of Statistics. 2009; 37:1705–1732.

Bühlmann P. Statistical significance in high-dimensional linear models. Bernoulli. 2013; 19:1212–1242.

Bühlmann, P.; van de Geer, S. Statistics for High-dimensional Data: Methods, Theory and Applications. Springer Series in Statistics. Springer; 2011.

Chung, FR. Spectral graph theory. Vol. 92. American Mathematical Soc; 1997.

de la Fuente A, Bing N, Hoeschele I, Mendes P. Discovery of meaningful associations in genomic data using partial correlation coefficients. Bioinformatics. 2004; 20:3565–3574. [PubMed: 15284096]

Fukuyama J, McMurdie PJ, Dethlefsen L, Relman DA, Holmes S. Comparisons of distance methods for combining covariates and abundances in microbiome studies. Pacific Symposium on Biocomputing. 2012:213–224. [PubMed: 22174277]

Hart GT, Ramani AK, Marcotte EM. How complete are current yeast and human protein-interaction networks? Genome Biology. 2006; 7:120. [PubMed: 17147767]

Huang J, Ma S, Li H, Zhang CH. The sparse Laplacian shrinkage estimator for high-dimensional regression. The Annals of Statistics. 2011; 39:2021–2046. [PubMed: 22102764]

Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: Current approaches and outstanding challenges. PLoS Comput Biology. 2012; 8:e1002375.

Kong SW, Pu WT, Park PJ. A multivariate approach for integrating genome-wide expression data and biological knowledge. Bioinformatics. 2006; 22:2373–2380. [PubMed: 16877751]

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

Li C, Li H. Variable selection and regression analysis for graph-structured covariates with an application to genomics. The Annals of Applied Statistics. 2010; 4:1498–1516. [PubMed: 22916087]

Liu D, Lin X, Ghosh D. Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. Biometrics. 2007; 63:1079–1088. [PubMed: 18078480]

Meinshausen N, Bühlmann P. Stability selection. Journal of the Royal Statistical Society: Series B. 2010; 72:417–473.

Michailidis G. Statistical challenges in biological networks. Journal of Computational and Graphical Statistics. 2012; 21:840–855.

Pan W, Xie B, Shen X. Incorporating predictor network in penalized regression with application to microarray data. Biometrics. 2010; 66:474–484. [PubMed: 19645699]

Randolph T, Harezlak J, Feng Z. Structured penalties for functional linear models—partially empirical eigenvectors for regression. Electronic Journals of Statistics. 2012; 6:323–353.

Shao J, Deng X. Estimation in high-dimensional linear models with deterministic design matrices. The Annals of Statistics. 2012; 40:812–831.

Shen X, Huang HC, Pan W. Simultaneous supervised clustering and feature selection over a graph. Biometrika. 2012; 99:899–914. [PubMed: 23843673]

Shojaie A, Michailidis G. Analysis of gene sets based on the underlying regulatory network. Journal of Computational Biology. 2009; 16:407–426. [PubMed: 19254181]

Shojaie A, Michailidis G. Network enrichment analysis in complex experiments. Statistical Applications in Genetics and Molecular Biology. 2010a; 9:Article 22.

Shojaie A, Michailidis G. Penalized principal component regression on graphs for analysis of subnetworks. Advances in Neural Information Processing Systems. 2010b; 23:2155–2163.

Slawski M, zu Castell W, Tutz G. Feature selection guided by structural information. The Annals of Applied Statistics. 2010; 4:1056–1080.

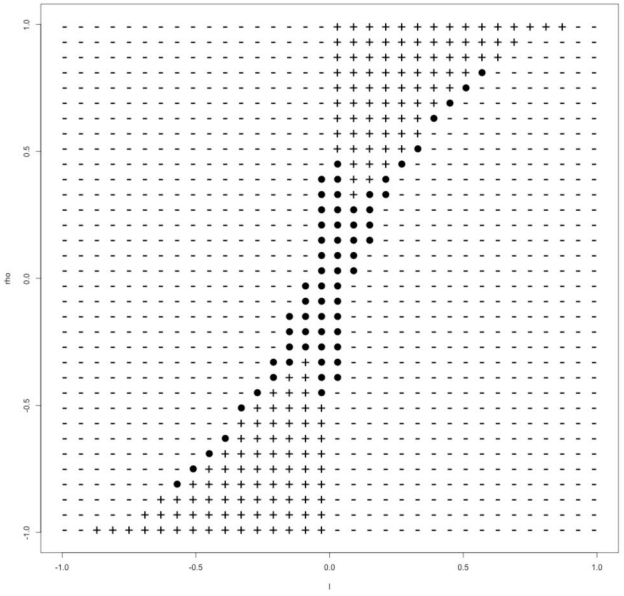
Sun T, Zhang CH. Scaled sparse linear regression. Biometrika. 2012; 99:879–898.

van de Geer S, Bühlmann P, Ritov Y, Dezeure R. On asymptotically optimal confidence regions and tests for high-dimensional models. The Annals of Statistics. 2014; 42:1166–1202.

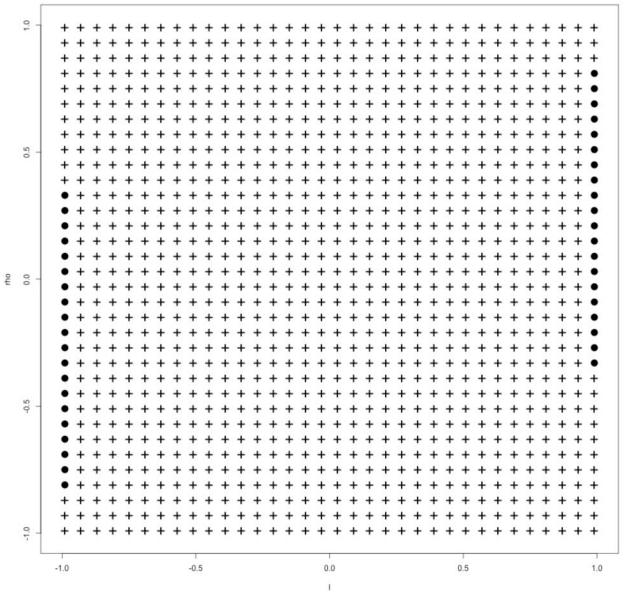
Wei P, Pan W. Incorporating gene networks into statistical tests for genomic data via a spatially correlated mixture model. Bioinformatics. 2008; 24:404–411. [PubMed: 18083717]

Zhang CH, Zhang S. Confidence intervals for low dimensional parameters in high dimensional linear models. Journal of the Royal Statistical Society: Series B. 2014; 76:217–242.

Zhu X, Gerstein M, Snyder M. Getting connected: analysis and principles of biological networks. Genes & Development. 2007; 21:1010–1024. [PubMed: 17473168]



(a) Grace versus GraceI



(b) Grace versus ridge

Figure 1.

(a) The ratio of $\Upsilon_{p,n}(h_n^G, l, \rho, |\beta_1^*|)$ over $\Upsilon_{p,n}(h_n^{GI}, 0, \rho, |\beta_1^*|)$ for different l and ρ with $h_n^G/n = h_n^{GI}/n = 10$, $[\log p/n]^{1/2-\xi} = 0.25$ and $\beta_1^* = 1$. A plus sign indicates the ratio is greater than 1.02, whereas a minus sign indicates the ratio is smaller than 0.98; filled circles indicate an intermediate value. (b) The log-ratio of $\Upsilon_{p,n}(h_n^G, l, \rho, |\beta_1|)$ over $\sqrt{1-\rho^2}$ for different l and ρ with $h_n^G/n = 10$, $[\log p/n]^{1/2-\xi} = 0.25$ and $\beta_1^* = 1$. A plus sign indicates the log-ratio is greater than 0.5 (ratio > 1.65), whereas a minus sign indicates the log-ratio is smaller than −0.5 (ratio < 0.61); filled circles indicate an intermediate value

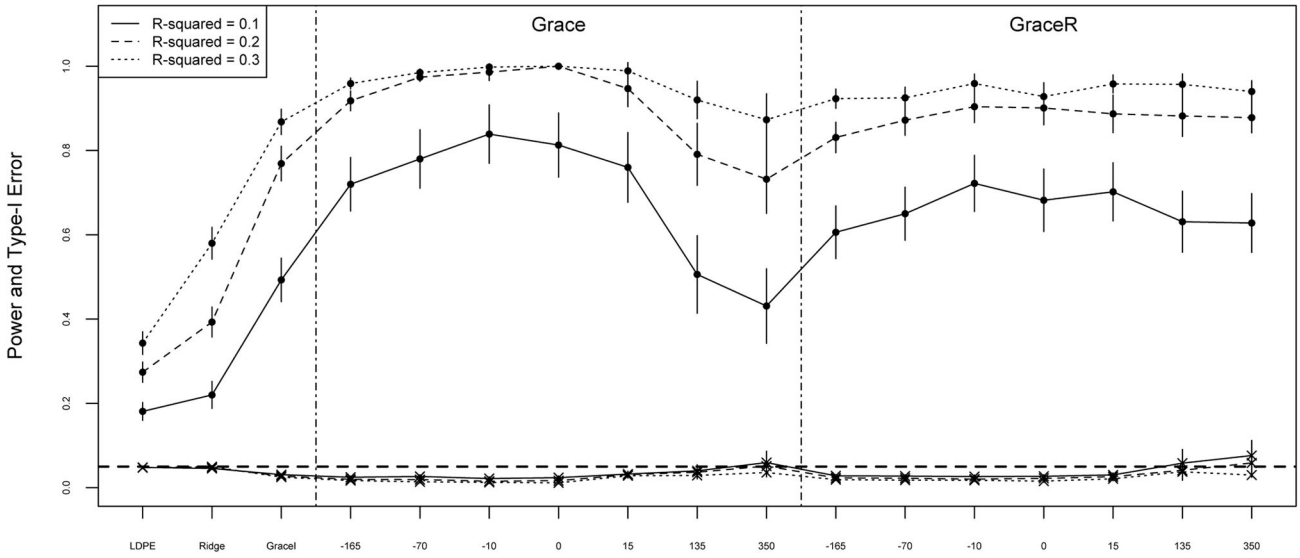
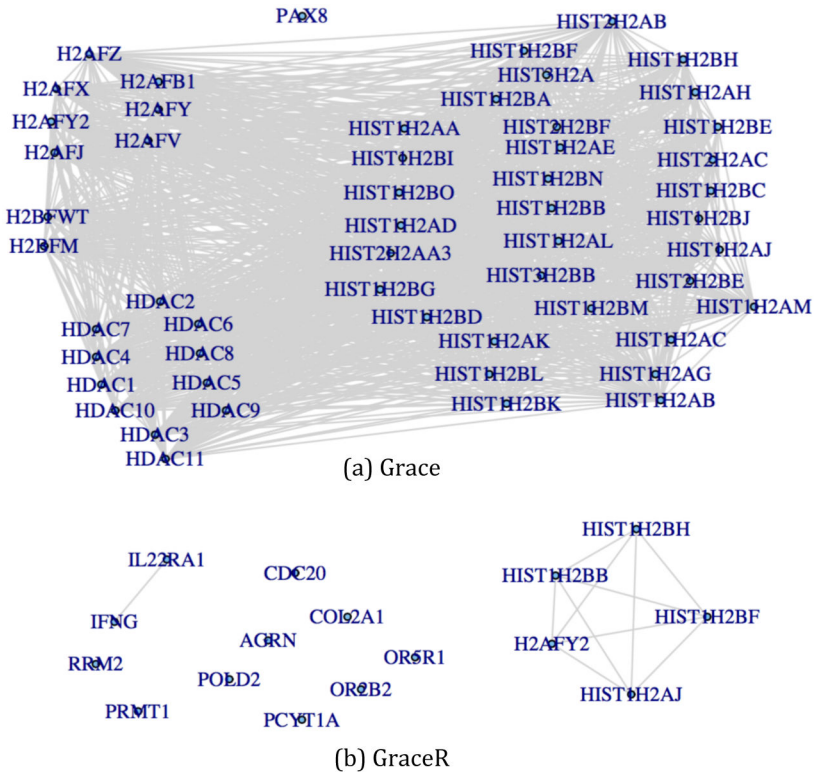


Figure 2.

Comparison of powers and type-I error rates of different testing methods, along with their 95% confidence bands. Testing methods include LDPE (Zhang and Zhang, 2014; van de Geer et al., 2014), ridge (Bühlmann, 2013), GraceI, Grace and GraceR tests. Filled circles (●) corresponds to powers, whereas crosses (×) are type-I error rates. Numbers on *x*-axis for Grace and GraceR tests refer to the number of perturbed edges (NPE) in the network used for testing, compared to the true network used to generate the data.



(a) Grace

(b) GraceR

Figure 3.

Results of analysis of TCGA prostate cancer data using the (a) *Grace* and (b) *GraceR* tests after adjusting for FDR at 0.05 level. In each case, genes found to be significantly associated with PSA level are shown, along with their interactions based on information from KEGG.