METHOD ARTICLE

REVISED *Space-log*: a novel approach to inferring gene-gene networks using SPACE model with log penalty [version 2; peer review: 2 approved, 1 approved with reservations]

Qian (Vicky) Wu [1,2], Wei Sun [2], Li Hsu [2]

[1] Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, WA, 98109, USA
[2] Public Health Division, Fred Hutchinson Cancer Research Center, Seattle, WA, 98109, USA

v2 **First published:** 21 Sep 2020, **9**:1159
https://doi.org/10.12688/f1000research.26128.1

Latest published: 05 Jan 2022, **9**:1159
https://doi.org/10.12688/f1000research.26128.2

Abstract

Gene expression data have been used to infer gene-gene networks (GGN) where an edge between two genes implies the conditional dependence of these two genes given all the other genes. Such gene-gene networks are of-ten referred to as gene regulatory networks since it may reveal expression regulation. Most of existing methods for identifying GGN employ penalized regression with *L1* (lasso), *L2* (ridge), or elastic net penalty, which spans the range of *L1* to *L2* penalty. However, for high dimensional gene expression data, a penalty that spans the range of *L0* and *L1* penalty, such as the log penalty, is often needed for variable selection consistency. Thus, we develop a novel method that em-ploys log penalty within the framework of an earlier network identification method space (Sparse PArtial Correlation Estimation), and implement it into a R package *space-log*. We show that the *space-log* is computationally efficient (source code implemented in C), and has good performance comparing with other methods, particularly for networks with hubs. *Space-log* is open source and available at GitHub, https://github.com/wuqian77/SpaceLog

Keywords

Gene-gene network, gene regulation, penalized regression, log penalty, partial correlation, R package, algorithm

This article is included in the RPackage gateway.

Open Peer Review

Reviewer Status ✓ ? ✓

Invited Reviewers	1	2	3
version 2 (revision) 05 Jan 2022	✓ report		✓ report
version 1 21 Sep 2020	? report	? report	? report

1. **Yuying Xie**, Michigan State University, East Lansing, USA
   Yuning Hao, Amazon, Seattle, USA
2. **Chi Song**, The Ohio State University, Columbus, USA
3. **Jessie Jeng**, North Carolina State University, Raleigh, USA

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding authors: Qian (Vicky) Wu (qwu@fredhutch.org), Li Hsu (lih@fredhutch.org)

Author roles: Wu Q(: Conceptualization, Data Curation, Formal Analysis, Methodology, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Sun W**: Conceptualization, Data Curation, Methodology, Software, Writing – Review & Editing; **Hsu L**: Conceptualization, Funding Acquisition, Methodology, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This research is supported by National Institutes of Health grant R01 CA189532.

How to cite this article: Wu Q(, Sun W and Hsu L. ***Space-log*: a novel approach to inferring gene-gene net-works using SPACE model with log penalty [version 2; peer review: 2 approved, 1 approved with reservations]** F1000Research 2022, **9**:1159 https://doi.org/10.12688/f1000research.26128.2

First published: 21 Sep 2020, **9**:1159 https://doi.org/10.12688/f1000research.26128.1

REVISED **Amendments from Version 1**

Updated background, and added discussion on run-time with a new table. Methods and references have also been revised.

Any further responses from the reviewers can be found at the end of the article

Introduction

Complex diseases, such as colorectal cancer (CRC), are caused by a combination of genetic, environmental and lifestyle factors, most of which have not yet been identified and explained. In the era of precision medicine, human genome project, especially breakthroughs in high throughput technologies, provides information based on patient's genetic and genomic data and have changed how researchers explain complex disease through personalized profiles. There is an urgent need to develop efficient statistical and computational tools to integrating genetic and genomic data to identify their contributions to complex diseases, such as gene expression and genetic regulatory networks. Analysis of gene expression data has led to the identification of novel disease-causing gene networks[1], thus contributing important new insights into understanding of these complex diseases. The objective of this paper is to introduce a novel method that constructs gene-gene network (GGN) based on high dimensional gene expression data. Popular methods for GGN include neighborhood selection[2], graphical Lasso[3], and space (Sparse PArtial Correlation Estimation)[4]. Neighborhood selection consistently estimates the non-zero entries of the partial correlation matrix, and provide an approximation of the maximum likelihood estimate of partial correlation matrix. Graphical Lasso improves on neighborhood selection by providing a maximum likelihood estimate of the partial correlation matrix. The space method exploits the symmetry of partial correlation matrix to improve the estimation accuracy. It also avoids potential conflicts in neighborhood selection, that is, Y_i is selected as a neighbor of Y_j but Y_j is not selected as a neighbor of Y_i, and one has to make a post-hoc decision for whether Y_i and Y_j are connected. Furthermore, those available methods employ L_1, L_2 or elastic net penalty. However, penalties in the range of L_0 to L_1 is often needed to improve the accuracy of variable selection for high-dimensional gene expression data[5]. In this paper, we propose a new statistical method to estimate GGN by implementing the log penalty for the space approach, which enhances sparsity by reweighted L_1 minimization, and we refer to our method as space-log.

We compared space-log with the space algorithm through extensive simulations, as well as the comparison with neighborhood selection methods using lasso or log penalty below. Peng *et al.* (2009)[4] have compared the space approach with gLasso and showed space outperformed gLasso in different simulation settings, thus we didn't include gLasso methods in our simulation studies here.

Methods

Suppose that we have data on n independent individuals and m genes. Assume the expression of m genes, after appropriate normalization, follow a multivariate Gaussian distribution $N(0, \Sigma)$.

Neighborhood selection using lasso or log penalty: NS-lasso, NS-log

The neighborhood selection (NS) approach considers each gene separately. Let Y_i be the gene expression value for the ith gene and $Y_{-i} = (Y_1, \ldots, Y_{i-1}, Y_{i+1}, Y_m)^T$. For the NS approach, Y_i is regressed on Y_{-i} by a penalized regression:

$$\hat{\beta} = argmin\left\{\frac{1}{2}(Y_i - Y_{-i}\beta)^T(Y_i - Y_{-i}\beta) + n\sum_{i\neq j} p(|\beta_{i,j}|;\omega)\right\} \qquad (1)$$

with penalty function $p(|\beta|;\omega)$. We will compare NS-lasso with lasso penalty $p(|\beta|;\lambda) = \lambda|\beta|$[6] and NS-log with log penalty $p(|\beta|;\lambda,\tau) = \lambda\log(|\beta|+\tau)$[7,8]. Source codes of NS-lasso and NS-log are available at https://github.com/Sun-lab/penalized_estimation/.

Joint modeling space using lasso penalty: space-lasso

The joint modeling approach space[3] is to estimate GGN, without the need to fit many (m) single gene regression models separately, but directly estimate partial correlation among all the genes. Denote the partial correlation between Y_i and Y_j by $\rho_{i,j}$. If we know the concentration matrix $\sum^{-1} = (\sigma^{ij})_{m\times m}$, then $\rho_{i,j} = -\frac{\sigma^{ij}}{\sqrt{\sigma^{ii}\sigma^{jj}}}$. Given $\beta_{ij} = -\frac{\sigma^{ij}}{\sigma^{ii}}$, we can easily get that $\rho_{i,j} = sign(\beta_{i,j})\sqrt{\beta_{i,j}\beta_{j,i}}$. Thus, the problem is translated into partial correlation matrix estimation. Specifically[4], proposed to minimize a penalized loss function

$$L_n(\beta,\sigma,Y) = \frac{1}{2}\sum_{i=1}^{m} w_i \left\| Y_i - \sum_{j\neq i}\beta_{ij}Y_j \right\|^2 + \sum_{i\neq j} p\left(\left|\rho_{i,j};\lambda\right|\right) \qquad (2)$$

$$= \frac{1}{2}\sum_{i=1}^{m} w_i \left\| Y_i - \sum_{j\neq i}\rho_{ij}\sqrt{\frac{\sigma^{jj}}{\sigma^{ii}}}Y_j \right\|^2 + \sum_{i\neq j} p\left(\left|\rho_{i,j};\lambda\right|\right) \qquad (3)$$

where $w_i \geq 0$ is the weight, e.g., uniform weights $w_i = 1$ for space-no, residual variance based weights $w_i = \sigma_{jj}$ for space-res, and degree based weights w_i = number of genes that $\{j : \rho_{i,j} \neq 0, j \neq i\}$ for space-df. In Peng *et al.* (2009)[4], $p(|\rho|;\lambda) = \lambda|\rho|$ and we call it as space-lasso.

New algorithm space-log: joint modeling space using log penalty

Inspired by Sun *et al.* (2010)[7] and Ha *et al.* (2016)[8], we extended the space approach with log penalty as space-log $p(|\rho|;\lambda,\tau) = \lambda\log(|\rho|+\tau)$ and used the active shooting algorithm[4] to update the coefficient estimates iteratively in space-log (Supplementary Materials). We determined the tuning parameters by using extended BIC (extBIC)[9].

Denote the target loss function as

$$\begin{aligned} f(\rho;\sigma) &= \frac{1}{2}\sum_{i=1}^{m} w_i \left\| Y_i - \sum_{j\neq i} \rho_{ij}\sqrt{\frac{\sigma^{jj}}{\sigma^{ii}}}Y_j \right\|^2 + \sum_{i\neq j} p\left(\left|\rho_{i,j}\right|;\tau,\lambda\right) \\ &= \frac{1}{2}\sum_{i=1}^{m} w_i \left\| Y_i - \sum_{j\neq i} \rho_{ij}\sqrt{\frac{\sigma^{jj}}{\sigma^{ii}}}Y_j \right\|^2 + \lambda\sum_{i\neq j} log\left(\left|\rho_{i,j}\right| + \tau\right) \end{aligned} \tag{4}$$

The goal is to estimate $\hat{\rho} = argmin_{\rho} f(\rho)$ for a given λ and τ. We implement the penalized estimation using space and Log penalties by Local Linear Approximation (LLA)[10].

$$p\left(\left|\rho_{i,j};\lambda,\tau\right|\right) \approx p\left(\left|\hat{\rho}_{i,j}^{(k)}\right|;\lambda,\tau\right) + p'\left(\left|\hat{\rho}_{i,j}^{(k)}\right|;\lambda,\tau\right)\left(\left|\rho_{i,j}\right| - \left|\hat{\rho}_{i,j}^{(k)}\right|\right) \tag{5}$$

Where $\left|\hat{\rho}_{i,j}^{(k)}\right|$ is the estimate of regression coefficient $\rho_{i,j}$ at the k-th iteration. After applying LLA for the penalty part, we can minimize loss function at the (k + 1)-th step, while solving for $\rho_{i,j}$ by

$$\begin{aligned} f^{(k+1)}(\rho_{i,j}) = \frac{1}{2}\sum_{i=1}^{m} w_i \left\| Y_i - \rho_{i,j}\sqrt{\frac{\sigma^{jj}}{\sigma^{ii}}}Y_j - \sum_{l\neq i;l\neq j} \hat{\rho}_{i,l}^{(k)}\sqrt{\frac{\hat{\sigma}_{(k)}^{ll}}{\hat{\sigma}_{(k)}^{ii}}}Y_l \right\|^2 \\ + \sum_{i\neq j} p'\left(\left|\hat{\rho}_{i,j}^{(k)}\right|;\lambda,\tau\right)\left|\rho_{i,j}\right| \end{aligned} \tag{6}$$

By letting $\partial f^{(k+1)}(\rho_{i,j}) / \partial \rho_{i,j} = 0$, we can find the solution for $\rho_{i,j}$ as follows:

$$\hat{\rho}_{i,j}^{(k+1)} = \begin{cases} 0 & if \left|z_j^{(k)}\right| \leq v_j^{-1} p'\left(\left|\hat{\rho}_{i,j}^{(k)}\right|;\lambda,\tau\right) \\ sgn\left(\hat{\rho}_{i,j}^{(k)}\right)[\left|z_j^{(k)}\right| - v_j^{-1} p'\left(\left|\hat{\rho}_{i,j}^{(k)}\right|;\lambda,\tau\right)] & if \left|z_j^{(k)}\right| > v_j^{-1} p'\left(\left|\hat{\rho}_{i,j}^{(k)}\right|;\lambda,\tau\right) \end{cases} \tag{7}$$

where $z_j^{(k)} = Y_j(Y_i - \sum_{l\neq i;l\neq j} \hat{\rho}_{i,l}^{(k)} \sqrt{\frac{\hat{\sigma}_{(k)}^{ll}}{\hat{\sigma}_{(k)}^{ii}}} Y_l) / V_j, V_j = Y_j^T Y_j$, and

$$p'\left(\left|\hat{\rho}_{i,j}^{(k)}\right|;\lambda,\tau\right) = sgn\left(\hat{\rho}_{i,j}^{(k)}\right)\frac{\lambda}{\left|\hat{\rho}_{i,j}^{(k)}\right| + \tau}.$$

Active-shooting

We adapted the same idea `active-shooting` algorithm from 4 to update the coefficient estimation iteratively in `space-log`. Without loss of generality, we kept most notation from 4 but tailored with `space-log`. The details are included in the *Supplementary Materials*.

Simulation studies

In this section, we present Monte Carlo simulation to evaluate the performance of the `space-log`, `space-lasso`, `NS-log`, and `NS-lasso`. Following[8], we studied two types of graphs: the traditional random graphs (ER model) where all the genes have the same expected number of neighbors[11,12], and hubs graphs where a few genes may have a large number of neighbors (BA model), and BA model is more frequently observed in gene networks[13].

We simulated GGN of m genes under both the BA and ER models, respectively. The initial graph had one gene and no edge. In the (k+1)th step, we added e edges between a new gene and e old genes. Under the BA model, there is a greater probability for the new gene to connect to an existing hub gene that has larger number of edges with the probability $p_E = v_i^{(t)} / \sum_j v_j^{(t)}$, where $v_i^{(t)}$ number of edges connected with the ith gene at the tth step. For the ER model, each edge of any gene pair (G_i, G_j) was added randomly in the GGN with probability pE independent from all other edges. After constructing the bone of GGN, we simulated gene expression based on multivariate Gaussian. Without loss of generality, we simulated data sets with n = 400 individuals, which is similar to the sample size in our real data examples. As shown in Table 1, we considered different number of genes m = 100, 200, 300 with various sparsity level determined by p_E = 1=m or 2=m for the ER model and e = 1 or e = 2 for the BA model.

We evaluated the performance of the methods by the following metrics: number of false positives (FP), false negatives (FN), FP+FN, F1 score, FDR, true positive rate (power). Note that there are three different weights used in joint modeling setting (`space-log`, `space-lasso`): (1) uniform weights; (2) residual variance based weights; and (3) degree freedom based weights. The corresponding methods are referred to as `sp_no`, `sp_res`, and `sp_df` with/without `log` respectively.

Under the BA model with m=100 and e=1 (Figure 1), we can see that `space-log` has smallest Errors (FP+FN), smallest FDR, and highest F1 score than other approaches, indicating that `space-log` controls overall false positive and false negative rates well. Under the ER model (Figure 2) with m=100 and e=1, `space-log` is slightly better than `space`, and `NS-log` shows lower Errors and higher F1 score than other approaches including `space-log`. Under both models, the log penalty has less false positives but slightly more false negatives compared to lasso penalty. We note that although log penalty performs well for both the ER and BA models, `space-log` is particularly powerful in identifying hub networks (such as BA models).

In the *Extended data*, Figures S3 and S4 show the results under the BA model for m=100,200,300 with low number of connections (e=1) and high number of connections (e=2), respectively. Figures S5 and S6 show the results under the ER model with low and high numbers of connections, respectively. Comparing with Figure 1, a similar pattern was noted with the increase of number of genes (m increases from 100, 200, to 300). In BA with low connections (Figure S3), `space-log` showed smallest FP+FN error and largest F1 score, which outperform all other methods. In BA with high connections (Figure S4), `NS-log` showed smallest FP+FN error and largest F1 score. For ER model with low and high connections, `NS-log` outperforms other methods in terms of FP+FN and F1 scores. It's

Table 1. Simulation settings.

m	*n*	p_E (ER)	*e* (BA)
400	100	1/100, 2/100	1,2
400	200	1/200, 2/200	1,2
400	300	1/300, 2/300	1,2

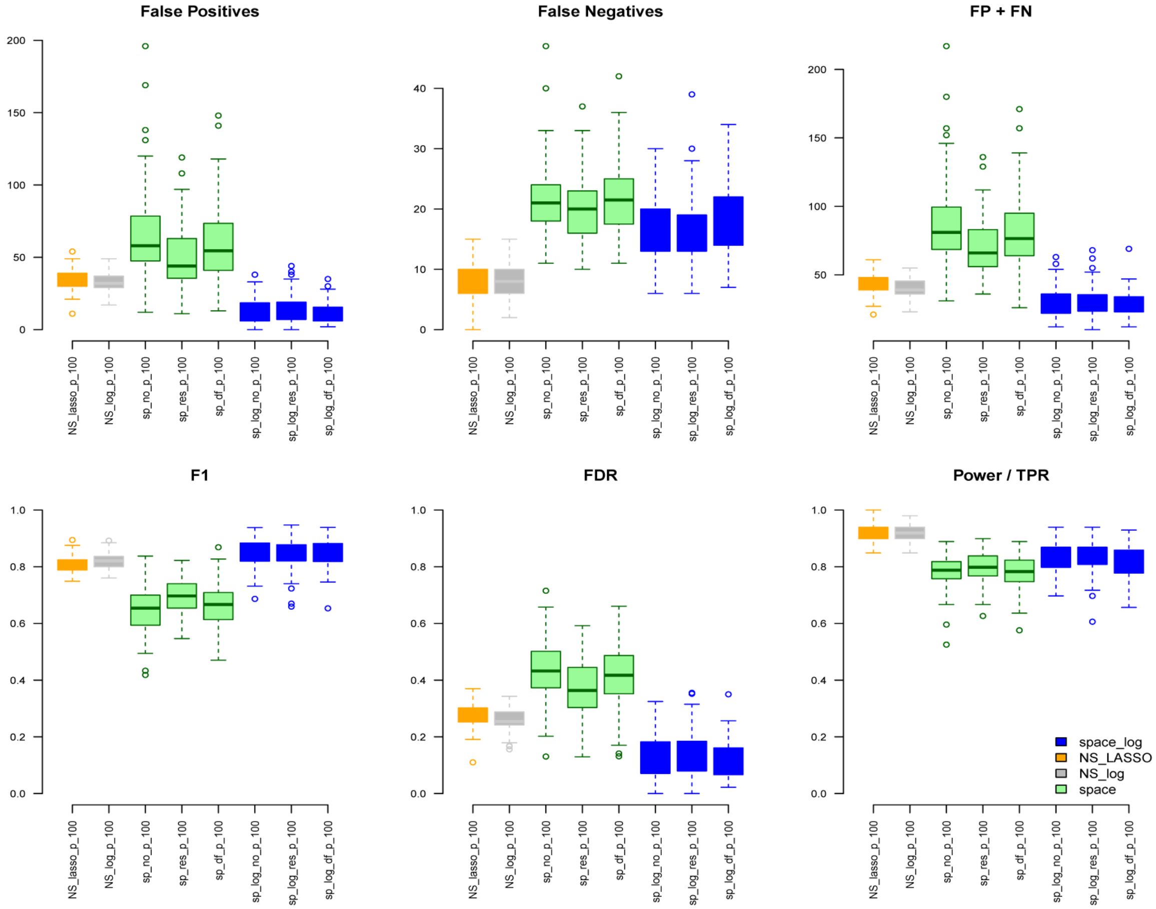

Figure 1. BA with 100 genes and e=1.

in line with our understanding that space-log is powerful at identifying hubs network, and NS-log is powerful at dealing with complex network with high number of gene-gene interactions and random networks.

We showed a simulated graph for the BA model with 400 subjects, 100 genes and each gene has only 1 connection (Figure 3). The GGN was estimated by 8 different approaches. In Figure 3, the true edges were indicated by black color, false positive (FP) edges by red color, and false negative (FN) edges by grey color. It's clear that space-log identified far fewer false positive edges (red line) comparing with space-lasso and NS approaches, while clearly indicating the hub structures. We observed that the FP edges by two NS approaches were quite randomly identified, and the FP edges by two space approaches were mostly within a hub and not between hubs.

We also compared the run-time under different simulation scenarios in Figure S7 and Table 2 under the BA model and low number of connections (e=1). Here the computing time for the space with no weight is presented. The computational time of the methods using log penalty is higher than that using lasso penalty, and the computational time for the space methods is much less than NS methods. Similar to what was observed in 8, the runtime increases approximately linearly with m. The difference of the run-time between the ER and BA model, the low and high numbers of connections or different weights of space methods are not significant (Figure S7).

In summary, the log penalty generally has better performance than the lasso penalty, and both space-log and NS-log control false positive and false negative rate well. For random networks, i.e., no hub, NS-log performs better than other methods. space-log performs best for hub-like gene networks (Figure 1 and Figure 3) with higher F1 score and less false positive edges. Identifying hub networks is generally considered of great interest in the GGN analysis, because a few of hubs connecting with a large proportion of genes, and those hub genes are thought to be master regulators and play a critical role in a biological system[14].

Application to GTEx and TCGA data

TCGA data

We applied both proposed space-log and existing methods (space-lasso, NS-log, and NS-lasso) to identify

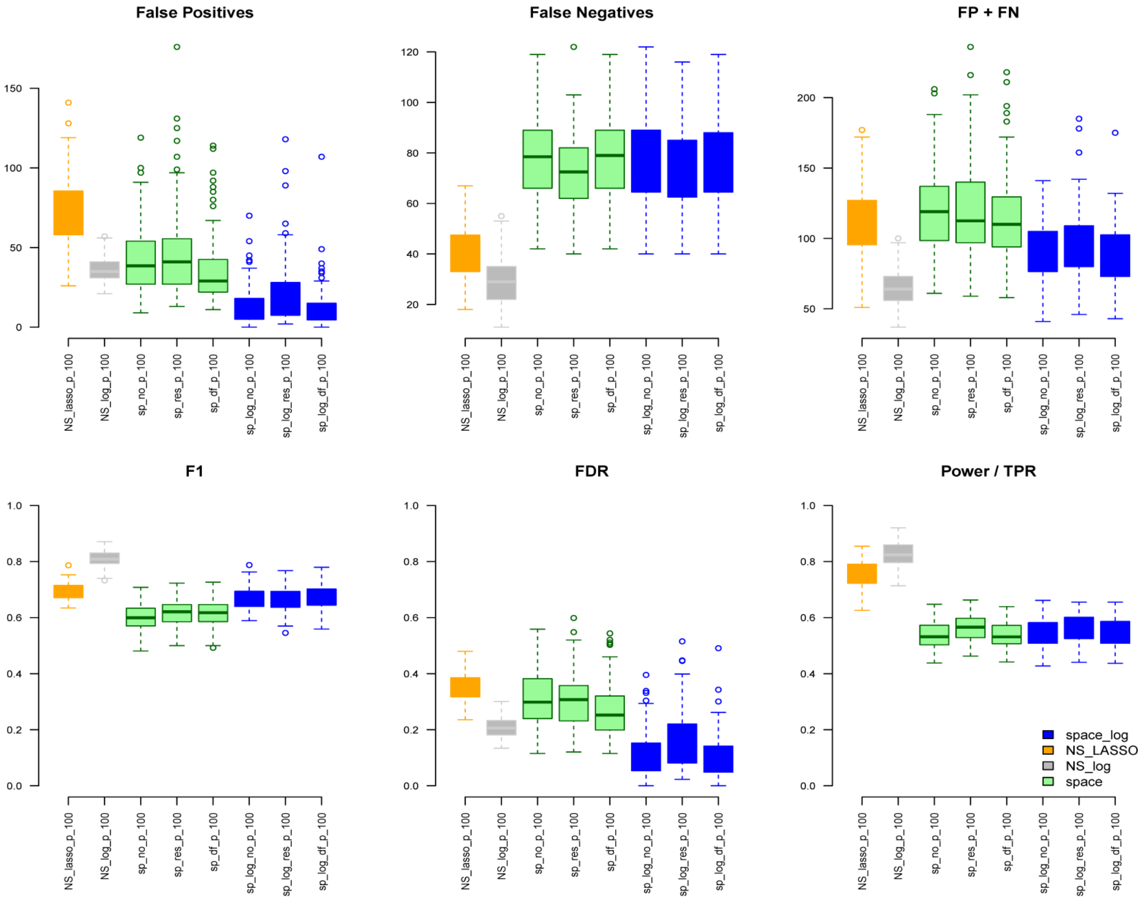

Figure 2. **ER with 100 genes and e=1.**

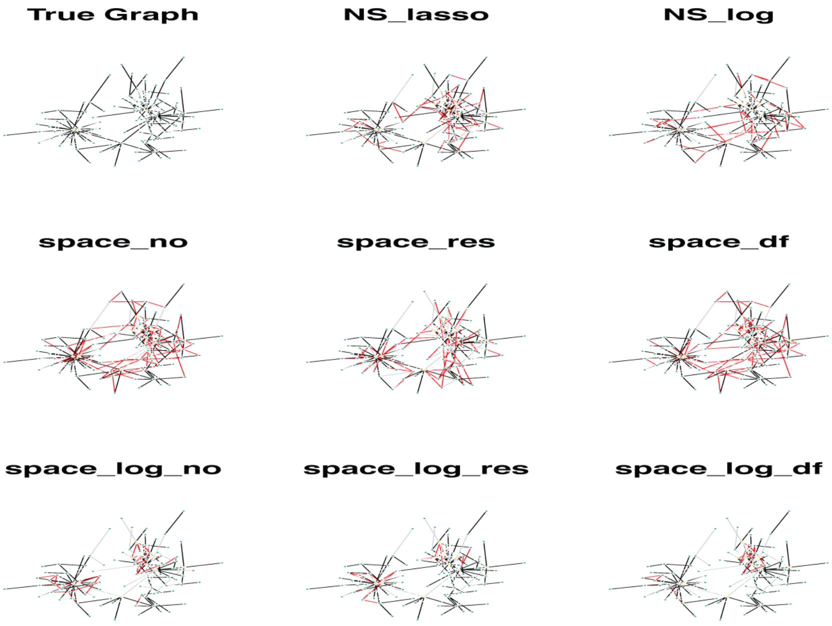

Figure 3. **A simulated graph for the BA model with 100 genes (e=1) and multiple hubs.** A total of 400 subjects were generated. The GGN was estimated by 8 different approaches. Black is true edges, red is false positive edges, and grey is false negative edges.

Table 2. Computational time in minutes with median (range) under the BA model with e=1 and n=400.

Methods	m=100	m=200	m=300
NS-lasso	6 (0.04, 8)	13 (0.2, 19)	20 (0.6, 30)
NS-log	71 (0.6, 94)	142 (3, 190)	226 (8, 293)
space	0.81 (0.1, 2,2)	2 (0.5, 3.3)	3 (1.1, 9)
space-log	9 (0.7,18)	16 (2.4, 27)	23 (5, 45)

GGN using RNA-seq data from tumor tissue of 550 TCGA (The Cancer Genome Atlas) Colon Adenocarcinoma (TCGA-COAD) cancer patients[15]. The preprocessing steps of RNA-seq data included: (1) transforming the expression of each gene by log(total read count) = *logTReC* (2) removing the confounding effects by taking residuals of logTReC from a linear regression with the following covariates: 75% of logTReC per sample (which captures read depth), plate, institution, age, and six PCs from the corresponding germline genotype data. After removing genes with low expression across most samples, we had 18,238 genes and 450 samples.

We considered gene sets C6 curated oncogenic pathways by MSigDB from the Broad Institute and inferred the GGN within each gene set. There were 189 gene pathways with a total of 8,737 unique genes for which TCGA have expression data. The sizes of gene sets ranged from 9 to 338 genes. Since we don't know the true GGN, we downloaded the common pathway version 10 from www.pathwaycommons.org to provide a partial "gold standard". The observed GGN by different methods were compared with the known edges from common pathway and calculated FP, FN, FP+FN, number of total discovery, F1 score, and true positive rate (Figure 4). The NS-based approach with both LASSO and log penalty discovered much more edges than space-based approach and `space-log` had fewer false positive (fewer FN+FP too) than `space-lasso`. There is almost no difference on number of false negative between different methods, as well as F1 score (Figure 4). Furthermore, in order to show the performance of these methods on the hub networks, we identified 17 pathways with hub-like genes (each hub gene set has < 50 genes and variance of the number of identified edges for each gene in the gene set > the first quartile of all 189 gene sets) and re-calculated the summary metrics in Figure 5. We noted that `space-log` approach has smallest Errors and slightly higher F1 than other approaches, which is in line with our finding in simulation that `space-log` is powerful in identifying hub networks, (such as BA models).

To further understand whether our proposed methods can better handle high dimensional data with larger number of genes, we generated k=71 larger network by combining gene sets with overlapping genes. The sizes of this new larger gene sets list ranged from 476 to 1253 genes, which is larger than the sample size n=450 (m > n). We applied `space-log`, `space-lasso`, `NS-log`, and NS-lasso approaches on this new larger list, and calculated similar metrics (FP, FN, FP+FN, etc in Figure S8). Comparing with TCGA data with smaller gene sets (Figure 4 and Figure 5), we observed that the NS-based approach with both lasso and log penalty discovered much more false positive edges than space-based approaches for this new larger gene sets setting, which is consistent with our finding in simulation that NS-based approach has more false positive in general.

GTEx data

The Genotype Tissue Expression (GTEx) project[16] aims to study tissue-specific gene expression and regulation in normal individuals. In this paper, we used gene expression data (RNA-seq) from blood tissue of 451 patients to identify GGN. We pre-processed gene expression data using the same procedure as for TCGA data. We mapped genes to gene pathways by MSigDB (https://www.gsea-msigdb.org/gsea/msigdb/index.jsp). A total of 189 gene pathways were represented with a total of 8097 unique genes. The size of gene sets ranged from 8 to 306 genes.

Again, we applied `space-log`, `space-lasso`, `NS-log`, and NS-lasso approaches to identify GGN. Using the same common pathway file used for the TCGA analysis as gold standard, we calculated FP, FN, FP+FN, # of discovery, F1, TPR (Figure 6). We obtained very similar results to the TCGA data. The NS-based approach with both LASSO and log penalty discovered much more edges than the space-based approach and `space-log` has fewer false positive (fewer FN+FP too) than `space-lasso`. There is almost no difference in the number of false negative between different methods, as well as F1 score. A similar sensitivity analysis was conducted to a subset of hub-type genes (Figure 7), where 30 pathways were selected to be in the first quartile of the variance of the number of identified genes with < 50. It also showed the `space-log` approach has smallest Errors (F1 is similar to other approaches).

Conclusions

In this paper, we proposed a new joint modeling method with log penalty, `space-log`, to identify gene-gene network. An assumption of the GGN analysis is that most of gene pairs do not directly interact with each other, and there are a few of master genes (hubs) for network that connect with many other genes, which are thought to play a critical role in a biological system[1,14]. Both simulation and real data analyses showed that `space-log` is particularly powerful in identifying hub networks and master genes, which is considered of great interest in gene-gene network analysis. In the *Extended data*, we compared several tuning parameter selection approaches, such as BIC Zou *et al.*[17], extBIC[9,18], and oracle[5], and showed that extBIC outperforms other methods in simulation. The R package "SpaceLog" on GitHub includes algorithms, simulation, and real data examples: https://github.com/wuqian77/SpaceLog.

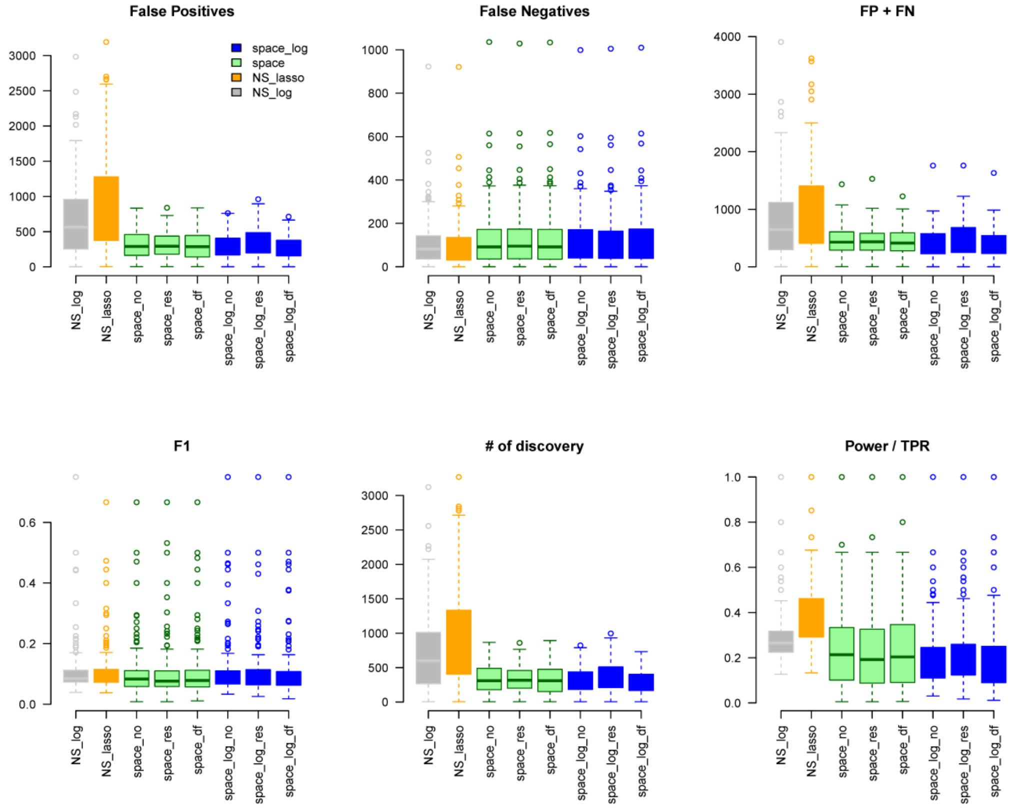

Figure 4. TCGA data analysis with ALL 189 Gene Sets.

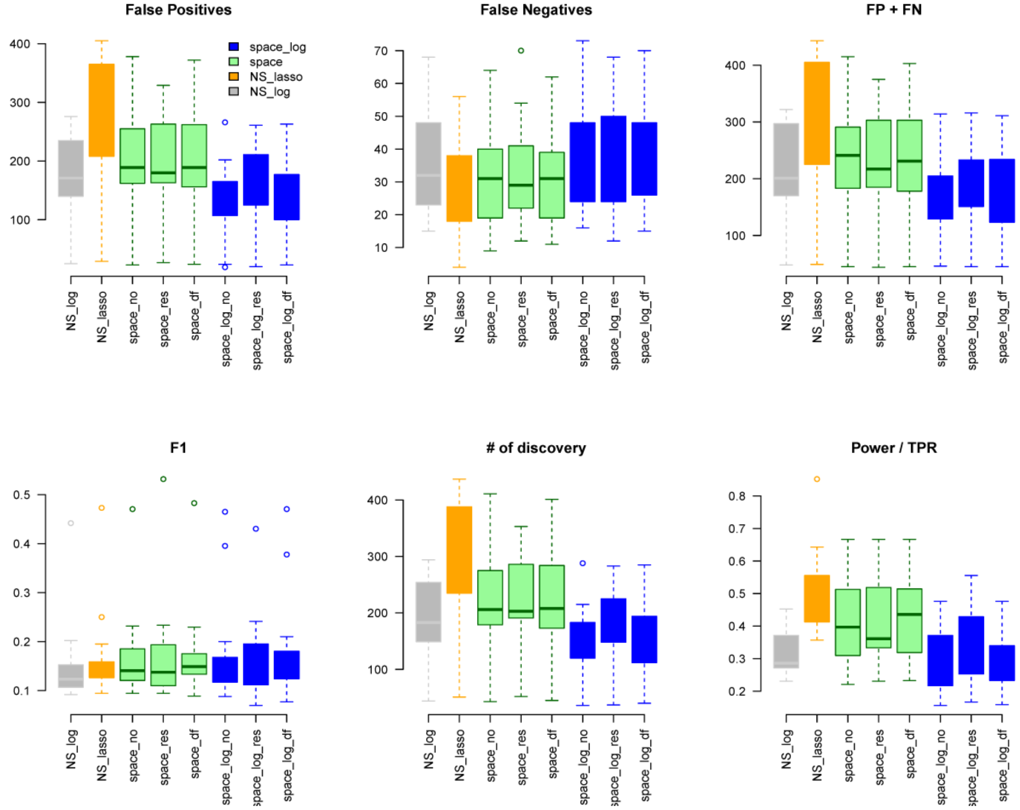

Figure 5. TCGA data analysis with BA hub-type Gene Sets.

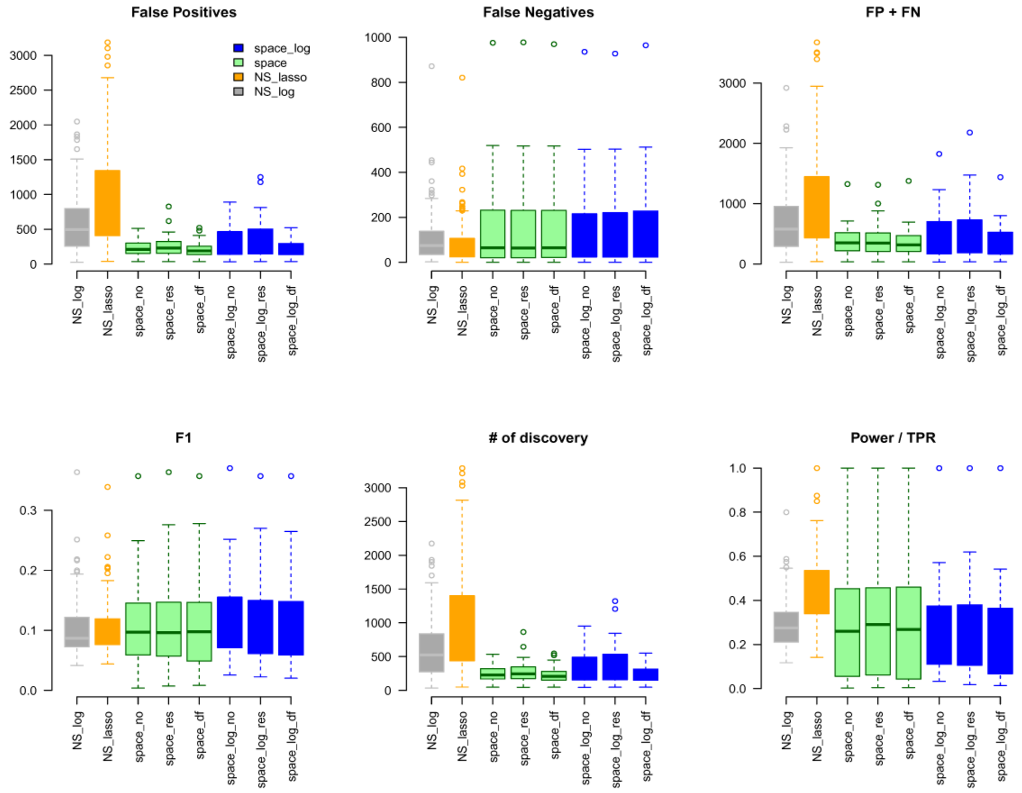

Figure 6. **GTEx data analysis with ALL 189 Gene Sets.**

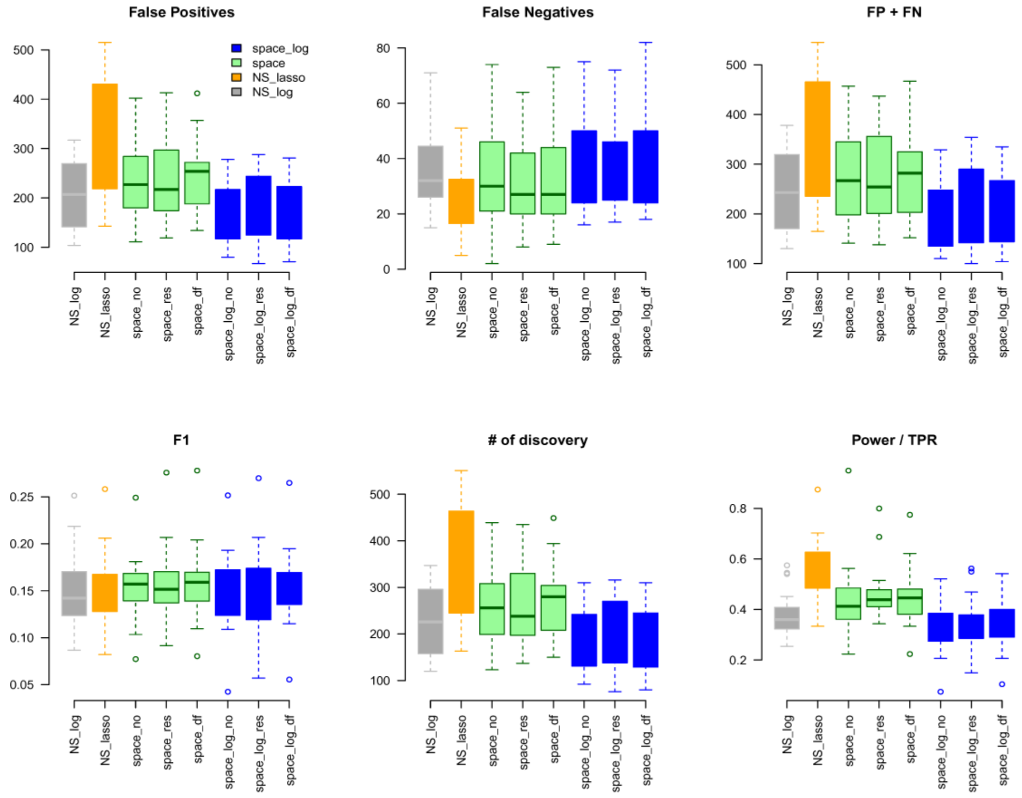

Figure 7. **GTEx data analysis with BA hub-type Gene Sets.**

Data availability

Underlying data

Simulation data. We used barabasi.game function from igraph R package to generate the skeleton of a BA model.

Source code, simulated data, and plots: https://github.com/wuqian77/SpaceLog/tree/master/Simulation.

TCGA data. The RNA-seq dataset from tumor tissue of 550 TCGA (The Cancer Genome Atlas) colon adenocarcinoma (TCGA-COAD) cancer patients[15] can be downloaded from dbGap phs000178.v1.p1.: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga.

Pre-processing and data analysis source code: https://github.com/wuqian77/SpaceLog/tree/ master/Analysis/TCGA.

GTex data. The RNA-seq dataset of blood tissue from the Genotype Tissue Expression (GTEx) project[16] can be downloaded from dbGap Genotype-Tissue Expression Project and the study accession is phs000424.v7.p2.: https://www.gtexportal.org/home/datasets.

Pre-processing and data analysis source code: https://github.com/wuqian77/SpaceLog/tree/master/Analysis/GTex.

Extended data

Zenodo: SpaceLog: First release of spacelog, http://doi.org/10.5281/zenodo.4002931[19].

This project contains the following extended data:

- the detailed algorithm for active shooting;
- simulation and figures on comparing methods to choose tuning parameters;
- simulation and figures on comparing different GGN methods under various scenarios.

License: GPL-3

Software availability

Source code for `space-log` available from: https://github.com/wuqian77/SpaceLog

Archived source code as at time of publication: http://doi.org/10.5281/zenodo.4002931[19].

License: GPL-3

Source code for `NS-log` and NS-lasso available from: https://github.com/Sun-lab/penalized_estimation

License: GPL-3

Existing methods `space-lasso` is available on R CRAN: https://cran.r-project.org/web/packages/space/index.html.

References

1. Liu Z, Meng J, Li X, *et al.*: **Identification of hub genes and key pathways associated with two subtypes of diffuse large b-cell lymphoma based on gene expression profiling via integrated bioinformatics.** *Biomed Res Int.* 2018; **2018**: 3574534.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**
2. Meinshausen N, Bühlmann P: **High-dimensional graphs and variable selection with the lasso.** *Ann Statist.* 2006; **34**(3): 1436–1462.
Publisher Full Text
3. Friedman J, Hastie T, Tibshirani R: **Sparse inverse covariance estimation with the graphical lasso.** *Biostatistics.* 2008; **9**(3): 432–441.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**
4. Peng J, Wang P, Zhou N, *et al.*: **Partial correlation estimation by joint sparse regression models.** *J Am Stat Assoc.* 2009; **104**(486): 735–746.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**
5. Chen TH, Sun W, Fine JP: **Designing penalty functions in high dimensional problems: The role of tuning parameters.** *Electron J Stat.* 2016; **10**(2): 2312–2328.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**
6. Tibshirani R: **Regression shrinkage and selection via the lasso.** *J R Statist Soc B.* 1996; **58**(1): 267–288.
Publisher Full Text
7. Sun W, Ibrahim JG, Zou F: **Genomewide multiple-loci mapping in experimental crosses by iterative adaptive penalized regression.** *Genetics.* 2010; **185**(1): 349–359.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**
8. Ha MJ, Sun W, Xie J: **Penpc: A two-step approach to estimate the skeletons of high-dimensional directed acyclic graphs.** *Biometrics.* 2016; **72**(1): 146–155.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**
9. Chen J, Chen Z: **Extended bayesian information criteria for model selection with large model spaces.** *Biometrika.* 2008; **95**(3): 759–771.
Publisher Full Text
10. Zou H, Li R: **One-step sparse estimates in nonconcave penalized likelihood models.** *Ann Stat.* 2008; **36**(4): 1509–1533.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**
11. Erdös P, Rényi A: **On the evolution of random graphs.** *Publ Math Inst Hung Acad Sci A.* 1960; **5**: 17–61.
Publisher Full Text
12. Kalisch M, Bühlmann P: **Estimating High-Dimensional Directed Acyclic Graphs with the PC-Algorithm.** *J Mach Learn Res.* 2007; **8**(Mar): 613–636.
Reference Source
13. Barabási AL, Albert R: **Emergence of scaling in random networks.** *Science.* 1999; **286**(5439): 509–512.
PubMed Abstract | **Publisher Full Text**
14. Barabasi AL, Oltvai ZN: **Network biology: understanding the cell's functional organization.** *Nat Rev Genet.* 2004; **5**(2): 101–13.
PubMed Abstract | **Publisher Full Text**
15. Cancer Genome Atlas Network: **Comprehensive molecular characterization of human colon and rectal cancer.** *Nature.* 2012; **487**(7407): 330–337.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**
16. Aguet F, Brown AA, Castel S, *et al.*: **Local genetic effects on gene expression across 44 human tissues.** *bioRxiv.* 2016; 074450.
Publisher Full Text
17. Zou H, Hastie T, Tibshirani R, *et al.*: **On the "degrees of freedom" of the lasso.** *Ann Statist.* 2007; **35**(5): 2173–2192.
Publisher Full Text
18. Chen J, Chen Z: **Extended bic for small-*n*-large-*p* sparse glm.** *Statistica Sinica.* 2012; 555–574.
Publisher Full Text
19. wuqian77: **wuqian77/SpaceLog: First release of spacelog (Version v.1.0.1).** *Zenodo.* 2020.
http://www.doi.org/10.5281/zenodo.4002931

Open Peer Review

Current Peer Review Status: ✓ ? ✓

Version 2

Reviewer Report 20 January 2022

https://doi.org/10.5256/f1000research.58701.r118658

✓

Jessie Jeng
Department of Statistics, North Carolina State University, Raleigh, NC, USA

Revisions are sufficient.

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 12 January 2022

https://doi.org/10.5256/f1000research.58701.r118659

✓

Yuying Xie
Department of Computational Mathematics, Science, and Engineering, Michigan State University, East Lansing, MI, USA
Yuning Hao
Amazon, Seattle, WA, USA

The authors have answered all my concerns. Congratulations!

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Statistics, Biostatistics, Genetics, Bioinformatics

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 21 June 2021

https://doi.org/10.5256/f1000research.28834.r86957

?

Jessie Jeng
Department of Statistics, North Carolina State University, Raleigh, NC, USA

This paper considers the problem of identifying gene regulatory networks based on high-dimensional gene expression data. A new method, space-log, is introduced to perform penalized regression with a log penalty. The new method is compared with several existing methods in simulation and real applications using GTEx and TCGA data. The new method seems to outperform other methods in identifying networks with hubs and master genes.

The overall presentation is clear. However, several improvements can be made:

1. I think it would be important to explain why the proposed method has advantages in identifying networks with hubs. I would imagine that a network with hubs tends to have more diverse sparsity levels of association across different genes than a network without hubs. Is the proposed method with a log penalty more effective for such scenarios? Or there are other reasons?

2. The Introduction is a bit too short. I would suggest adding more biological background and a motivation from a real application point of view.

3. Some discussions on the computational complexity or a comparison of the computational times of different methods should be helpful.

Is the rationale for developing the new method (or application) clearly explained?

Partly

Is the description of the method technically sound?

Yes

Are sufficient details provided to allow replication of the method development and its use by others?

Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Statistics, Statistical Genomics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 02 Dec 2021

Qian Wu, Fred Hutchinson Cancer Research Center, Seattle, USA

1. I think it would be important to explain why the proposed method has advantages in identifying networks with hubs. I would imagine that a network with hubs tends to have more diverse sparsity levels of association across different genes than a network without hubs. Is the proposed method with a log penalty more effective for such scenarios? Or there are other reasons?

We totally agree with the reviewer that hub type network has more diverse sparsity (BA type graph has a larger variation (number of edges per gene) than ER type graph) and one additional evidence that log penalty works better for diverse sparsity are the results of an earlier paper https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5628772/ e.g., the results in tables 1 and 2.

2. The Introduction is a bit too short. I would suggest adding more biological background and motivation from a real application point of view.

Good comment. We added biological background in the introduction section accordingly.

3. Some discussions on the computational complexity or a comparison of the computational times of different methods should be helpful.

We thank the reviewer for pointing out this. We have added the run-time discussion with a table in the online manuscript, and a figure in Supplementary Materials on Github.
https://github.com/wuqian77/SpaceLog/blob/master/Document/F1000Research_Journal_Article_Spacelog_Supplem

Competing Interests: NA

Reviewer Report 19 October 2020

https://doi.org/10.5256/f1000research.28834.r71864

? **Chi Song**
College of Public Health, Division of Biostatistics, The Ohio State University, Columbus, OH, USA

In this paper, the authors developed “space-log”, a new method to infer gene-gene network (GGN), by incorporating log penalty into the “space” method. The authors compared their method to the original “space” with LASSO penalty and the neighborhood selection (NS) methods, using simulation and real data. In general, the description of the proposed method is clear, and the simulation and real application settings made sense. However, I believe this paper needs to be revised to (1) provide the rationale for choosing the combination of “space” and log penalty, (2) compare with more state-of-the-art methods, and (3) improve the presentation and writing. Here are some specific comments.

1. The authors need to discuss why they decided to combine “space” and log penalty for GGN detection. What is the benefit of using log penalty over other penalty choices, including concave penalties and other nonconcave penalties such as SCAD, MCP, and TLP? The introduction should be extended to include more state-of-the-art methods, such as adaptive LASSO or high-dimensional regression methods with sparse precision matrix estimation.

2. Although briefly mentioned gLasso in the introduction, the authors did not talk about this class of methods in the rest of the paper. I am curious to see how space-log compares to gLasso and its extensions with nonconcave penalties in simulation and real application (see Fan *et al.*, 2009). [1]

3. Based on my understanding, the major benefit of using nonconcave penalties is to reduce the estimation bias of the correlation coefficients. I suggest the authors include comparisons based on the coefficient estimation. In addition, the authors should provide details about how the precision matrices are simulated. Currently, only the simulation methods for its bone structure are provided.

4. In conclusion/discussion, the authors should provide more insights or heuristics about why the proposed method performed better than other methods. What aspects of the data makes the proposed method favorable? Is there any limitation that the users should pay attention to when applying this method in application?

Minor comments:

1. The citations in this paper are apparently converted from an author-year citation format. The authors need to make changes to the writing to adapt to numbered citation format.

2. Presentation and writing issues, for example:

- The authors need to decide if a comma is needed between two subscripts in the notation (e.g. \beta_{ij} or \beta_{i,j} on page 3 left column).
- On page 4, left column, third paragraph, the sentence “The initial paragraph ...” only applies

to BA model, but not ER model.

- Also, it should be “in the (t+1)th step” instead of “in the (k+1)th step”.
- In the last sentence of this paragraph, it should be “1/m” and “2/m”.
- This is not an exhaustive list of writing issues. I suggest the authors proofread carefully.

References

1. Fan J, Feng Y, Wu Y: Network exploration via the adaptive LASSO and SCAD penalties. *The Annals of Applied Statistics*. 2009; **3** (2): 521-541 Publisher Full Text
2. Mohsen Pourahmadi: Sparse Gaussian Graphical Models. 2013. 121-140 Publisher Full Text

Is the rationale for developing the new method (or application) clearly explained?

Partly

Is the description of the method technically sound?

Yes

Are sufficient details provided to allow replication of the method development and its use by others?

Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Partly

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Biostatistics, Bioinformatics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 14 October 2020

https://doi.org/10.5256/f1000research.28834.r71868

? **Yuying Xie**
Department of Computational Mathematics, Science, and Engineering, Michigan State University, East Lansing, MI, USA
Yuning Hao
Amazon, Seattle, WA, USA

The paper introduces an extension of the SPACE (Sparse Partial Correlation Estimation) method, which is used to estimate a gene-gene network (GNN). The proposed framework, space-log, relies on the log penalty, which delivers better variable selection performance than LASSO, especially for GGN with hubs. The authors also have created a very efficient R package for the proposed method. The paper is clearly written, and the proposed method showed promising results for real data analyses. I list my questions below:

1. In both TCGA and GTEx analyses, the dimensions m is smaller than the sample size n. Since log penalty can handle high dimension low sample size data, can you also include a real application with m larger than n?

2. The numerical results are all based on the tuned model using extended BIC, which is one-shot of the result. It is better to also include ROC curves reflecting the whole spectrum of the results for a range of tuning parameters.

Minor comments:

1. In formula (3), the ‘square root’ should not include Yj.

2. In t page 3 section ‘Joint modeling space using lasso penalty: space-lasso’, it is better to include the relationship between \beta_{ij} and \simga_{ij} before making the conclusion that \rho_{i,j} = sign(\beta_{i,j}) \sqrt{\beta{ij}\beta_{ji}}.

3. For the log penalty, please also cite ‘Enhancing sparsity by reweighted L1 minimization’.

Is the rationale for developing the new method (or application) clearly explained?

Yes

Is the description of the method technically sound?

Yes

Are sufficient details provided to allow replication of the method development and its use by others?

Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Statistics, Biostatistics, Genetics, Bioinformatics

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Author Response 08 Dec 2021

Qian Wu, Fred Hutchinson Cancer Research Center, Seattle, USA

1. In both TCGA and GTEx analyses, the dimensions m is smaller than the sample size n. Since log penalty can handle high dimension low sample size data, can you also include a real application with m larger than n?

We thank the reviewer for pointing out this limitation. We "create" some larger network by combining some gene sets with overlapping genes (m > n). An updated "larger size network" graph is added in supplementary.
https://github.com/wuqian77/SpaceLog/blob/master/Document/F1000Research_Journal_Article_Spacelog_S

1. The numerical results are all based on the tuned model using extended BIC, which is one-shot of the result. It is better to also include ROC curves reflecting the whole spectrum of the results for a range of tuning parameters.

Good comment. Cross-validation (CV) approach is commonly used to choose the tuning parameter but time-consuming, and recent literature (Wang et al., 2009) showed BIC-type approach has better performance than CV. Thus, we tried different tuning parameters and used grid search to generate an "Oracle" result (based on maximize F1 score or minimize FDR) in supplementary (section S2.3). It showed extBIC performs better than BIC and its performance is close to Oracle for space approach.
https://github.com/wuqian77/SpaceLog/blob/master/Document/F1000Research_Journal_Article_Spacelog_S

Minor comments:

1. In formula (3), the 'square root' should not include Yj.

2. In t page 3 section 'Joint modeling space using lasso penalty: space-lasso', it is better to include the relationship between \beta_{ij} and \simga_{ij} before making the conclusion that \rho_{i,j} = sign(\beta_{i,j}) \sqrt{\beta{ij}\beta_{ji}}.

3. For the log penalty, please also cite 'Enhancing sparsity by reweighted L1 minimization'.

We thank the reviewer for pointing out this oversight. We have updated the manuscript accordingly. _

Competing Interests: No competing interests were disclosed.