Space-log: a novel approach to inferring gene-gene networks using SPACE model with log penalty [version 1; peer review: awaiting peer review]

Qian (Vicky) Wu¹,², Wei Sun², Li Hsu²

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

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 often referred to as gene regulatory networks since it may reveal expression regulation. Most of existing methods for identifying GGN employ penalized regression with $L_1$ (lasso), $L_2$ (ridge), or elastic net penalty, which spans the range of $L_1$ to $L_2$ penalty. However, for high dimensional gene expression data, a penalty that spans the range of $L_0$ and $L_1$ penalty, such as the log penalty, is often needed for variable selection consistency. Thus, we develop a novel method that employs 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.
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.

Copyright: © 2020 Wu Q et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

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 1; peer review: awaiting peer review] F1000Research 2020, 9:1159
https://doi.org/10.12688/f1000research.26128.1

Introduction
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\textsuperscript{1}, graphical Lasso\textsuperscript{2}, and space (Sparse Partial Correlation Estimation)\textsuperscript{3}. 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\textsuperscript{4}. In this paper, we propose a new statistical method to estimate GGN by implementing the log penalty for the space approach, and we refer to our method as \textit{space-log}.

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: \textit{NS-lasso, NS-log}

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

\[
\hat{\beta} = \arg \min \left\{ \frac{1}{2} \left( Y - Y \beta \right)^T \left( Y - Y \beta \right) + \sum_{i,j} p(\beta_{ij}; \omega) \right\}
\]

(1)

with penalty function \( p(\beta; \omega) \). We will compare \textit{NS-lasso} with lasso penalty \( p(\beta; \lambda) = \lambda \| \beta \|_1 \) and \textit{NS-log} with log penalty \( p(\beta; \lambda, \tau) = \lambda \log(\| \beta \| + \tau) \). Source codes of \textit{NS-lasso} and \textit{NS-log} are available at https://github.com/Sun-lab/penalized_estimation/.

Joint modeling \textit{space\-lasso}: \textit{space\-log}

The joint modeling approach \textit{space\-lasso} 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 \( \rho_{ij} \). If we know the concentration matrix \( \Sigma^{-1} = (\sigma_{ij})_{m \times m} \), then \( \rho_{ij} = -\frac{\sigma_{ij}}{\sqrt{\sigma_{ii} \sigma_{jj}}} \).

So we can easily get that \( \rho_{ij} = \text{sign}(\beta_{ij}) \sqrt{\rho_{ij}} \). Thus, the problem is translated into partial correlation matrix estimation. Specifically\textsuperscript{7}, proposed to minimize a penalized loss function

\[
L_n(\beta, \sigma, Y) = \frac{1}{2} \sum_{j=1}^m \left( Y_j - \sum_{i \neq j} \hat{\beta}_{ij} Y_i \right)^2 + \sum_{i,j} p(\rho_{ij}; \lambda)
\]

(2)

where \( w_j \geq 0 \) is the weight, e.g., uniform weights \( w_j = 1 \) for space-no, residual variance based weights \( w_j = \sigma_j^2 \) for space-res, and degree based weights \( w_j = \text{number of genes that } \{j : \rho_{ij} \neq 0, j \neq i \} \) for space-deg. In 3, \( p(\rho_{ij}; \lambda) = \lambda |\rho_{ij}| \) and we call it as \textit{space\-lasso}.

New algorithm \textit{space-log}: joint modeling space using log penalty

Inspired by 6 and 7, we extended the space approach with log penalty as \textit{space-log} \( p(\rho; \lambda, \tau) = \lambda \log(\| \rho \| + \tau) \) and used the active shooting algorithm\textsuperscript{8} to update the coefficient estimates iteratively in \textit{space-log} (Extended Data). We determined the tuning parameters by using extended BIC (extBIC)\textsuperscript{9}.

Denote the target loss function as

\[
f(\rho, \sigma) = \frac{1}{2} \sum_{j=1}^m w_j \left( Y_j - \sum_{i \neq j} \hat{\beta}_{ij} Y_i \right)^2 + \sum_{i,j} p(\rho_{ij}; \lambda, \tau)
\]

(3)

where \( \rho_{ij} \) is the estimate of regression coefficient \( \rho_{ij} \) at the \( k \)-th step, while solving for \( \rho_{ij} \) by

\[
f^{(k+1)}(\rho_{ij}) = \frac{1}{2} \sum_{j=1}^m w_j \left( Y_j - \hat{\rho}_{ij} Y_i \right)^2 + \sum_{i,j} p'(\rho_{ij}; \lambda, \tau) |\rho_{ij}| \rho_{ij}
\]

(4)

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

\[
\rho_{ij}^{(k+1)} = \begin{cases} 0, & |z_{ij}^{(k)}| \leq y_j \cdot p'(\hat{\rho}_{ij}^{(k)}; \lambda, \tau) \\
\text{sgn}(\hat{\rho}_{ij}^{(k)}) |z_{ij}^{(k)}| - y_j \cdot p'(\hat{\rho}_{ij}^{(k)}; \lambda, \tau), & |z_{ij}^{(k)}| > y_j \cdot p'(\hat{\rho}_{ij}^{(k)}; \lambda, \tau) \end{cases}
\]

(5)

where \( z_{ij}^{(k)} = Y_j (Y_i - \sum_{i \neq j} \hat{\rho}_{ij}^{(k)} Y_i) / y_j \), \( y_j = Y_j^T Y_j \), and \( p'(\rho_{ij}; \lambda, \tau) = \text{sgn}(\rho_{ij}) \frac{\lambda}{|\rho_{ij}|^{1+\tau}} \).
Active-shooting
We adapted the same idea active-shooting algorithm from\(^3\) to update the coefficient estimation iteratively in space-log. Without loss of generality, we kept most notation from\(^1\) but tailored with space-log. The details are included in the Extended data.

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\(^1\), we studied two types of graphs: the traditional random graphs (ER model) where all the genes have the same expected number of neighbors\(^{10,11}\), 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\(^3\).

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 \(v_j^{(t)}\) number of edges connected with the \(r\)th gene at the \(r\)th step. For the ER model, each edge of any gene pair \((G_i,G_j)\) was added randomly in the GGN with probability \(p_E\) 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. 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 \(e=2\) 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 outperforms 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 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. 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\(^1\).

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 GGN using RNA-seq data from tumor tissue of 550 TCGA (The Cancer Genome Atlas) Colon Adenocarcinoma (TCGA-COAD) cancer patients\(^14\). 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.

---

**Table 1. Simulation settings.**

<table>
<thead>
<tr>
<th>(m)</th>
<th>(n)</th>
<th>(p_E) (ER)</th>
<th>(e) (BA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>400</td>
<td>100</td>
<td>1/100, 2/100</td>
<td>1,2</td>
</tr>
<tr>
<td>400</td>
<td>200</td>
<td>1/200, 2/200</td>
<td>1,2</td>
</tr>
<tr>
<td>400</td>
<td>300</td>
<td>1/300, 2/300</td>
<td>1,2</td>
</tr>
</tbody>
</table>
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 (Extended data: Figure S10). 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).

GTEx data

The Genotype Tissue Expression (GTEx) project\textsuperscript{15} 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.

Figure 1. BA with 100 genes and e=1.
**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.
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.
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$^{1,16}$ . 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$^{18}$, and oracle$^4$, 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.

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/Analysis/GTex.


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

GTEx data. The RNA-seq dataset of blood tissue from the Genotype Tissue Expression (GTEx) project$^{15}$ 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

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


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


The benefits of publishing with F1000Research:

• Your article is published within days, with no editorial bias
• You can publish traditional articles, null/negative results, case reports, data notes and more
• The peer review process is transparent and collaborative
• Your article is indexed in PubMed after passing peer review
• Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com