JOURNAL OF APPLIED STATISTICS
2020, VOL. 47, NO. 6, 1064–1083
https://doi.org/10.1080/02664763.2019.1667311

Taylor & Francis
Taylor & Francis Group

Check for updates

TGCnA: temporal gene coexpression network analysis using a low-rank plus sparse framework

Jinyu Li[a]*, Yutong Lai[a]*, Chi Zhang[b] and Qi Zhang [a]

[a]Department of Statistics, University of Nebraska-Lincoln, Lincoln, NE, USA; [b]School of Biological Sciences, University of Nebraska-Lincoln, Lincoln, NE, USA

ABSTRACT
Various gene network models with distinct physical nature have been widely used in biological studies. For temporal transcriptomic studies, the current dynamic models either ignore the temporal variation in the network structure or fail to scale up to a large number of genes due to severe computational bottlenecks and sample size limitation. Although the correlation-based gene networks are computationally affordable, they have limitations after being applied to gene expression time-course data. We proposed Temporal Gene Coexpression Network Analysis (TGCnA) framework for the transcriptomic time-course data. The mathematical nature of TGCnA is the joint modeling of multiple covariance matrices across time points using a 'low-rank plus sparse' framework, in which the network similarity across time points is explicitly modeled in the low-rank component. We demonstrated the advantage of TGCnA in covariance matrix estimation and gene module discovery using both simulation data and real transcriptomic data. The code is available at https://github.com/QiZhangStat/TGCnA.

ARTICLE HISTORY
Received 4 September 2018
Accepted 9 September 2019

KEYWORDS
Gene coexpression; transcriptomic time course; covariance matrix estimation; low-rank plus sparse; WGCNA; KEGG

1. Background

High throughput manner and low cost of sequencing technologies enable the biologists to generate an enormous wealth of data for discovering and quantifying the relationship among large amount and various types of molecular elements, such as gene expressions, proteins, metabolites and epigenetic marks. These elements and their relationship or interactions could be modeled as nodes and edges in a network model. Specifically, the gene co-expression network (GCN) models have been used for exploration, interpretation and visualization of the relationship among genes in a wide range of biological applications, including disease-gene association [43], identification of genes responding to environmental change, tissue-specific gene identification [37], and functional gene annotation [19]. GCNs can also be combined with other biological data in various analyses, such as identifying functional eQTLs [41] and studying gene-phenotype association [13]. Partially due

CONTACT Chi Zhang zhang.chi@unl.edu School of Biological Sciences, University of Nebraska-Lincoln, Lincoln, 68588 NE, USA; Qi Zhang qi.zhang@unl.edu Department of Statistics, University of Nebraska-Lincoln, Lincoln, 68583 NE, USA
*The two authors contributed equally to this work.

© 2019 Informa UK Limited, trading as Taylor & Francis Group

to this, many GCN databases were developed as annotation resources (e.g. GeneFriends [40], COXPRESdb [29] and PlaNex [44]).

Many construction tools and analysis tools were developed for GCN, and the physical nature of the resultant networks are different. WGCNA [46] was developed based on the marginal correlation of the gene pairs. GeneTS [34] and BicMix [15] built Gaussian Graphical Models (GGM) of genes, which were based on partial correlations. Under the multivariate Gaussian assumption, GGM captures the conditional dependence among genes. Mutual information has also been used in defining gene networks [8]. Comparing with Pearson's correlation, it could also capture the nonlinear association [36]. Different from the un-directed networks produced by the above methods, Bayesian Network (BN)-based methods infer directed networks from gene expression data [14,28].

Temporal transcriptomic data are extremely useful in biological studies, such as in developmental biology [17] and in stress biology [27]. GCN models have been utilized for these kinds of analysis. For instance, BNs could also be extended as Dynamic Bayesian Network (DBN) for time-course data, which models the directed gene-gene relationship across time points [23]. Besides DBN, another two lines of dynamic network models are based on the differential equations [47] and dynamic system model [5]. These methods, however, mostly focus on the temporal variation of the mean processes of the gene expression, but not the changes in the covariance structure. Additionally, these computational algorithms search in high-dimensional parameter space and require a large number of replicates and/or long computational time that does not scale to large networks. For example, Zhu and Wang [48] reported that it took 9 min for their proposed method HMDBN to learn a simulated dynamic network among 10 nodes using 1019 observations, while its competitors all took 11–58 h. Even though HMDBN has made tremendous progress along this direction, it may be still unrealistic to fit large DBN without further computational improvement, as the parameter space grows exponentially with the number of nodes. In practice, many biologists simply built one static GCN by aggregating all data together [4,27]. Such strategies mix various sources of heterogeneity and could describe neither the dynamics as in DBNs nor the time-specific snapshots of the biological systems. To construct the latter, the naive approach is simply building one GCN at each time. Then they can be compared them using differential network analysis [3]. It also has many drawbacks. First, the replicates at each time point may not be large enough to build a reliable and robust GCN. Second, ignoring the similarity of the gene expression at different time points within the same time course often results in false positives in differential network analysis across time points. In this paper, we presented a new method for building time-point specific GCNs, which we refer to as Temporal Gene Coexpression Network Analysis (TGCnA). It jointly estimates all GCNs for all time points using a novel 'low-rank plus sparse' framework [12,25]. In both simulations and the real data, we showed that the resulting GCNs were more robust and accurate than the separate modeling at each time point and thus led to more accurate downstream analysis.

2. Methods

2.1. Model setup

In a typical transcriptomic time-course study, suppose we collect n_t samples and observe the $p \times n_t$ data matrix $X_t = (x_{t1}, \ldots, x_{t,n_t})$ for $t \in \{t_1, \ldots, t_T\}$, where p is the number of

genes, n_t is the number of samples and T is the total number of time points. Reconstructing the GCN at each time point can be achieved by estimating Σ_t and the covariance matrix of the rows of X_t. Since this is our focus, we assume the rows of X_t have zero mean.

One naive estimate of Σ_t is simply the raw sample covariance matrix $\hat{\Sigma}_t$. This approach may lead to extremely noisy estimates, because the sample sizes at each time point are usually small. It also ignores the natural partial ordering of the samples (by time), and the network structural similarity across time points. Such similarity comes from various sources. For example, the genes in the same pathway tend to be coexpressed, and the pathway membership of the genes are fixed (e.g. KEGG database). The temporal change in GCN, however, could be caused by the common genes among these pathways share genes, and the time-varying pathway activity strength.

The above biological observations motivate us to model GCN with time-invariant latent factors and their time-varying loadings, an approach connected with the low-rank approximation of high-dimensional covariance matrix [11]. In the literature of matrix recovery, it has been noted that the low-rank approximation may be too restrictive and not robust, and a natural extension is the 'low-rank plus sparse' framework [12,25]. Towards this end, we also included a time-specific sparse component to reserve the significant links that cannot be captured by the factor model. In the context of gene networks, these sparse components are expected to capture the important time-specific gene-gene interactions that cannot be explained by the latent factors.

To summarize, we propose the following 'low-rank plus sparse' estimator for Σ_t

$$\hat{\Sigma}_t = UD_tU^{\mathrm{T}} + \hat{R}_t \tag{1}$$

Here U is a $p \times K$ matrix whose columns are the time-invariant latent factors learned from the transcriptomic data itself, the $K \times K$ diagonal matrix D_t are their loadings that change smoothly through time, and $\hat{R}_t$ denotes the estimated sparse component of the time-specific links at time point t. Our model shares the 'low-rank plus sparse' framework as in [12], but we extend their framework to the joint modeling of an ordered sequence of covariance matrices instead of one single covariance matrix.

2.2. TGCnA: temporal gene coexpression network analysis using low-rank plus sparse framework

We propose the Temporal Gene Coexpression Network Analysis (TGCnA, Algorithm 1) framework based on the 'low-rank plus sparse' model (Equation (1)).

Algorithm 1: TGCnA
Input: Time-course gene expression data matrices $X_{t_1}, X_{t_2}, \ldots, X_{t_T}$ (RNA-Seq or micro-array data)
Output: Estimated time-specific coexpression networks $\hat{\Sigma}_t$ for $t \in \{t_1, \ldots, t_T\}$
Steps:

(1) Extract the time-invariant latent factors U by applying SVD to the pooled normalized data matrix.

(2) Estimate the time-specific weights D_t via spline.
(3) Estimate the time-specific sparse component $\hat{R}_t$ by adaptive thresholding.
(4) Re-construct the covariance matrix at time t according to Equation (1) and calculate the correlation matrix.

When performing SVD on pooled data (across all time points) for extracting U in the first step, normalization is critical. If not done properly, the leading principal directions could be heavily influenced by the time points with larger sample sizes or larger variability. Data normalization typically involves centering and scaling. Our goal in this step is to find the principal directions that are representative at all time points. Thus, we scaled X_t with the leading singular value of X_t and then multiplied the square root of the number of genes $\sqrt{p}$. This approach assigned equal weights to the principal directions of the data at each time point. Write the pooled normalized data as

$$\tilde{X} = \sqrt{p}(\lambda_1^{-1} X_{t_1}, \ldots, \lambda_T^{-1} X_{t_T})$$

where λ_ℓ is the leading singular value of X_{t_ℓ}. Then we applied SVD to $\tilde{X}$ and used its top K left singular vectors as the columns of the time-invariant latent factors matrix U. Choosing the rank for low-rank matrix approximation is a difficult task [39], because fewer PCs will include incomplete information of the process while more PCs will cause the model over-parameterized and include noise. This problem could be posed as a model selection problem, and many information criteria have been applied, including Akaike's entropy-based Information Criterion (AIC, [1,2]) and Minimum Description Length (MDL [31,32,39], see Appendix A.1 for details). There is no dominant procedure for this problem, and both AIC and MDL are consistent under certain regularity conditions. However, Valle *et al.* [39] suggested that AIC tended to overestimate the dimension of the low-rank approximation in certain scenarios. This is consistent with our data analysis, where we found that both of AIC and MDL selected the correct number of components in simulations (Appendix Figure 1), while MDL selected a more compact model in the real data analysis (Appendix Figure 2). Thus, we recommend using MDL at this step.

In the second step, let U_k be the kth latent factor, and one natural raw estimate of its weight at time point t is $\tilde{d}_{tk} = U_k^{\mathrm{T}} X_t X_t^{\mathrm{T}} U_k$. For fixed $k = 1, \ldots, K$, we propose to further smooth $(\tilde{d}_{t_1,k}, \ldots, \tilde{d}_{t_T,k})$, the time-varying weight curve of U_k, at its log scale using spline, and report resultant sequence $(d_{t_1,k}, \ldots, d_{t_T,k})$ as the temporal weights of this factor. Finally, we define

$$D_t = \mathrm{diag}(d_{t,1}, \ldots, d_{t,K})$$

In the third step, we estimated the time-specific sparse component $\hat{R}_t$ using the adaptive thresholding rule [6] as used in [12]. To facilitate the presentation, we introduce the following two notations: we will use $A(i, j)$ to denote the element of A in its ith row and jth column, and $(b)_+$ to denote the positive part of b, i.e. it is 0 if b is negative. Then, let

$$\tilde{R}_t = \frac{1}{n_t - 1} X_t X_t^{\mathrm{T}} - U D_t U^{\mathrm{T}}$$

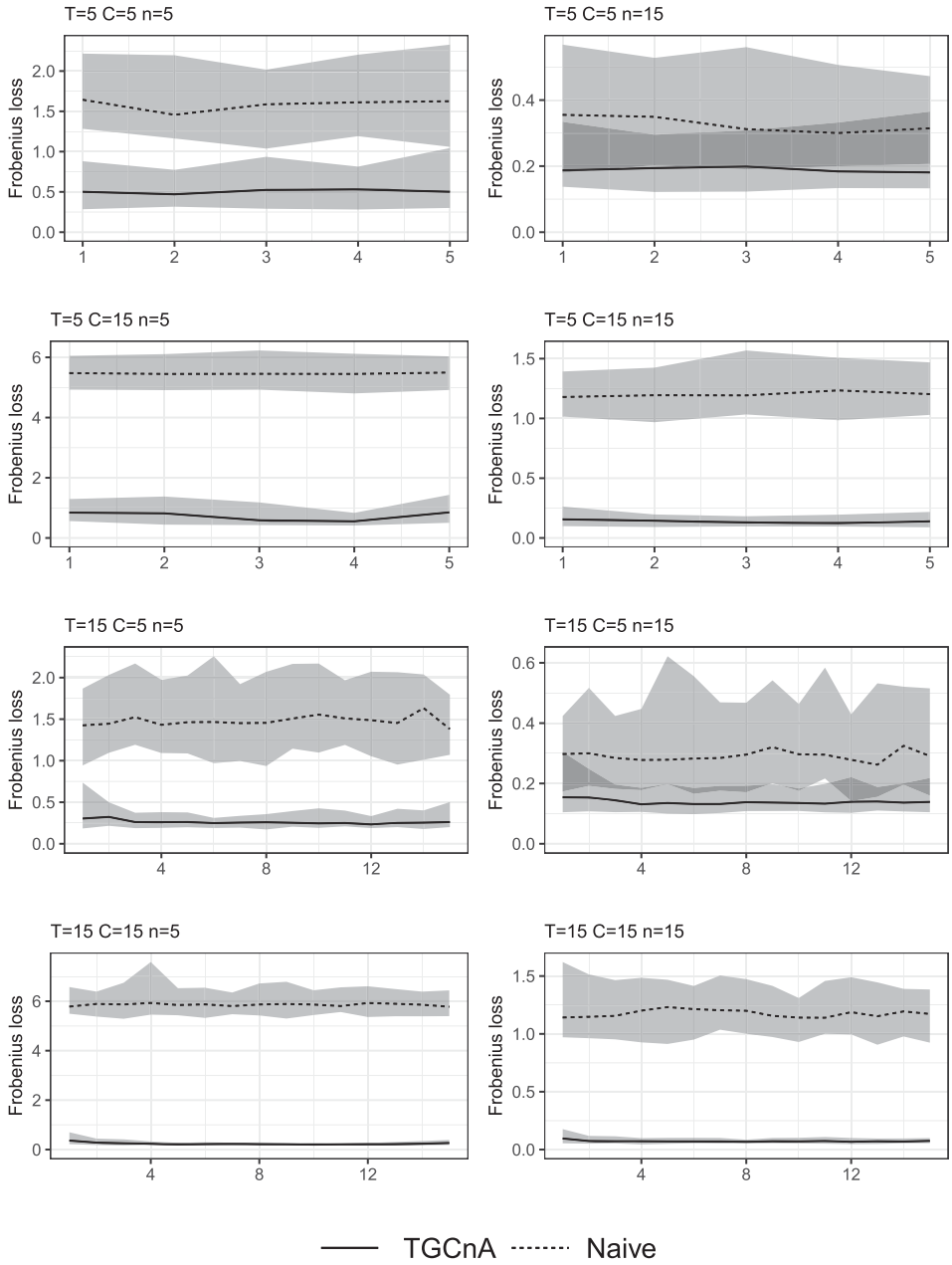

Figure 1. Mean Frobenius loss in covariance matrix estimation at each time point with 95% prediction band for simulated continuous data. The *y*-axis represents the values after being normalized by the Frobenius norm of their corresponding true covariance matrices.

be the original residual matrix after removing the low-rank component from the sample covariance at t. The elements of the sparse component matrix are then

$$\hat{R}_t(i,j) = \tilde{R}_t(i,j)(1 - \tau_{t,i,j}/|\tilde{R}_t(i,j)|)_+^{\eta}$$

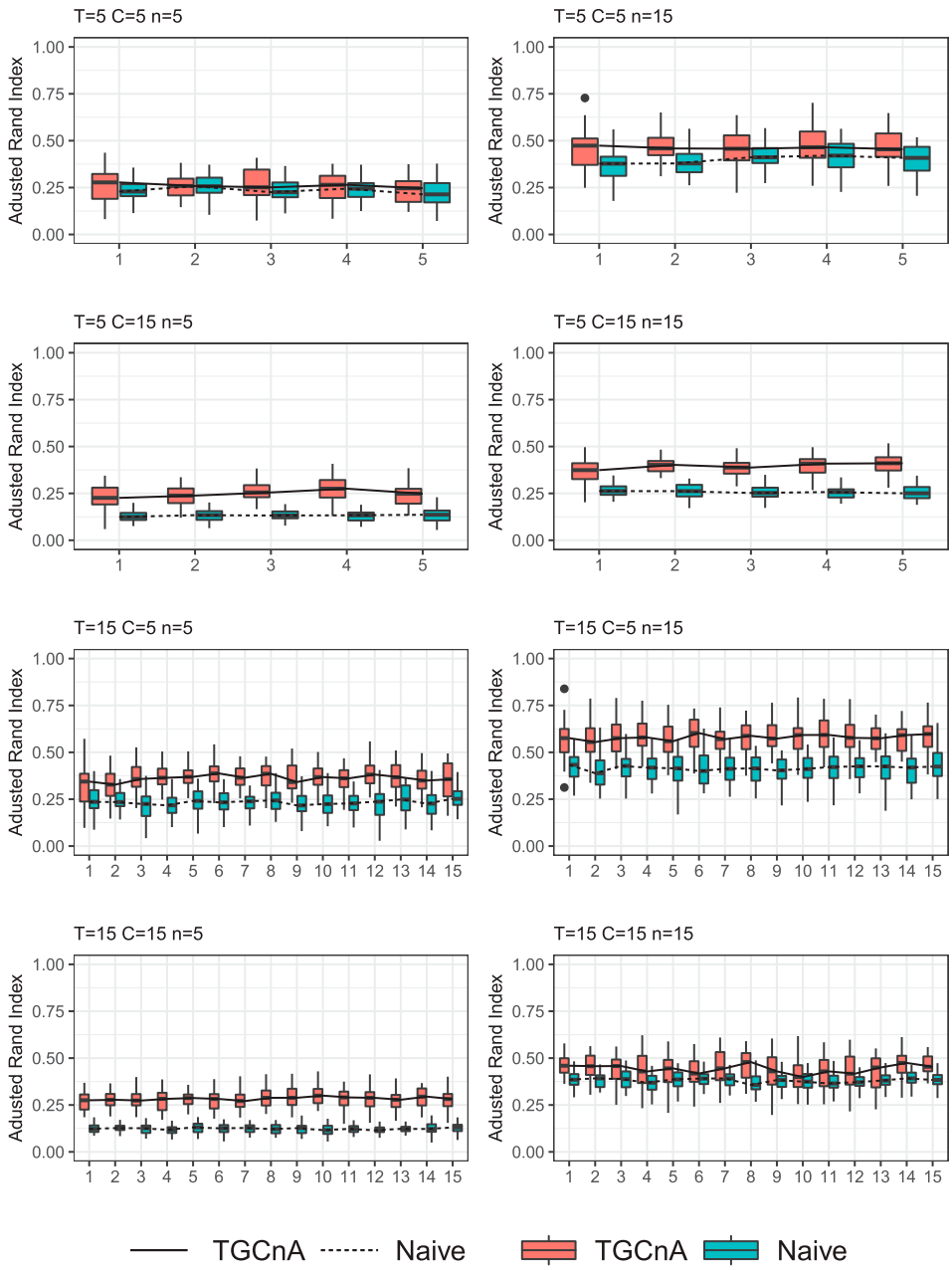

Figure 2. Adjusted Rand Index between the discovered modules and the 'true' simulated modules at each time point for simulated continuous data.

for $1 \leq i, j \leq p$ and $i \neq j$, where the adaptive threshold

$$\tau_{t,i,j} = c\sqrt{\tilde{R}_t(i,i)\tilde{R}_t(j,j)}$$

for some non-negative constant $c \in [0, 1]$ and $\eta > 0$. In our analysis, we set $c = 0.5$, and $\eta = 4$ as suggested in [6,12].

Finally, the time-specific gene-gene covariance matrices can be assembled according to Equation (1) and converted to the corresponding correlation matrices.

2.3. Implementation and computation time

We implemented TGCnA in R. The computational time increases with the number of genes and time points. For our real data analysis with more than 3000 genes and 12 time points, TGCnA only took about 7 min on a laptop computer.

2.4. Downstream analysis of TGCnA

The output of TGCnA are time-specific gene-gene correlation matrices which could serve as the input of gene co-expression network analysis procedures such as *weighted correlation network analysis (WGCNA)* [22]. WGCNA is a method for finding clusters/modules of highly correlated genes and then describing the correlation patterns among genes across different samples. In this study, we fed our outputs to WGCNA as adjacency matrices for constructing scale-free gene networks at each time point using power transformation and for module discovery by hierarchical clustering with adjacency-based dissimilarity.

In the real data analysis, we further investigated the biological interpretation of the discovered modules using R/Bioconductor package `clusterProfiler` [45]. `clusterProfiler` performs a hypergeometric test for enrichment analyses of given gene lists. We only performed the enrichment analysis of KEGG pathways instead of the other gene ontology terms, as we only analyzed the genes in KEGG pathways (see Section 3.2.1 for details).

3. Results

In the following, we applied TGCnA to both simulation data and a brain data set and compared with the static gene coexpression networks constructed separately at each time point, which is called, in this paper, 'Naive' estimate due to its independence assumption among time points. We showed that TGCnA achieves more accurate covariance matrix estimation and better-discovered gene modules.

3.1. Simulation-based evaluation

3.1.1. Simulation model

We simulated the gene expression data whose covariance matrices Σ_t were generated based on Equation (1). Let p be the number of genes. In our simulation model (see Appendix A.2 for details), C is the number of true underlying gene groups, which equals to the true number of low-rank component for our simulation model. Roughly speaking, the larger the C, the more complex the covariance matrices are. We simulated the time-invariant latent factor matrix U in a way that it encoded C potentially overlapping gene groups, and each factor was associated with a smooth weight curve. The time point-specific sparse components R_t's were generated such that they also contained the same grouping structure. The continuous transcriptomic time-course data were then generated from $N(0, \Sigma_t)$. We additionally simulated length p count vectors from correlated Poisson distributions that mimic

RNA-Seq data. In the log-scale, the expectations of these Poisson models are simulated from $N(\mu_0, \sigma_0^2 \Sigma_t)$ where Σ_t is as specified above, and $(\mu_0, \sigma_0^2) = (\log(20), 0.2^2)$ are simulation parameters that scale the simulated count data so that they resemble the range as seen in real RNA-Seq data (see Appendix A.2 for details). For the analysis with count data, we used $\log(count + 1)$ as the input of TGCnA and the Naive method. We remark that the estimated GCNs from such count data are generally not comparable with Σ_t in terms of covariance estimation loss due to their difference in physical meanings. However, they should have a similar group structure.

Let T be the total number of time points, and n the number of replicates at each time point. In our simulation studies, we fixed $p = 2000$ and considered $T = 5, 15$, $C = 5, 15$, and $n = 5, 15$. For each setting, we repeated the simulations for 40 times.

3.1.2. TGCnA improves the time point specific gene-gene covariance matrices estimation

The statistical nature of a gene coexpression network is a covariance matrix. Thus we first compared TGCnA with the naive method in terms of the Frobenius loss in covariance matrix estimation using the simulated continuous data (Figure 1). We found that TGCnA outperformed the naive methods in all simulation settings. In particular, even though both TGCnA and the naive method perform worse when the number of replicates per time point decreases (e.g. $n = 5$), the comparative advantage of TGCnA actually becomes bigger. This is because TGCnA enables the time-specific covariance matrix estimates to use the information from the other time points through the low-rank component. The thresholded sparse component of TGCnA also effectively de-noises the estimates by trimming the spurious correlations, which could explain the more significant contrast between TGCnA and the Naive method when the true covariance matrices have more complex structure (e.g. $C = 15$).

3.1.3. TGCnA achieves more accurate module discovery

One of the major goals of gene coexpression analysis is module discovery, which is essentially a clustering problem, whose performance could be measured by the Adjusted Rand Index (ARI) of the discovered modules and the true module membership. In our simulation studies, the 'groups' defined in the simulation setting are time-invariant and overlapping, and their group cohesiveness can also vary across time points. Thus it is not a suitable measure of the time-varying gene module architecture. Instead, we calculated the ARI between the identified modules (using the same clustering algorithm) from the simulated true time-specific correlation matrices, and those discovered from the corresponding estimated correlation matrices (either by TGCnA or the Naive method), the former of which were treated as the 'ground truth' as the true simulated correlations were used. Figure 2 and Appendix Figure 3 compared the modules discovered from the TGCnA and the Naive estimates of the time-specific correlation matrices in terms of their ARI's against the 'ground truth', and we found that TGCnA led to more accurate modules estimated in most settings. Similar to our observation in the comparison in covariance matrix estimation, TGCnA became more preferable when the structure of the true covariance matrices was more complex. But the number of replicates did not appear to have a clear impact on the difference between the two methods, except in the case where $T = 15$ and $C = 15$.

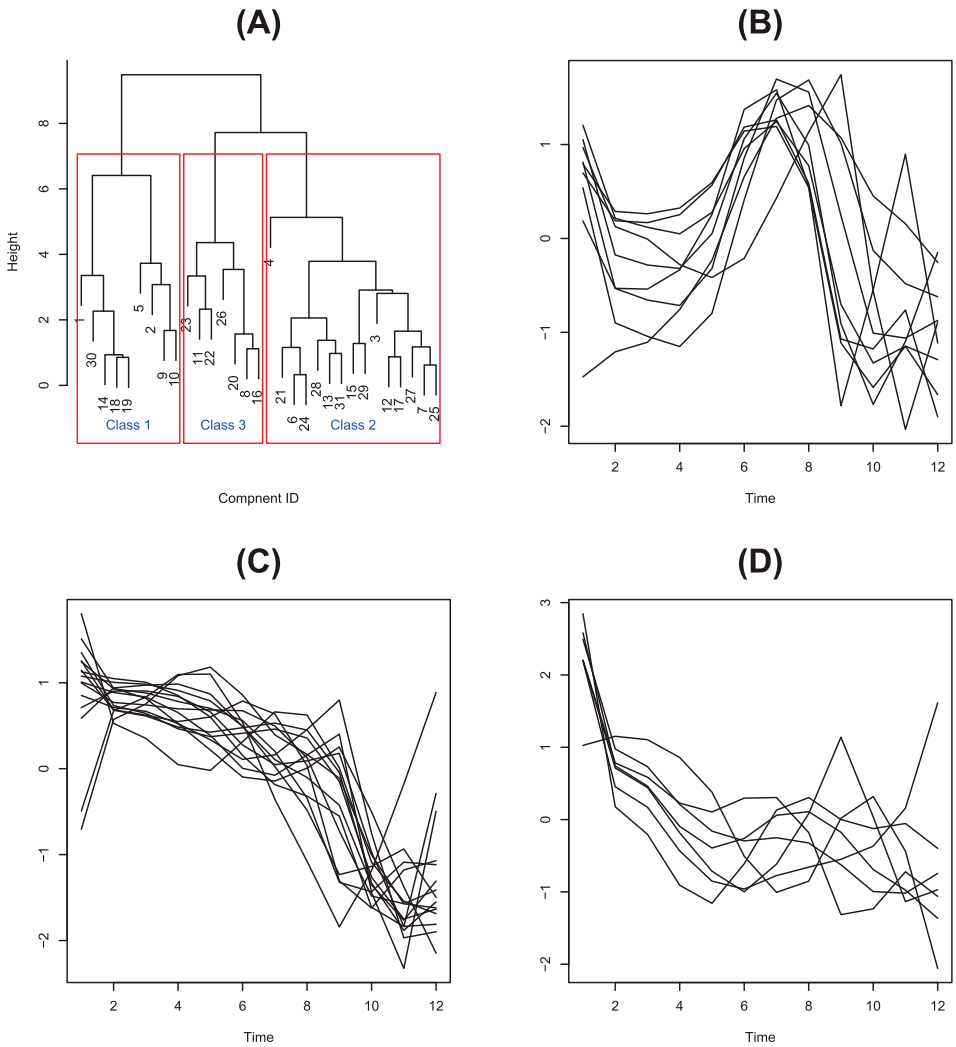

Figure 3. Hierarchical clustering of the normalized weight curves of the 31 TGCnA factors. (A) Dendrogram of the hierarchical clustering. (B)–(D) The weight curves in the three clusters, each featured (B) a peak around early infancy, (C) a monotone decreasing trend, and (D) a peak at early prenatal state, respectively.

3.2. Analysis of BrainSpan RNA-Seq data

3.2.1. Data description and preprocessing

The RNA-Seq data were the Developing Human Brain dataset obtained from the BrainSpan Atlas (http://www.brainspan.org/static/download.html). This data set consists of 524 brain samples in total, grouped into 12 developmental stages ranging from 8 post-conceptional weeks (pcw) to 40 years of age, including 6 prenatal time points and 6 time points after birth (see Appendix Table A1 for details). The number of samples in each stage ranges from 22 to 93. The expression values of this dataset were RNA-sequencing reads in the units of Per Kilobase of transcript per Million (RPKM) for 52,376 genes in total. We applied a log

transformation, $\log_2(RPKM + 1)$ on the expression values, and filtered out the genes with low variation in expression and consistently low expressions in the proceeding of development. The cutoffs were the third quartile value less than $\log_2(5)$ and the interquartile range less than $\log_2(1.5)$. Even though all genes are potentially interesting, including non-coding genes and genes without known functions, building a network with all genes will unavoidably add noise and unimportant nodes to the network. Since our goal is benchmarking GCN methods, we only analyzed the 3114 genes annotated by the KEGG pathway database [20] and filtered the rest to simplify the biological interpretations. Since our goal was covariance matrix estimation, the genes expressions were centered at each time point.

3.2.2. TGCnA extracted interpretable latent factors

Time-invariant latent factors were learned by TGCnA, and they were related to the gene group structure. The biological meanings of time-invariant latent factors were studied from all 31 extracted components for the BrainSpan data in this section.

As per its definition, a time-invariant factor k is associated with a time-varying weight curve $(d_{t_1,k}, \ldots, d_{t_T,k})$ representing the importance of this factor to the covariance matrices. Therefore, the similarity in the weight curves indicates that the corresponding factors have similar behaviors and then relevant biological functions. All the normalized curves were clustered by a hierarchical clustering method, and three metaclusters were discovered (Figure 3(A)). Each metacluster has individual characteristic: a peak around early infancy (Figure 3(B)), a monotone decreasing trend (Figure 3(C)), and a peak at early prenatal state (Figure 3(D)), respectively.

Based on KEGG pathway enrichment analysis of the top 25% most relevant genes for each metacluster (see Appendix A.3 for details), their enriched pathways were generally consistent with the overall pattern of the weight curves. For Cluster 1 (Figure 3(B)) featuring a peak around early infancy, the enriched pathways are generally relevant to the brain development of infants. For example, one enriched pathway, synaptic vesicle cycle pathway, regulates the dendritic and synaptic density, which has been shown to reach peak in infancy and early childhood and decline from 2 to 16 years [9]. The discovered pathway, PPR pathway, is associated with white-matter development and modulates brain development in preterm infants [16,21]. Cluster 2 (Figure 3(C)) had monotone decreasing weight curves and was enriched with the pathways becoming inactive during aging, such as neuroactive ligand-receptor interaction [10]. In Cluster 3 (Figure 3(D)) with a peak at the early prenatal state, enriched pathways are relevant to the embryonic brain development such as Cell adhesion molecules (CAMs), Gap junction, and Mucin type O-glycan biosynthesis [38].

3.2.3. Module discovery and annotation

Module identification and comparison are the most widely used downstream analysis of GCN, as they reveal the potential co-regulation relationship among genes. In this section, we explored the biological interpretation of the modules discovered from the time-specific coexpression networks constructed by TGCnA.

We first investigated the module conservation across time points. A new adjacency matrix was built whose edge weights were the proportion of total time points that this pair of genes were in the same module. We called a gene-gene interaction to be time-invariant if they were always in the same module. There were 1156 genes involved in such

time-invariant connections, which led to a reduced adjacency matrix. Clustering based on its associated TOM distance matrix yielded 14 modules. KEGG enrichment of these modules showed that regulation of actin cytoskeleton, protein processing in endoplasmic reticulum (ER), Ras signaling pathway, MAPK signaling pathway, and Rap1 signaling pathway were enriched ($FDR = 0.1$ for each module). The regulation of the actin cytoskeleton pathway is critical for the development of the neural system, especially for neuronal migration [35]. Endoplasmic reticulum is related to various acute disorders and degenerative diseases of the brain [30]. The Ras and MAPK signaling pathways regulate many cell functions such as cell proliferation, survival and apoptosis. Rap1 pathway is important for Neuronal Progenitor Cell Differentiation [33]. Overall, these pathways encode fundamental cell functions that are expected to have strong effects at all time points, which could explain why these genes were always connected.

Differential network analysis is a popular downstream analysis after gene network construction. Thus we designed a conservative differential analysis of the pathways enriched in modules discovered at different time points (Appendix A.4). In our study, we compared the TGCnA outputs of the 2nd (10–12 weeks prenatal) and the 11th (adolescence) time points based on this method as a showcase. It was striking to see that Huntington's disease was enriched in two modules for the adolescence stage. Symptoms of Huntington's disease usually begin between 30 and 50 years of age [42]. Our results suggested that its molecular signature could be found in transcriptomic data at an even earlier age. On the other hand, the pathway enriched at the 10–12 weeks prenatal stage was an Adherens junction, which was 'important for maintaining tissue architecture and cell polarity and can limit cell movement and proliferation' [20].

3.2.4. TGCnA yields more robust gene modules in real data analysis

We also evaluated the robustness of the TGCnA module output using the consistency between the modules discovered from the original data, and the clustering using the data after being sub-sampled or with additional noise. Specifically, we half-sampled the original

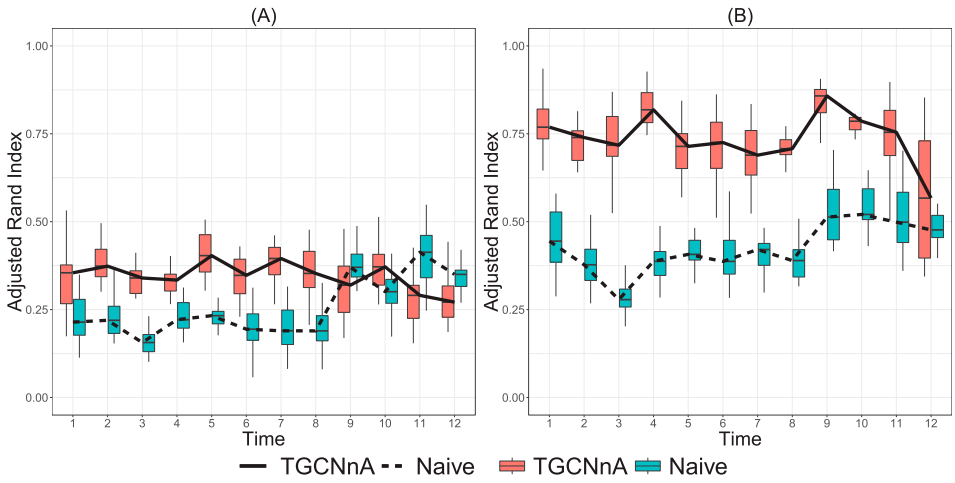

Figure 4. Adjusted Rand Index for (A) half-sampling the real data and (B) the real data-driven simulation with 10% added noise.

data at each time points and ran TGCnA, which repeated for 20 times. Figure 4(A) compared the TGCnA and the Naive method in terms of the ARI between the gene module outputs of the half-sampled data with the original data output. We found that TGCnA yielded more consistent ARI across time points, and they were higher than the results of the naive method at the majority of the time points. We remarked that the time difference between the last four times points are much larger than those between the earlier time points and the time span for each of these groups are also much wider (see Appendix Table A1), which could potentially explain the vanishing advantage of TGCnA at these time points as the information of the other time points became less useful. Nevertheless, TGCnA provided overall more robust modules. We also ran a real data-driven simulation by adding white noise to the data. The standard deviation of the added white noise for each gene at each time point is 0.1 times the standard deviation of its expression. In this analysis, we again found that the ARI of TGCnA output is higher than that of the naive estimate (Figure 4(B)).

4. Conclusion and discussion

Both of gene coexpression network analysis and temporal transcriptomic studies have been widely used. There has not been any appropriate and computationally feasible time-specific GCN inference from time-course data. Most of the existing studies either model the GCN at each time point completely separately, or pool the data across time points to build one single network. In this paper, we proposed the Temporal Gene Coexpression Network Analysis (TGCnA) that jointly model the temporal transcriptomic data when the samples at different time points are potentially from different subjects. The outputs of TGCnA are time-specific gene-gene correlation matrices, which allow the users to perform various downstream analysis flexibly with other computational tools such as WGCNA. Using both simulation and real data examples, we have shown that TGCnA achieves more accurate correlation matrix estimation and more robust module identification.

The statistical nature of TGCnA is a 'low-rank plus sparse' estimator of the covariance matrices, and it could be viewed as an extension of the Principal Orthogonal Complement Thresholding (POET, [12]) method. While POET focused on one single covariance matrix, TGCnA jointly estimates multiple covariance matrices simultaneously under the structural assumption that their low-rank components share the same factors.

One issue with TGCnA, along with other correlation-based network models, is the interpretability of the network links, as there are many biological and technical factors that may contribute to the empirical gene-gene correlations. Besides improving the data pre-processing to reduce the impact of the undesirable confounders, the interpretability can also be potentially improved via data integration [7]. Thus we will explore the incorporation of the epigenetic data, metabolic pathway, gene oncology and protein-protein interactions in our modeling framework. The input samples from different points should be under similar conditions except for the time effect. We will also explore TGCnA's potential in the meta-analysis of the time-course samples from different studies by studying proper preprocessing procedures to remove the study-specific factors. In this paper, we focused on the analysis of the coexpression networks themselves. Many other works in the literature attempted to incorporate the coexpression information in clustering of the gene expression curves [18,26]. But they usually considered the coexpression across time

points for the same gene, and rarely explicitly modeled the gene-gene correlations as in TGCnA. Another potential future research direction would be utilizing TGCnA to improve the clustering of the mean gene expression curves.

Acknowledgements

We thank the Holland Computing Center (HCC) at UNL for computation resources and technical supports. A previous version of this paper is available on bioRxiv [24].

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work has been supported by NSF ABI (Division of Biological Infrastructure) (Award# DBI-1564621), NSF EPSCoR (RII) Track II (Office of Integrative Activities) (Award# OIA-1736192) and NU Collaborative System Science Seed Grant to CZ and QZ.

ORCID

Qi Zhang http://orcid.org/0000-0001-6197-0973

References

[1] H. Akaike, *Information theory and an extension of the maximum likelihood principle*, in *Selected Papers of Hirotugu Akaike*, E. Parzen, K. Tanabe, and G. Kitagawa, eds. Springer, New York, 1998, pp. 199–213.

[2] H. Akaike, *Maximum likelihood identification of Gaussian autoregressive moving average models*, Biometrika 60 (1973), pp. 255–265.

[3] H. Azevedo, N.A. Khaled, P. Santos, F.B. Bertonha, and C.A. Moreira-Filho, *Temporal analysis of hippocampal CA3 gene coexpression networks in a rat model of febrile seizures*, Dis. Model. Mech. 11 (2018), p. dmm02907.

[4] S. Ballouz, W. Verleyen, and J. Gillis, *Guidance for RNA-seq co-expression network construction and analysis: Safety in numbers*, Bioinformatics 31 (2015), pp. 2123–2130.

[5] M.J. Beal, F. Falciani, Z. Ghahramani, C. Rangel, and D.L. Wild, *A Bayesian approach to reconstructing genetic regulatory networks with hidden factors*, Bioinformatics 21 (2004), pp. 349–356.

[6] T. Cai and W. Liu, *Adaptive thresholding for sparse covariance matrix estimation*, J. Am. Stat. Assoc. 106 (2011), pp. 672–684.

[7] X. Chen, J. Gu, X. Wang, J.G. Jung, T.L. Wang, L. Hilakivi-Clarke, R. Clarke, and J. Xuan, *CRNET: An efficient sampling approach to infer functional regulatory networks by integrating large-scale chip-seq and time-course rna-seq data*, Bioinformatics 34 (2017), p. btx827.

[8] C.O. Daub, R. Steuer, J. Selbig, and S. Kloska, *Estimating mutual information using b-spline functions – an improved similarity measure for analysing gene expression data*, BMC Bioinf. 5 (2004), p. 118.

[9] P. Dehm, S.A. Jimenez, B.R. Olsen, and D.J. Prockop, *A transport form of collagen from embryonic tendon: Electron microscopic demonstration of an NH2-terminal extension and evidence suggesting the presence of cystine in the molecule*, Proc. Natl. Acad. Sci. 69 (1972), pp. 60–64.

[10] H.M. Dönertaş, H. İzgi, A. Kamacıoğlu, Z. He, P. Khaitovich, and M. Somel, *Gene expression reversal toward pre-adult levels in the aging human brain and age-related loss of cellular identity*, Sci. Rep. 7 (2017), p. 5894.

[11] J. Fan, Y. Fan, and J. Lv, *High dimensional covariance matrix estimation using a factor model*, J. Econom. 147 (2008), pp. 186–197.
[12] J. Fan, Y. Liao, and M. Mincheva, *Large covariance estimation by thresholding principal orthogonal complements*, J. R. Stat. Soc. Ser. B (Stat. Methodol.) 75 (2013), pp. 603–680.
[13] S.P. Ficklin, F. Luo, and F.A. Feltus, *The association of multiple interacting genes with specific phenotypes in rice using gene coexpression networks*, Plant Physiol. 154 (2010), pp. 13–24.
[14] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, *Using Bayesian networks to analyze expression data*, J. Comput. Biol. 7 (2000), pp. 601–620.
[15] C. Gao, I.C. McDowell, S. Zhao, C.D. Brown, and B.E. Engelhardt, *Context specific and differential gene co-expression networks via Bayesian biclustering*, PLoS Comput. Biol. 12 (2016), p. e1004791.
[16] J.H. Gilmore, R.C. Knickmeyer, and W. Gao, *Imaging structural and functional brain development in early childhood*, Nat. Rev. Neurosci. 19 (2018), p. 123.
[17] M.J. Hawrylycz, E.S. Lein, A.L. Guillozet-Bongaarts, E.H. Shen, L. Ng, J.A. Miller, L.N. van de Lagemaat, K.A. Smith, A. Ebbert, Z.L. Riley, C. Abajian, C.F. Beckmann, A. Bernard, D. Bertagnolli, A.F. Boe, P.M. Cartagena, M.M. Chakravarty, M. Chapin, J. Chong, R.A. Dalley, B.D. Daly, C. Dang, S. Datta, N. Dee, T.A. Dolbeare, V. Faber, D. Feng, D.R. Fowler, J. Goldy, B.W. Gregor, Z. Haradon, D.R. Haynor, J.G. Hohmann, S. Horvath, R.E. Howard, A. Jeromin, J.M. Jochim, M. Kinnunen, C. Lau, E.T. Lazarz, C. Lee, T.A. Lemon, L. Li, Y. Li, J.A. Morris, C.C. Overly, P.D. Parker, S.E. Parry, M. Reding, J.J. Royall, J. Schulkin, P.A. Sequeira, C.R. Slaughterbeck, S.C. Smith, A.J. Sodt, S.M. Sunkin, B.E. Swanson, M.P. Vawter, D. Williams, P. Wohnoutka, H.R. Zielke, D.H. Geschwind, P.R. Hof, S.M. Smith, C. Koch, S.G.N. Grant, and A.R. Jones, *An anatomically comprehensive atlas of the adult human brain transcriptome*, Nature 489 (2012), p. 391.
[18] W. Huang, X. Cao, and S. Zhong, *Network-based comparison of temporal gene expression patterns*, Bioinformatics 26 (2010), pp. 2944–2951.
[19] H.N. Kadarmideen and N.S. Watson-Haigh, *Building gene co-expression networks using transcriptomics data for systems biology investigations: Comparison of methods using microarray data*, Bioinformation 8 (2012), p. 855.
[20] M. Kanehisa and S. Goto, *KEGG: Kyoto encyclopedia of genes and genomes*, Nucleic Acids Res. 28 (2000), pp. 27–30.
[21] M.L. Krishnan, Z. Wang, P. Aljabar, G. Ball, G. Mirza, A. Saxena, S.J. Counsell, J.V. Hajnal, G. Montana, and A.D. Edwards, *Machine learning shows association between genetic variability in pparg and cerebral connectivity in preterm infants*, Proc. Natl. Acad. Sci. 114 (2017), p. 201704907.
[22] P. Langfelder and S. Horvath, *WGCNA: An R package for weighted correlation network analysis*, BMC Bioinf. 9 (2008), p. 559.
[23] S. Lebre, J. Becq, F. Devaux, M.P. Stumpf, and G. Lelandais, *Statistical inference of the time-varying structure of gene-regulation networks*, BMC Syst. Biol. 4 (2010), p. 130.
[24] J. Li, Y. Lai, C. Zhang, and Q. Zhang, *Temporal gene coexpression network analysis using a low-rank plus sparse framework*, bioRxiv (2018), p. 359612.
[25] X. Luo, *High dimensional low rank and sparse covariance matrix estimation via convex minimization*, preprint (2011).
[26] P. Ma, C.I. Castillo-Davis, W. Zhong, and J.S. Liu, *A data-driven clustering method for time course gene expression data*, Nucleic Acids Res. 34 (2006), pp. 1261–1269.
[27] Z. Miao, Z. Han, T. Zhang, S. Chen, and C. Ma, *A systems approach to a spatio-temporal understanding of the drought stress response in maize*, Sci. Rep. 7 (2017), p. 6590.
[28] Y. Ni, F.C. Stingo, and V. Baladandayuthapani, *Bayesian nonlinear model selection for gene regulatory networks*, Biometrics 71 (2015), pp. 585–595.
[29] T. Obayashi, S. Hayashi, M. Shibaoka, M. Saeki, H. Ohta, and K. Kinoshita, *COXPRESdb: A database of coexpressed gene networks in mammals*, Nucleic Acids Res. 36 (2007), pp. D77–D82.
[30] W. Paschen, *Endoplasmic reticulum: A primary target in various acute disorders and degenerative diseases of the brain*, Cell Calcium 34 (2003), pp. 365–383.
[31] J. Rissanen, *Modeling by shortest data description*, Automatica 14 (1978), pp. 465–471.

[32] J. Rissanen, *A universal prior for integers and estimation by minimum description length*, Ann. Stat. (1983), pp. 416–431.
[33] D. Rueda, B. Navarro, A. Martínez-Serrano, M. Guzmán, and I. Galve-Roperh, *The endocannabinoid anandamide inhibits neuronal progenitor cell differentiation through attenuation of the Rap1/B-Raf/ERK pathway*, J. Biol. Chem. 277 (2002), pp. 46645–46650.
[34] J. Schäfer and K. Strimmer, *An empirical Bayes approach to inferring large-scale gene association networks*, Bioinformatics 21 (2004), pp. 754–764.
[35] D.J. Solecki, N. Trivedi, E.E. Govek, R.A. Kerekes, S.S. Gleason, and M.E. Hatten, *Myosin II motors and F-actin dynamics drive the coordinated movement of the centrosome and soma during CNS glial-guided neuronal migration*, Neuron 63 (2009), pp. 63–80.
[36] L. Song, P. Langfelder, and S. Horvath, *Comparison of co-expression measures: Mutual information, correlation, and model based indices*, BMC Bioinf. 13 (2012), p. 328.
[37] M. Tan, D. Cheng, Y. Yang, G. Zhang, M. Qin, J. Chen, Y. Chen, and M. Jiang, *Co-expression network analysis of the transcriptomes of rice roots exposed to various cadmium stresses reveals universal cadmium-responsive genes*, BMC Plant Biol. 17 (2017), p. 194.
[38] D.T. Tran and K.G. Ten Hagen, *Mucin-type O-glycosylation during development*, J. Biol. Chem. 288 (2013), pp. 6921–6929.
[39] S. Valle, W. Li, and S.J. Qin, *Selection of the number of principal components: The variance of the reconstruction error criterion with a comparison to other methods*, Ind. Eng. Chem. Res. 38 (1999), pp. 4389–4401.
[40] S. van Dam, T. Craig, and J.P. de Magalhaes, *GeneFriends: A human RNA-seq-based gene and transcript co-expression database*, Nucleic Acids Res. 43 (2014), pp. D1124–D1132.
[41] N. Villa-Vialaneix, L. Liaubet, T. Laurent, P. Cherel, A. Gamot, and M. SanCristobal, *The structure of a gene co-expression network reveals biological functions underlying eQTLs*, PLoS ONE 8 (2013), p. e60045.
[42] F.O. Walker, *Huntington's disease*, Lancet 369 (2007), pp. 218–228.
[43] Y. Yang, L. Han, Y. Yuan, J. Li, N. Hei, and H. Liang, *Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types*, Nat. Commun. 5 (2014), p. 3231.
[44] W.C. Yim, Y. Yu, K. Song, C.S. Jang, and B.M. Lee, *PLANEX: The plant co-expression database*, BMC Plant Biol. 13 (2013), p. 83.
[45] G. Yu, L.G. Wang, Y. Han, and Q.Y. He, *clusterProfiler: An R package for comparing biological themes among gene clusters*, OMICS: J. Integr. Biol. 16 (2012), pp. 284–287.
[46] B. Zhang and S. Horvath, *A general framework for weighted gene co-expression network analysis*, Stat. Appl. Genet. Mol. Biol. 4 (2005), p. 17.
[47] Q. Zhang, Y. Yu, J. Zhang, and H. Liang, *Using single-index odes to study dynamic gene regulatory network*, PLoS ONE 13 (2018), p. e0192833.
[48] S. Zhu and Y. Wang, *Hidden Markov induced dynamic Bayesian network for recovering time evolving gene regulatory networks*, Sci. Rep. 5 (2015), p. 17841.

Appendices

Appendix 1. Supplementary notes

A.1 AIC and MDL for selecting K

When the number of principal components taking the value of $k = 1, \ldots, m$, AIC and MDL have the form of

$$\mathrm{AIC}(k) = -2n(m-k)\log\rho(k) + 2k(2m-k),$$

$$\mathrm{MDL}(k) = -n(m-k)\log\rho(k) + \frac{k}{2}(2m-k)\log n,$$

respectively, where $m = N - T$ is the total number of samples $N = nT$ subtracted by T, the number of time points. The internal function $\rho(k)$ is

$$\rho(k) = \frac{(l_{k+1} l_{k+2}, \cdots, l_m)^{1/(m-k)}}{\frac{1}{m-k}(l_{k+1} l_{k+2}, \cdots, l_m)},$$

where $l_k = \hat{\sigma}^2 + d_k^2$ represents the corresponding eigenvalues, and d_k is the kth singular value, the value of $\hat{\sigma}^2$ is estimated as the mean of diagonal elements of the covariance matrix after subtracting the low-rank component UDU^{T}, where D is the $K \times K$ diagonal loading matrix.

A.2 Simulation model

We simulated gene expression data whose covariance matrices Σ_t were generated based on Equation (1).

Let p be the number of genes and C the number of underlying gene groups. We first simulated the $p \times C$ time-invariant latent factor matrix U as the following. As we have discussed, our model was motivated by the observation that the genes usually belong to certain functional groups with time-varying effects, and these groups may overlap. Thus we first simulated a $p \times C$ binary group membership matrix S, where $S(j, g) = 1\{\text{gene } j \text{belong to group } g\}$. For each group, we picked a random integer between $[1/C, 2/C]$ as the group size, and its members are randomly selected. Note that its rows could contain more than one non-zero elements as the groups could overlap. Then we used the left singular vectors of $S + 0.01A$ as the latent factors U, where A is a $p \times C$ random matrix whose entries are i.i.d. from $Unif(-1, 1)$.

Next, we simulated the time-varying weights of these latent factors. For $t = 1, \ldots, T$, let $d_{g,t}$ be the gth diagonal element of D_t. We defined $(\log(d_{g,1}), \ldots, \log(d_{g,T}))$ as a random linear combination of B-spline basis defined on $(0, T + 1)$, whose coefficients were i.i.d. samples from $Unif(-1, 1)$. Then $L_t = UD_tU^{\mathrm{T}}$ is the low-rank component of the simulated covariance matrix.

The most naive way of simulating the sparse component R_t is simply generating a sparse random symmetric matrix. But such matrix does not contain any information about the structure. To generate an informative sparse component, we defined the elements of the upper triangle of the symmetric matrix R_t as the following:

$$R_t(i, j) = b_{t,i,j} \cdot e_{t,i,j} \cdot sign\left(L_t(i, j)\right)$$

where $b_{t,i,j} \overset{iid}{\sim} Unif(0.1, 0.3)$, $e_{t,i,j} \overset{iid}{\sim} Bernoulli(p_{t,i,j})$ with the probability $p_{t,i,j} = 0.005 \cdot 1\{|L_t(i, j)| \geq q_t\} + 0.0005 \cdot 1\{|L_t(i, j)| < q_t\}$, and the diagonal elements of R_t were set to 1. Here $b_{i,j}$ modeled the magnitude of the time-point specific sparse component, and the definition of $p_{t,i,j}$ ensured that R_t and L_t contained non-contradicting information about the underlying structure. q_t were set as the 0.9 quantile of the absolute values of the off-diagonal elements of L_t in out simulation.

Finally, we defined the time-specific covariance matrix $\Sigma_t = UD_tU^{\mathrm{T}} + R_t$ and simulated $X_{t,r}$ from $N(0, \Sigma_t)$ for $r = 1, \ldots, n$.

We also additionally simulated length p count vectors $\tilde{X}_{t,r}$ from Poisson distributions with expectations $20 \exp(0.2X_{t,r})$, so that the simulated counts are correlated through Σ_t, the covariance matrix of their expectations in log scale. Σ_t is as specified for simulated continuous data. The scaling parameters 20 and 0.2 in the transformation are chosen so that the output resemble the range as seen in real RNA-Seq data.

In our simulation studies, we fixed $p = 2000$ and considered $T = 5, 15$, $C = 5, 15$, and $n = 5, 15$. For each setting, the simulations were repeated for 40 times.

A.3 Identifying the most relevant genes for latent factors

According to Equation (1), for each pair of genes (i, j),

$$\hat{\Sigma}_t(i, j) = \sum_{k=1}^{K} D_t(k, k) U(i, k) U(j, k) + \hat{R}_t(i, j)$$

The contribution of one single factor k to the correlation pattern among gene i and other genes could be measured by $|U(i,k)|$, because factor k becomes irrelevant to the covariances involving gene i at all time points if $|U(i,k)|$ is close to 0. Following a similar spirit, the joint relevance of a group of factors $k_1, \ldots, k_A$ to the covariances involving a particular gene i could be measured by $\ell_i = \sum_{a=1}^{A} U(i,k_a)^2$. It is essentially the L_2 row norm of the corresponding submatrix of U. For a group of factors, we said a gene i was more relevant to these factors if ℓ_i is large. For our KEGG enrichment analysis, we considered the top 25% genes with the largest ℓ_i for each cluster of latent factors.

A.4 Conservative differential enrichment analysis

In a pairwise comparison of time points t_1 and t_2, a pathway is said to be specific to t_1 if it satisfies the following conditions. (1) It is enriched for some modules at t_1 but not at t_2 nor for the above time-invariant modules; and (2) at most 25% of the genes that are involved in this pathway are also in any of the pathways enriched at t_2 or in the time-invariant modules. The second criterion is to avoid the case where two pathways shared a large proportion of genes, but were enriched at different time points due to their minor differences in gene composition and the statistical cutoff in enrichment analysis.

Appendix 2. Supplementary figures and tables

Table A1. Description of developmental stages for the Brainspan data available at http://www.brainspan.org/static/download.html.

Stage	Age	Developmental stage	Replicates number
1	8–9 pcw	Early prenatal	30
2	10–12 pcw	Early prenatal	45
3	13–15 pcw	Early mid-prenatal	44
4	16–18 pcw	Early mid-prenatal	53
5	19–24 pcw	Late mid-prenatal	43
6	25–38 pcw	Late prenatal	22
7	Birth–5 months	Early infancy	33
8	6–18 months	Late infancy	26
9	19 months–5 years	Early childhood	44
10	6–11 years	Late childhood	41
11	12–19 years	Adolescence	50
12	20–40 years	Adulthood	93

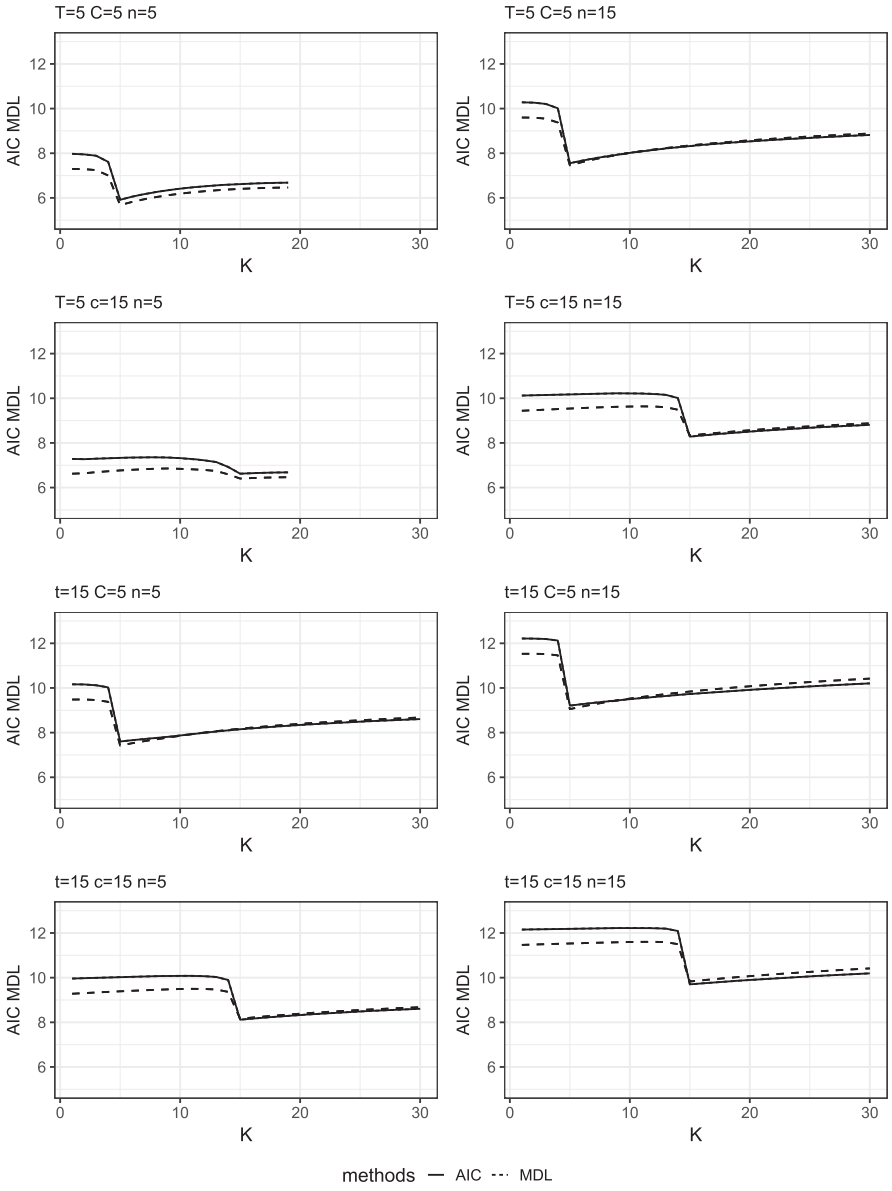

Figure A1. Selecting the number of latent factors using AIC and MDL for the simulation studies.

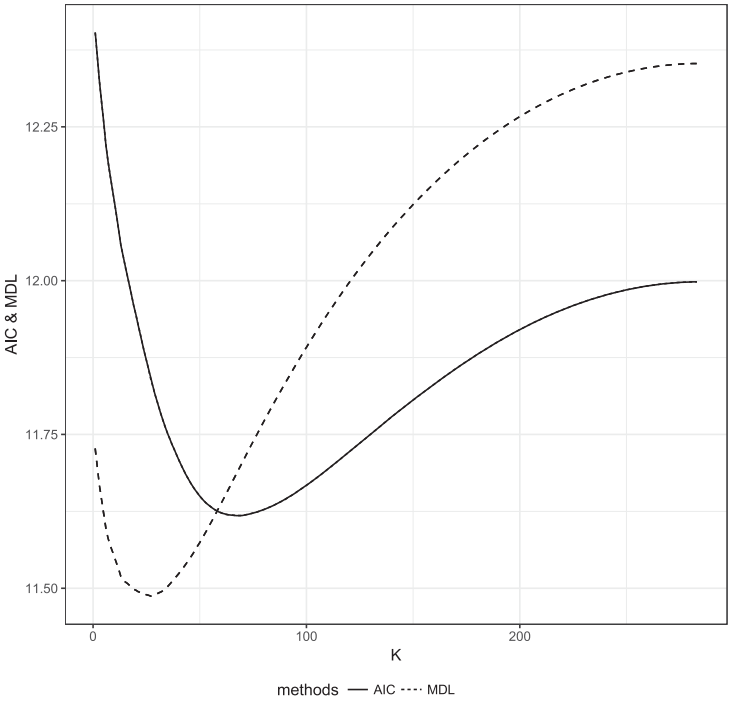

Figure A2. Selecting the number of latent factors using AIC and MDL for the real data analysis.

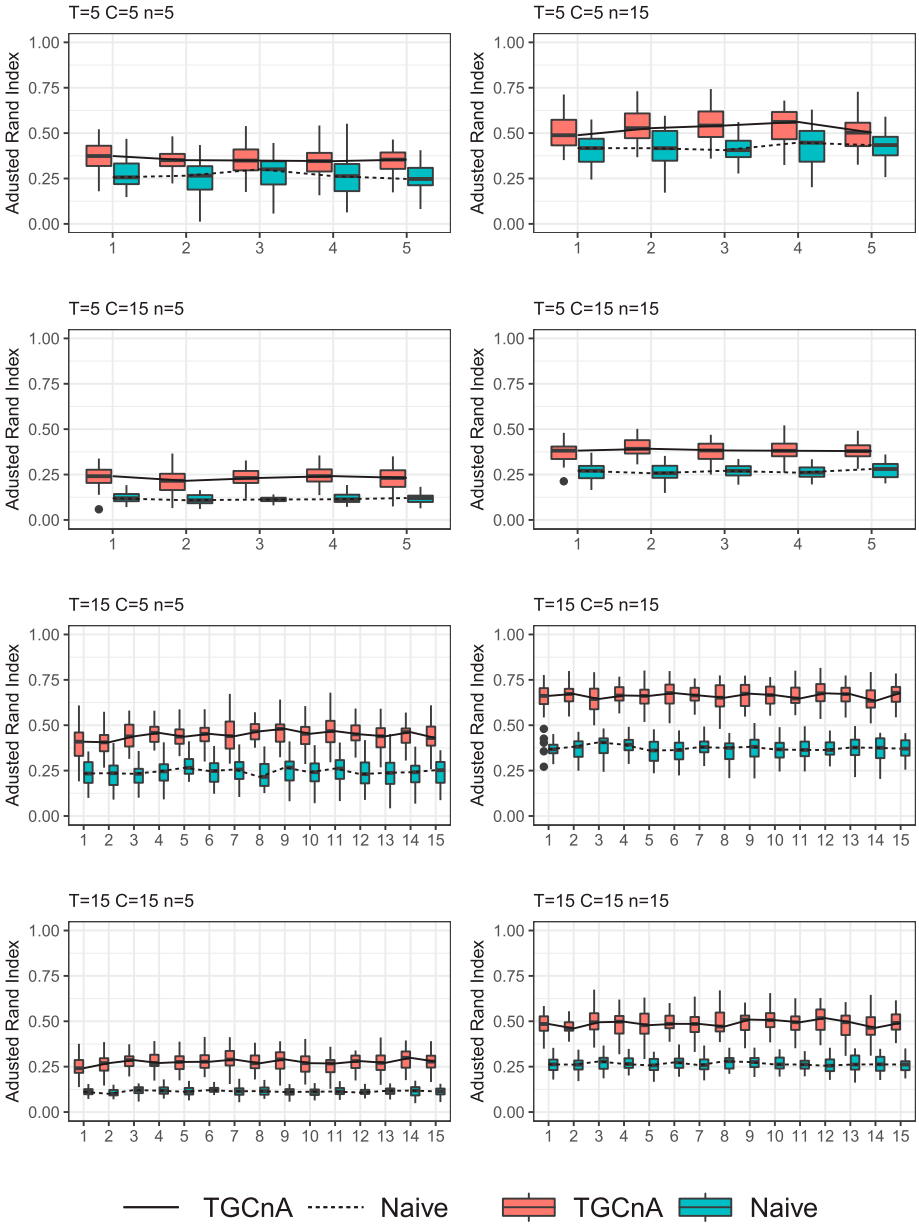

Figure A3. Adjusted Rand Index between the discovered modules and the 'true' simulated modules at each time point for simulated count data.