Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Network-Based Neighborhood Regression

Yaoming Zhen & Jin-Hong Du

To cite this article: Yaoming Zhen & Jin-Hong Du (22 May 2025): Network-Based Neighborhood Regression, Journal of the American Statistical Association, DOI: 10.1080/01621459.2025.2485342

To link to this article: https://doi.org/10.1080/01621459.2025.2485342

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2025, VOL. 00, NO. 0, 1–14: Theory and Methods
https://doi.org/10.1080/01621459.2025.2485342

Network-Based Neighborhood Regression

Yaoming Zhen[a] and Jin-Hong Du[b,c]

[a]Department of Statistical Sciences, University of Toronto, Toronto, Ontario, Canada; [b]Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA; [c]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA

ABSTRACT

Given the ubiquity of modularity in biological systems, module-level regulation analysis is vital for understanding biological systems across various levels and their dynamics. Current statistical analysis on biological modules predominantly focuses on either detecting the functional modules in biological networks or subgroup regression on the biological features without using the network data. This article proposes a novel network-based neighborhood regression framework whose regression functions depend on both the *global* community-level information and *local* connectivity structures among entities. An efficient community-wise least square optimization approach is developed to uncover the strength of regulation among the network modules while enabling asymptotic inference. With random graph theory, we derive non-asymptotic estimation error bounds for the proposed estimator, achieving exact minimax optimality. Unlike the root-n consistency typical in canonical linear regression, our model exhibits linear consistency in the number of nodes n, highlighting the advantage of incorporating neighborhood information. The effectiveness of the proposed framework is further supported by extensive numerical experiments. Application to whole-exome sequencing and RNA-sequencing Autism datasets demonstrates the usage of the proposed method in identifying the association between the gene modules of genetic variations and the gene modules of genomic differential expressions. Supplementary materials for this article are available online, including a standardized description of the materials available for reproducing the work.

ARTICLE HISTORY
Received June 2024
Accepted March 2025

KEYWORDS
Autism spectrum disorder; Gene co-expressions; Neighborhood regression; Network data; Stochastic block model

1. Introduction

In various biological systems, it is more than common for biological units to interplay with each other and form functional modules, such as in the gene co-expression networks (Liu, Lei, and Roeder 2015), protein-protein interaction networks (Brohee and Van Helden 2006), and functional connectivities in brain regions (Zhang, Sun, and Li 2020; Paul and Chen 2020). Measurements of a single biological unit depend not only on its own features but also on those of other units it interacts with. Understanding the module-level regulation relationships could provide crucial insights into the biological development processes. It is thus of scientific interest to investigate the evolution of the biological units while incorporating their local neighborhood information (Zhang, Levina, and Zhu 2017) and global cluster-level relationships (Le and Li 2022) into a unified framework.

Our motivation is an study of genetic and genomic associations related to autism spectrum disorder (ASD). Characterized by compromised social interactions and repetitive behaviors, ASD is significantly influenced by genetic variation, which is usually quantified as the genetic risk (GR) scores computed from the whole exome sequencing datasets (Liu et al. 2013). Although there are typically thousands of genes, the genetic evidence indicated by the GR scores is scarce. Using hidden Markov random field (HMRF) models, Liu et al. (2014) and Liu, Lei, and Roeder (2015) incorporated gene co-expression networks to identify clusters of autism risk genes. Recently, Gandal et al. (2022) study the genomic differentially expressed (DE) scores by contrasting the gene expressions between ASD and neurotypical individuals. When comparing the GR scores in previous studies with DE scores, only a small portion of overlaps is observed, and the interplay between genetic evidence and genomic evidence remains unknown. Further, the evolution of a gene's expression levels shall relate not only to its neighboring genes but also to the functional module in which it is located. Therefore, a better modeling strategy is desired to quantify the directional causal effect from the genetic evidence to the genomics evidence while incorporating both the neighborhood and community information.

Existing methods, such as the random effects and subgroup effects models, only provide information about heterogeneity within each community but lack the capability to model inter-community interactions. Additionally, approaches like network-assisted regression proposed by Li, Levina, and Zhu (2019) and Le and Li (2022) do not directly incorporate neighborhood information or account for the heterogeneity of regression coefficients across different communities. Although these methods attempt to leverage network data, they fall short of comprehensively modeling the complex dependencies that exist among samples and network modules.

To bridge these gaps, we propose a novel network-based neighborhood regression model that predicts the response of a node based on the covariates of all nodes within its

CONTACT Yaoming Zhen *yaoming.zhen@utoronto.ca* Department of Statistical Sciences, University of Toronto, Toronto, Ontario, M5G 1X6, Canada.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

neighborhood. The significant challenge lies in addressing the dependency among samples within the network, leading to potential overparameterization issues. To mitigate this, our model employs a block structure in the neighborhood regression coefficient matrix to reflect community-wise common effects, making estimation and inference feasible. We demonstrate that the community-wise least squares objective function can be decomposed into multiple nonoverlapping linear regression objective functions, which allows for efficient estimation and inference despite the complexities posed by network data with community-wise interactions.

The main contribution of this article is 3-fold: (a) Aiming to better understand the directional effect of the autism genetic factors on their differential expressions, the proposed network-based neighborhood regression framework incorporates not only the *local* connectivity patterns of the genes but also their *global* community-wise common effects. (b) Theoretically, we develop random-design and non-asymptotic analyses for the network-based design matrix to derive concentration behavior for the Hessian matrix of neighborhood regression, which further leads to the asymptotic consistency of the proposed community-wise least square estimator. Most importantly, our theory, along with minimax optimality, suggests the blessing of neighborhood information aggregation, yielding that the convergence rate of the neighborhood regression coefficients is almost linear in the number of nodes if the network is dense enough. This finding substantially distinguishes from the root-n consistency in canonical linear regression setups, highlighting the potential of leveraging neighborhood information in network-based regression models. We further study the influence of misspecified community memberships. (c) Simulation studies showcase the feasibility and necessity of the proposed method, and application to the Autism gene datasets identifies interpretable community-wise common effects among the genes under investigation.

1.1. Other Applications

While this article primarily addresses biomedical applications, other potential applications of the proposed network-based neighborhood regression model are also worth considering.

- In social networks, such as Twitter, a person's retweet behavior depends not only on how actively she/he posts original tweets but also on how actively her/his friends post original tweets. People of the same age and occupation group tend to share similar tweets and exhibit similar retweet patterns.
- In the city transportation network, the number of arrival rides at any station depends not only on the number of departure rides from this station but also on the number of departure rides from its neighboring stations. Stations in urban communities or suburban communities have similar arrival and departure patterns.
- In the world trading network, the among of goods that a country imports depends on the among of goods it exports and those of its neighboring countries export. Trading patterns in developed countries or developing countries can be similar.

In these examples, predicting individual response based on the covariates in its neighborhood and community necessitates a new regression method.

1.2. Related Work

Statistical analysis of biological data with module structure primarily focuses on sub-group identification, such as community detection and sub-group regression. Common community detection approaches include likelihood-based approaches under stochastic block model (Celisse, Daudin, and Pierre 2012), latent space model (Raftery et al. 2012; Zhang, Xu, and Zhu 2022), and random dot graph model (Athreya et al. 2018), spectral clustering under stochastic block model and degree-corrected stochastic block model (Jin 2015; Lei and Rinaldo 2015), and modularity maximization (Shang et al. 2013). For details, we refer interested readers to the comprehensive review papers Abbe (2018) and Gao and Ma (2021). Besides, there has been a notable shift in research focusing on integrating network structure and node attributes to identify communities more accurately. Related works include (Newman and Clauset 2016; Zhang, Levina, and Zhu 2016; Yan and Sarkar 2021; Xu, Zhen, and Wang 2023) and Hu and Wang (2024).

In the research line of sub-group regression, Zhou et al. (2022) and Wang et al. (2023) propose sub-group regression models to analyze personal treatment effects and low-dimensional latent factors, respectively, without using network information. However, using the network information in predictive models has not yet been well-studied. Recently, Li, Levina, and Zhu (2019) study linear regression with network cohesion regularizer on the individual node effects; Le and Li (2022) further extend it to a semi-parametric regression model by incorporating network spectral information. However, neither method directly incorporates node-wise neighborhood information and the heterogeneity of regression coefficients in different communities.

1.3. Notations

Denote $[n] = \{1, \ldots, n\}$ and $\mathbb{1}(A) \in \{0, 1\}$ as the indicator function for any event A. Let $\mathbf{0}_n, \mathbf{1}_n \in \mathbb{R}^n$ be the vectors of all zeros and ones, respectively, and $\boldsymbol{I}_n \in \{0, 1\}^{n\times n}$ the nth order identity matrix. For a vector $\boldsymbol{x} \in \mathbb{R}^n$, denote by $\|\boldsymbol{x}\|_p$ its l_p-norm with $p \in \mathbb{N} \cup \{\infty\}$. Conventionally, we write $\|\boldsymbol{x}\|$ as the l_2-norm of $\boldsymbol{x}$ without the subscript. In addition, $\mathrm{diag}(\boldsymbol{x}) \in \mathbb{R}^{n\times n}$ denotes the diagonal matrix whose diagonals are $x_1, \ldots, x_n$. For a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $\boldsymbol{A}_{i,\cdot} \in \mathbb{R}^n$ and $\boldsymbol{A}_{\cdot,j} \in \mathbb{R}^m$, respectively represent its ith row and jth column, and we denote $\boldsymbol{A}^\dagger$ as its Moore-Penrose pseudoinverse. Moreover, we denote $\lambda_k(\boldsymbol{A})$ as the kth largest eigen-value of a symmetric matrix $\boldsymbol{A}$, and the smallest and largest eigenvalues are also denoted by $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$, respectively. If $\boldsymbol{A}$ is positive definite, we have $\lambda_{\max}(\boldsymbol{A}) = \|\boldsymbol{A}\|$, the spectral norm of $\boldsymbol{A}$, while $\lambda_{\min}(\boldsymbol{A}) = \|\boldsymbol{A}^{-1}\|^{-1}$. The regular matrix product, Hadamard product, Kronecker product, and Khatri-Rao product (column-wise Kronecker product) between two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are denoted by $\boldsymbol{AB}$, $\boldsymbol{A}*\boldsymbol{B}$, $\boldsymbol{A}\otimes\boldsymbol{B}$, and $\boldsymbol{A}\odot\boldsymbol{B}$, respectively. For convenience, we place the lowest operation priority on Hadamard products among the above products, for

example, $\boldsymbol{AB} * \boldsymbol{C} = (\boldsymbol{AB}) * \boldsymbol{C}$. Suppose $\boldsymbol{A}$ and $\boldsymbol{B}$ are conformable symmetric matrices, we write $\boldsymbol{A} \preceq \boldsymbol{B}$ if $\boldsymbol{B} - \boldsymbol{A}$ is positive semi-definite. Finally, for two positive sequences a_n and b_n, $a_n = O(b_n)$ implies there exists an absolute constant C such that $a_n \leq Cb_n$ for all n, $a_n = \Omega(b_n)$ means $b_n = O(a_n)$, and $a_n = o(b_n)$ stands for $\lim_{n\to\infty} a_n/b_n = 0$. For convenience, we denote $a_n \asymp b_n$ if $a_n = O(b_n)$ and $a_n = \Omega(b_n)$. The terms "module", "cluster", and "community" will be used interchangeably.

2. Network-based Neighborhood Regression

2.1. Genetic Risk and Differential Expressed Scores Modeling

Consider two sources of evidence from statistical tests: the genetic risk (GR) score $\boldsymbol{x} = (x_i)_{i\in[n]} \in \mathbb{R}^n$ and the differentially expressed (DE) score $\boldsymbol{y} = (y_i)_{i\in[n]} \in \mathbb{R}^n$ for n genes. These scores can be derived from previous genetic studies (Fu et al. 2022) and genomic analyses (Gandal et al. 2022). Additionally, a gene co-expression network, revealing bivariate dependencies between gene expression patterns and their corresponding sub-networks (modules) (Liu, Lei, and Roeder 2015), often serves as auxiliary information. Let $\boldsymbol{A} = (A_{i,j})_{i,j\in[n]} \in \{0,1\}^{n\times n}$ denote the adjacency matrix of an undirected and unweighted gene co-expression network. To characterize the directional effect from genetic evidence to genomic evidence with network information, we consider the following network-based neighborhood regression model:

$$y_i = \sum_{j\in N_i} \widetilde{\beta}_{i,j} x_j + \epsilon_i, \text{ for } i \in [n], \tag{1}$$

where $N_i = \{j \in [n] : A_{i,j} = 1\}$ is the neighborhood of gene $i \in [n]$, $\widetilde{\beta}_{i,j}$ is the effect sizes from gene j to gene i for $i, j \in [n]$, and $\boldsymbol{\epsilon} = (\epsilon_i)_{i\in[n]}$ contains independent additive noises. We further denote the coefficient matrix by $\widetilde{\boldsymbol{\beta}} = (\widetilde{\beta}_{i,j})_{i,j\in[n]} \in \mathbb{R}^{n\times n}$.

Model (1) formulates y_i as a linear combination of its neighbors' covariate x_j's, up to additive noise. Specifically, the DE score of gene i is affected by the GR scores of its neighboring j's who is connected to i, for $i, j \in [n]$. Because y_i is supposed to be affected by x_i, we assume $A_{i,i} \equiv 1$ and thus $i \in N_i$. When the network information is not available, this model reduces to the simple linear regression model with $N_i = \{i\}$ and $\widetilde{\beta}_{i,i} = \beta$ for some constant $\beta \in \mathbb{R}$, and $i \in [n]$. For simplicity, model (1) does not include an intercept term as one can always center the data prior to model fitting; see Appendix A for more details.

Compared to high-dimensional regressions, estimating the coefficient matrix $\widetilde{\boldsymbol{\beta}}$ in model (1) is particularly challenging, even with the presence of only a *single* covariate. The fundamental difficulty stems from the model's overparameterization: while the model includes n^2 unknown parameters, there are merely n pairs of GR and DE scores available, creating a significant disparity. This imbalance makes the estimation of $\widetilde{\boldsymbol{\beta}}$ impractical without introducing additional structural constraints. To address this challenge, we turn to an inherent characteristic in network data-community structure. Community structure reflects the tendency of nodes within the same community to exhibit similar linking patterns and usually more intense connections compared to nodes in different communities, particularly in assortative networks. By using community information from the gene co-expression network, we can enhance our predictive modeling and address the over-parameterization issue effectively.

Suppose there are K communities among the genes. We use $\boldsymbol{Z} \in \{0,1\}^{n\times K}$ to denote the community membership such that $Z_{i,k} = 1$ if gene i is in the kth community. Given the motivation above, we impose a block structure of $\widetilde{\boldsymbol{\beta}}$ according to the community structure

$$\widetilde{\boldsymbol{\beta}} = \boldsymbol{Z}\boldsymbol{\beta}\boldsymbol{Z}^\top, \tag{2}$$

for $\boldsymbol{\beta} \in \mathbb{R}^{K\times K}$. The diagonal entries of $\boldsymbol{\beta}$ reflect the within-community causal effects, while the off-diagonal entries represent the between-cluster effect strengths. We remark that different from the conventional stochastic block model for network data, the core coefficient matrix $\boldsymbol{\beta}$ is not necessarily symmetric as the effect β_{k_2,k_1} from the GR scores in the k_1th community to the DE scores in the k_2th community are directional, for $k_1, k_2 \in [K]$. In genetic studies, genes within functional modules or pathways often behave similarly and exhibit coordinated expression and regulation (Langfelder and Horvath 2008). For instance, in an etiologically active community, genes frequently co-express, leading to a strong positive causal effect from GR to DE scores, whereas this relationship may be weaker or even opposite in an etiologically inactive community. Therefore, imposing block structure in the regression coefficient matrix, or module-level genetic studies in general, could increase statistical power and biological insight compared to individual gene analyses, especially for single-cell data analysis with limited sample sizes (Adamson et al. 2016; Jin et al. 2020). With the block structure of $\widetilde{\boldsymbol{\beta}}$ in (2), the neighborhood regression model (1) can be rewritten as

$$\boldsymbol{y} = (\boldsymbol{Z}\boldsymbol{\beta}\boldsymbol{Z}^\top * \boldsymbol{A})\boldsymbol{x} + \boldsymbol{\epsilon}, \tag{3}$$

where $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^\top$ is the noise vector. The proposed network-based neighborhood regression framework is illustrated in Figure 1. Compared to fused lasso in high-dimensional regression (Hesamian, Johannssen, and Chukhrova 2024) and graph-fused lasso (Yu, Yang, and Zhang 2025), model (3) directly accommodates neighborhood effects and community structure by sparsity and community-wise constancy, which reduces model complexity and enables computational feasibility. Relaxation to soft sparsity and constancy would be interesting and requires further exploration.

In what follows, we assume the community membership $\boldsymbol{Z}$ is known. If it is unknown, it can be exactly recovered with high probability, which is called strong consistency (Zhao, Levina, and Zhu 2012), or recovered up to a vanishing fraction with high probability, which is called weak consistency (Zhao, Levina, and Zhu 2012), from the network data under relatively mild conditions, provided the averaged degree of the nodes diverges as n goes to infinity. We will inspect the issue of membership misspecification in Theorem 7.

2.2. Community-Wise Least Square Estimation

To estimate the core community-level coefficient $\boldsymbol{\beta}$ in (2), we consider the following least square objective as in conventional linear regression framework

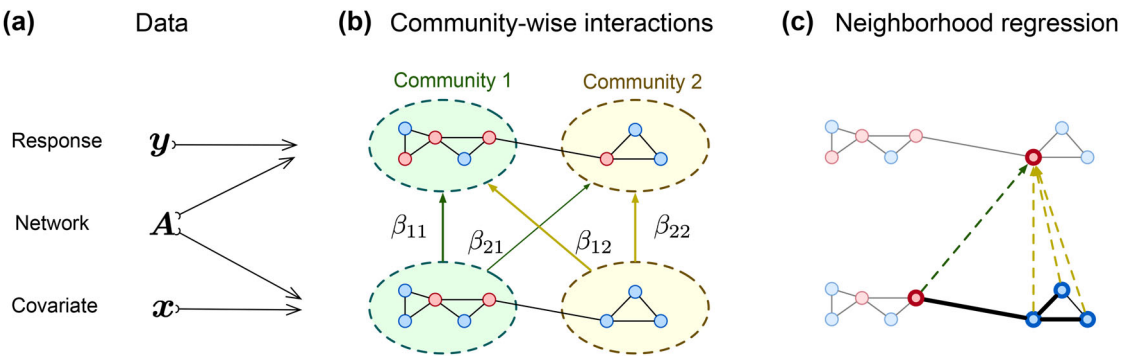

Figure 1. Network-based neighborhood regression. (a) The covariate $\boldsymbol{x}$ and response $\boldsymbol{y}$ are observed for each node in a network $\boldsymbol{A}$. (b) The community-wise interactions are modeled by the coefficient matrix $\boldsymbol{\beta}$. (c) The conditional mean of a particular response is the average of the covariates in its neighborhood weighted by the community-wise interaction strengths.

$$\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{2n}\|\boldsymbol{y} - (\boldsymbol{Z}\boldsymbol{\beta}\boldsymbol{Z}^\top * \boldsymbol{A})\boldsymbol{x}\|_2^2. \tag{4}$$

At first glance, the objective function (4) looks a bit counterintuitive since the parameters to be estimated $\boldsymbol{\beta}$ and the predictor vector $\boldsymbol{x}$ respectively take the places of the design matrix and regression coefficients in the classical multiple linear regression setup. Fortunately, simple algebra yields that the objective function (4) can be decomposed into the weighted average of K nonoverlapping objective functions in that

$$\mathcal{L}(\boldsymbol{\beta}) = \sum_{k=1}^{K} \frac{n_k}{n}\mathcal{L}_k(\boldsymbol{\beta}_{k,\cdot}), \text{ with}$$
$$\mathcal{L}_k(\boldsymbol{\beta}_{k,\cdot}) = \frac{1}{2n_k}\|\boldsymbol{y} * \boldsymbol{Z}_{\cdot,k} - \boldsymbol{M}_k\boldsymbol{\beta}_{k,\cdot}\|_2^2,$$

where $n_k = \sum_{i=1}^{n} Z_{i,k}$ is the size of the kth community and the design matrix for the linear regression objective with transformed response $\boldsymbol{y} * \boldsymbol{Z}_{\cdot,k}$ is

$$\boldsymbol{M}_k = \operatorname{diag}(\boldsymbol{Z}_{\cdot,k})\left(\boldsymbol{1}_n\boldsymbol{x}^\top * \boldsymbol{A}\right)\boldsymbol{Z}. \tag{5}$$

We say the objective functions $\mathcal{L}_k$'s are nonoverlapping because $\mathcal{L}_k$ is solely a function of the kth row of the core coefficient matrix $\boldsymbol{\beta}$ but not any other entries. Therefore, minimizing $\mathcal{L}(\boldsymbol{\beta})$ is equivalent to minimizing each $\mathcal{L}_k(\boldsymbol{\beta}_{k,\cdot})$ individually, for $k \in [K]$. Intuitively, only those y_i's with i inside the kth community contain useful information in estimating $\boldsymbol{\beta}_{k,\cdot}$ while every x_j can contribute to y_i. This is why, in $\mathcal{L}_k(\cdot)$, we can zero out the responses not in the kth community by $\boldsymbol{y} * \boldsymbol{Z}_{\cdot,k}$ and the corresponding rows in the design matrix $\boldsymbol{M}_k$ by a factor $\operatorname{diag}(\boldsymbol{Z}_{\cdot,k})$, while the information of every x_j is encoded in the jth column of the factor $\boldsymbol{1}_n\boldsymbol{x}^\top * \boldsymbol{A}$ in $\boldsymbol{M}_k$. Moreover, $(\boldsymbol{1}_n\boldsymbol{x}^\top * \boldsymbol{A})\boldsymbol{Z}$ aggregates the samples' neighborhood effects according to their community memberships. One can essentially eliminate $n - n_k$ zero entries of $\boldsymbol{y} * \boldsymbol{Z}_{\cdot,k} - \boldsymbol{M}_k\boldsymbol{\beta}_{k,\cdot}$ to save memory and speed up the computation, but we stick with this expression to avoid introducing excessive notations. Before ending this paragraph, we remark that the community-wise least square estimation can be readily extended to multiple regression settings through tensor decomposition, as demonstrated in Appendix B.

2.3. Fixed-Design Analysis

From the discussion above, it is sufficient to minimize $\mathcal{L}_k(\boldsymbol{\beta}_{k,\cdot})$ for each community individually. We, therefore, coin the proposed optimization problem as community-wise least square optimization. The corresponding solution, *Community-wise Least Square Estimator* (CLSE), is defined as

$$\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_{1,\cdot}, \ldots, \widehat{\boldsymbol{\beta}}_{K,\cdot})^\top \text{ with } \widehat{\boldsymbol{\beta}}_{k,\cdot} = \operatorname*{argmin}_{\boldsymbol{\beta}_{k,\cdot}\in\mathbb{R}^K} \mathcal{L}_k(\boldsymbol{\beta}_{k,\cdot}). \tag{6}$$

In the fixed-design scenario, both the covariate $\boldsymbol{x}$ and the network $\boldsymbol{A}$, and hence $\boldsymbol{M}_k$, are treated as deterministic while only the responses are random, we can recover various desired properties as in the classical multiple linear regression setups. For instance, the objective function $\mathcal{L}_k$ is convex with respect to $\boldsymbol{\beta}_{k,\cdot} \in \mathbb{R}^K$ and is strongly convex with parameter $\lambda_{\min}(\boldsymbol{M}_k^\top\boldsymbol{M}_k)$ if $\boldsymbol{M}_k$ has full column rank. Moreover, a closed-form expression for the CLSE estimator (6) can be derived as in the following proposition.

Proposition 1 (Stationary point). The stationary point of $\partial\mathcal{L}_k(\boldsymbol{\beta}_{k,\cdot})/\partial\boldsymbol{\beta}_{k,\cdot}$ is the solution to the following system of normal equations:

$$\boldsymbol{M}_k^\top\boldsymbol{M}_k\boldsymbol{\beta}_{k,\cdot} = \boldsymbol{M}_k^\top\boldsymbol{y}, \tag{7}$$

for any $k \in [K]$. Particularly, when $\boldsymbol{M}_k$ has rank K, it follows that $\widehat{\boldsymbol{\beta}}_{k,\cdot} = (\boldsymbol{M}_k^\top\boldsymbol{M}_k)^{-1}\boldsymbol{M}_k^\top\boldsymbol{y}$.

Proposition 1 resembles the solution to the least square estimator in multiple linear regression. When $\boldsymbol{M}_k$ is singular, there are infinitely many solutions to (7). In this case, one may use the minimum l_2-norm solution $\widehat{\boldsymbol{\beta}}_{k,\cdot} = (\boldsymbol{M}_k^\top\boldsymbol{M}_k)^\dagger\boldsymbol{M}_k^\top\boldsymbol{y}$ (Hastie et al. 2022; Patil, Du, and Tibshirani 2024). Comparing to imposing general low-rank, sparse, and/or fusion structure on the regression coefficient matrix as in Zhang, Sun, and Li (2020, 2023), where flexibility for data-adaptive learning increases while careful optimization algorithm design is required, the proposed block structure on the regression coefficient matrix directly enables an explicit solution for the CLSE.

Denote the Hessian matrix of $\mathcal{L}_k(\cdot)$ by $\boldsymbol{H}_k = \boldsymbol{M}_k^\top\boldsymbol{M}_k$ and the underlying true parameter by $\boldsymbol{\beta}_{k,0} = \{\mathbb{E}(\boldsymbol{H}_k \mid \boldsymbol{M}_k)\}^{-1}\mathbb{E}(\boldsymbol{M}_k^\top\boldsymbol{y} \mid \boldsymbol{M}_k) = \boldsymbol{H}_k^{-1}\boldsymbol{M}_k^\top\mathbb{E}(\boldsymbol{y} \mid \boldsymbol{M}_k)$. A consequence of Proposition 1 is the asymptotic normality of the estimator, as stated in the following corollary.

Corollary 2 (Asymptotic normality). Conditioned on $\boldsymbol{A}$ and $\boldsymbol{x}$, for any $k \in [K]$, assume that $\text{rank}(\boldsymbol{M}_k) = K$ and ϵ_i's are independent and identically distributed with $\mathbb{E}(\epsilon_i) = 0$ and $\text{var}(\epsilon_i) = \sigma_k^2$, for those i's in the kth community. It then follows that

$$\widehat{\boldsymbol{\beta}}_{k,\cdot} \to \mathcal{N}_K(\boldsymbol{\beta}_{k,0}, \sigma_k^2 \boldsymbol{H}_k^{-1}),$$

in distribution, where $\mathcal{N}_K$ stands for K-dimensional multivariate normal distribution.

Corollary 2 can be readily obtained from classical results for linear regression estimator (Eicker 1963; Fahrmeir and Kaufmann 1985). Under the full column rank assumption on $\boldsymbol{M}_k$, both Proposition 1 and Corollary 2 suggest that the CLSE $\widehat{\boldsymbol{\beta}}_{k,\cdot}$ shares the same form and properties as the least squares estimator for classical multiple linear regression.

The fixed-design analysis mandates the design matrix $\boldsymbol{M}_k$ to be nondegenerate. In our formulation, the rank of the Hessian is influenced not only by the covariates $\boldsymbol{x}$ but also by the network $\boldsymbol{A}$, which is typically a single, noisy sample. For instance, gene co-expression networks are derived by binarizing correlation matrices with measurement errors (Liu, Lei, and Roeder 2015). When $\boldsymbol{A}$ is considered as a random network, its symmetry causes the rows of $\boldsymbol{M}_k$ to be neither independent nor identically distributed. This inherent sampling randomness and unique structure of $\boldsymbol{A}$ complicate the rank conditions, consistency, and optimality analysis, setting our neighborhood regression framework apart from conventional random-design linear regression. Therefore, an in-depth investigation of random-design analysis for our proposed framework is essential, as elaborated in the next section.

3. Random-Design Analysis

3.1. Assumptions

We begin by introducing several technical assumptions. First, similar to the majority of literature for network data analysis (Abbe 2018; Lee and Wilkinson 2019), we assume the network data $\boldsymbol{A}$ follows the stochastic block model up to deterministic self-loops.

Assumption 1 (Stochastic block model). Assume that $A_{i,j} = A_{j,i}$'s are independent Bernoulli random variables with success probability $P_{i,j}$'s, for $i < j$, where $\boldsymbol{P} = \boldsymbol{Z}\boldsymbol{B}\boldsymbol{Z}^\top$ and $\boldsymbol{B} \in \mathbb{R}^{K\times K}$ is a symmetric community level probability matrix.

In Assumption 1, the probability matrix $\boldsymbol{B}$ determines the connectivity strengths between and within communities, and $P_{i,j}$ only depends on the community memberships of vertexes i and j, for $i \neq j$. Independent connectivities may not faithfully capture the actual relationships among genes, and Bahadur representation has been proposed by Yuan and Qu (2021) to model the dependent connectivities for network data under the stochastic block model. In addition, The network sparsity can be characterized by $s_n = \max_{k_1,k_2\in[K]} B_{k_1,k_2}$. As real-life networks are usually sparse, community detection and other tasks in network data can be feasible only when the average degree of the vertexes diverges. We, therefore, require the following assumption on the network sparsity.

Assumption 2 (Network sparsity). Assume that the network sparsity satisfy $s_n \geq \frac{\log n}{n}$.

Similar definitions of network sparsity and assumption have been popularly employed in network data analysis, including hypergraph networks (Zhen and Wang 2023) and multi-layer networks (Lei, Chen, and Lynch 2020; Xu, Zhen, and Wang 2023).

Unlike most literature on network data analysis that requires the community sizes to be balanced or asymptotically balanced (i.e., n_k's are asymptotically of the same order), we only require the following much weaker assumption that tailors the proposed network-based neighborhood regression framework.

Assumption 3 (Community sizes). There exists some positive constant δ such that

$$\|\boldsymbol{x}\|_\infty \max_{k\in[K]} \left(\frac{\log n}{n_k s_n (\|\boldsymbol{x}\|^2 + \|\boldsymbol{x}^{(k)}\|^2)} \right)^{1/2} \leq \delta, \tag{8}$$

where $\boldsymbol{x}^{(k)} = \boldsymbol{Z}_{\cdot,k} * \boldsymbol{x}$ for $k \in [K]$.

Assumption 3 is relatively mild, and we can understand it through the following examples. As the first example, if $|x_i|$'s are upper bounded and lower bounded away from 0, then the quantity on the left hand side of (8) is essentially of the order $\{\log(n)/(n_k n s_n)\}^{1/2} = \mathcal{O}(n_k^{-1/2})$ according to Assumption 2. As another example, if $x_1, \ldots, x_n$ are independent standard normal random variables, and thus $\|\boldsymbol{x}\|^2$ is a Chi-square random variable with degree n, then it can be shown that $\|\boldsymbol{x}\|_\infty \leq c \log n$ and $|\|\boldsymbol{x}\|^2 - n| \leq c(n \log n)^{1/2}$ with high probability for some constant c. Therefore, the left hand side of (8) is $\mathcal{O}_p(\log n \{\log(n)/(n_k n s_n)\}^{1/2}) = \mathcal{O}_p(n_k^{-1/2} \log n)$. In both examples, (8) holds as long as the community size n_k diverges faster than $(\log n)^2$. Moreover, both examples allow the upper bound δ to vanish. We also remark that including the $\|\boldsymbol{x}^{(k)}\|^2$ in the denominator can better accommodate imbalanced communities because the optimal k in the left-hand side of (8) does not necessarily lead to $n_{\min}$ in the denominator, the smallest community size.

The next assumption concerns the tail behavior of the additive noise $\boldsymbol{\epsilon}$.

Assumption 4 (Additive noise). Conditional on $\boldsymbol{x}$ and $\boldsymbol{A}$, we assume $\epsilon_1, \ldots, \epsilon_n$ are independent centered sub-exponential random variables with uniform parameters $(\sigma_\epsilon^2, b_\epsilon)$. Precisely,

$$\mathbb{E}(e^{t\epsilon_i}) \leq e^{t^2\sigma_\epsilon^2/2}, \text{ for any } |t| < b_\epsilon^{-1} \text{ and } i \in [n].$$

The conditional independent assumption on ϵ_i's still allows for the dependence between ϵ_i and ϵ_j, for any $i \neq j$. This makes the assumption more realistic to genomic data, where the genes are adjacency with each other in a chromatin that has a 3-dimensional spatial structure. Alternatively, we may directly model the dependence of $\epsilon_1, \ldots, \epsilon_n$ as a weakly dependence sequence satisfying τ-mixing property (Merlevède, Peligrad, and Rio 2011). The sub-exponential assumption on ϵ_i's is mild and includes the classes of bounded random variables and sub-Gaussian random variables (Wainwright 2019, Definition 2.7). Also, Assumption 4 does not require ϵ_i's to be identically distributed, while most classic regression setups do. Finally, to

better illustrate the rate of convergence, sometimes it will be convenient to assume the following tali bound for the maximum deviation of $\boldsymbol{x}$'s entries.

Assumption 5. Conditional on $\boldsymbol{A}$, x_i's are independent zero-mean random variables, and there exists a constant γ and a quantity κ_n vanishing with n such that with probability at least $1-\kappa_n$, it holds that $\|\boldsymbol{x}\|_\infty \le \gamma(\log n)^{1/2}$.

Note that the x_i's are not necessarily independent under Assumption 5, and a wide range of classes of distributions, such as sub-exponential random variables, shall satisfy such exponentially decaying probabilistic tail bound.

3.2. Non-Asymptotic Analysis of the Hessian

In this section, we study the spectral property of the Hessian $\boldsymbol{H}_k$ by investigating its mean $\mathbb{E}(\boldsymbol{H}_k)$ and its concentration behavior of $\boldsymbol{H}_k$ to $\mathbb{E}(\boldsymbol{H}_k)$. We first decompose $\boldsymbol{H}_k$ as

$$\boldsymbol{H}_k = \boldsymbol{I}_1 + \boldsymbol{I}_2 + \boldsymbol{I}_3 + \boldsymbol{I}_4, \tag{9}$$

where $\boldsymbol{I}_1 = \boldsymbol{Z}^\top \left[\boldsymbol{x}\boldsymbol{x}^\top * (\boldsymbol{A}-\mathbb{E}(\boldsymbol{A}))\mathrm{diag}(\boldsymbol{Z}_{\cdot,k})(\boldsymbol{A}-\mathbb{E}(\boldsymbol{A}))\right]\boldsymbol{Z}$ is a matrix quadratic form of $\boldsymbol{A}-\mathbb{E}(\boldsymbol{A})$, $\boldsymbol{I}_2 = \boldsymbol{Z}^\top \left[\boldsymbol{x}\boldsymbol{x}^\top * (\boldsymbol{A}-\mathbb{E}(\boldsymbol{A}))\mathrm{diag}(\boldsymbol{Z}_{\cdot,k})\mathbb{E}(\boldsymbol{A})\right]\boldsymbol{Z}$ has zero-mean, $\boldsymbol{I}_3 = \boldsymbol{I}_2^\top$, and $\boldsymbol{I}_4 = \boldsymbol{Z}^\top \left[\boldsymbol{x}\boldsymbol{x}^\top * \mathbb{E}(\boldsymbol{A})\mathrm{diag}(\boldsymbol{Z}_{\cdot,k})\mathbb{E}(\boldsymbol{A})\right]\boldsymbol{Z}$ is deterministic and positive semidefinite when $\boldsymbol{x}$ is given. Moreover, we can further decompose $\boldsymbol{I}_1 = \boldsymbol{S}_1 + \boldsymbol{S}_2$, where

$$\begin{aligned}\boldsymbol{S}_1 = &\sum_{k_1=1}^{K}\sum_{k_2=1}^{K}\sum_{\psi_i=k_1}\sum_{\psi_{i'}=k_2}\sum_{\psi_j=k} x_i x_{i'} \\ &\mathbb{1}\left(i\neq i'\right)(\boldsymbol{A}-\mathbb{E}(\boldsymbol{A}))_{i,j}(\boldsymbol{A}-\mathbb{E}(\boldsymbol{A}))_{i',j}\boldsymbol{e}_{k_1}\boldsymbol{e}_{k_2}^\top, \text{and}\end{aligned}$$

$$\boldsymbol{S}_2 = \sum_{k_1=1}^{K}\sum_{\psi_i=k_1}\sum_{\psi_j=k} x_i^2(A_{i,j}-\mathbb{E}(A_{i,j}))^2\boldsymbol{e}_{k_1}\boldsymbol{e}_{k_1}^\top. \tag{10}$$

Herein, $\boldsymbol{e}_j \in \mathbb{R}^n$ is a unit vector whose jth entry being one, $\psi_i = \mathrm{argmax}_{k\in[K]} Z_{i,k}$ is the community assignment of the ith sample. Note that the notation $\sum_{\psi_i=k_1}$ is a shorthand notation for $\sum_{i:,\psi_i=k_1}$, which sums terms across all the indexes i in the k_1th community, similarly for $\sum_{\psi_{i'}=k_2}$ and $\sum_{\psi_j=k}$. Clearly, $\boldsymbol{S}_1$ is a zero-mean symmetric matrix, and $\boldsymbol{S}_2$ is a diagonal matrix that has nonnegative diagonals. Intuitively, the spectral information of $\boldsymbol{H}_k$ is mainly encoded in $\boldsymbol{S}_2$ and $\boldsymbol{I}_4$, while the zero-mean terms shall have small spectral norms. In fact, it follows from the above decomposition that $\mathbb{E}(\boldsymbol{H}_k) = \mathbb{E}(\boldsymbol{S}_2) + \boldsymbol{I}_4$, and

$$\boldsymbol{H}_k - \mathbb{E}(\boldsymbol{H}_k) = \boldsymbol{S}_1 + \{\boldsymbol{S}_2 - \mathbb{E}(\boldsymbol{S}_2)\} + (\boldsymbol{I}_2 + \boldsymbol{I}_2^\top) \tag{11}$$

Since $\boldsymbol{I}_4$ is positive semidefinite, we can verify straightforwardly that

$$\begin{aligned}\lambda_{\min}\left(\mathbb{E}(\boldsymbol{H}_k)\right) &\ge \lambda_{\min}\left(\mathbb{E}(\boldsymbol{S}_2)\right) \\ &\ge \min_{k'\in[K]} B_{k',k}(1-B_{k',k})(n_k-1)\|\boldsymbol{x}^{(k')}\|^2.\end{aligned} \tag{12}$$

From (12), the lower bound of $\lambda_{\min}(\mathbb{E}(\boldsymbol{H}_k))$ shall be of order $\Omega(n_k n_{\min} s_n)$ with $n_{\min} = \min_{k'\in[K]} n_{k'}$ if $B_{k',k}$ is of order $\Omega(s_n)$. This indicates that $\mathbb{E}(\boldsymbol{H}_k)$ is guaranteed to have full rank. If the perturbation of $\boldsymbol{H}_k - \mathbb{E}(\boldsymbol{H}_k)$ can be further controlled, we are able to infer that $\boldsymbol{H}_k$ is nonsingular with high probability. The following theorem provides a careful perturbation analysis of $\boldsymbol{H}_k - \mathbb{E}(\boldsymbol{H}_k)$.

Theorem 3 (Fisher information concentration). Under Assumption 1–3, for the Hessian $\boldsymbol{H}_k = \boldsymbol{M}_k^\top \boldsymbol{M}_k$ with $\boldsymbol{M}_k$ defined in (5), it holds that $\|\boldsymbol{H}_k - \mathbb{E}(\boldsymbol{H}_k)\| \le r_k$ with probability at least $1 - 2K(C_1 n_k + 2C_1 + 2)/n^2$, for some universal constant C_1, where

$$\begin{aligned}r_k = &\alpha(\delta)(C_1 + {s_n}^{1/2})(\|\boldsymbol{x}\|^2 + \|\boldsymbol{x}^{(k)}\|^2)^{1/2} \\ &\times \left\{\left(\sum_{k'=1}^{K}\|\boldsymbol{x}^{(k')}\|_\infty^2\right)^{1/2} + \frac{\|\boldsymbol{x}\|}{n^{1/2}}\right\} s_n(n_k n)^{1/2}\log n \\ &+ 2^{3/2}\left(\frac{2}{3^{1/2}}C_1 {s_n}^{1/2} + 1\right)\max_{k'\in[K]}\|\boldsymbol{x}^{(k')}\|_4^2 (n_k s_n \log n)^{1/2} \\ &+ 4(C_1+1)\|\boldsymbol{x}\|_\infty^2 \log n,\end{aligned}$$

and $\alpha(\delta) = \{(8\delta + 4(4\delta^2+9)^{1/2})\}/3$.

The universal constant C_1 comes from the decoupling constant in de la Pena and Montgomery-Smith (1995), and our result and technical proof do not induce any other unclear constant. In addition, δ comes from Assumption 3 and $\alpha(\delta)$ will decrease to 4 if δ vanishes. As non-asymptotic analysis is conducted, the probabilistic upper bound looks a bit complicated. The next corollary details the asymptotic order of $\|\boldsymbol{H}_k - \mathbb{E}(\boldsymbol{H}_k)\|$.

Corollary 4. Under Assumptions 1–3 and 5, there exists a universal constant C_2, such that

$$\|\boldsymbol{H}_k - \mathbb{E}(\boldsymbol{H}_k)\| \le C_2\gamma^2 K^{1/2} s_n n n_k^{1/2}(\log n)^2,$$

with probability at least $1 - 2K(C_1 n_k + 2C_1 + 2)/n^2 - \kappa_n$.

Since $\lambda_{\min}(\boldsymbol{H}_k) = \Omega(s_n n_k n_{\min})$, Corollary 4 allows us to conclude concentration if $n(\log n)^2 = O(n_k^{1/2} n_{\min})$ and $K = O(1)$. It then follows from Weyl's inequality that $\lambda_{\min}(\boldsymbol{H}_k)$ is asymptotically of the same order as $\lambda_{\min}(\mathbb{E}(\boldsymbol{H}_k))$.

Based on the decomposition (11), the proof for Theorem 3 relies on a decoupling approach (de la Pena and Montgomery-Smith 1995) to bound the matrix quadratic form $\boldsymbol{S}_1$, the usual Bernstein's inequality together with the union bound to bound the diagonal matrix $\boldsymbol{S}_2 - \mathbb{E}(\boldsymbol{S}_2)$ and matrix Bernsten's inequality (Tropp 2012) to bound $\boldsymbol{I}_2$, which are done in Lemmas 11-12, Lemma 13, and Lemma 14 in the supplementary materials, respectively. Our proof technique is related to, but substantially different from, the technique in Lei and Lin (2023) or Hanson-Wright type inequality (Rudelson and Vershynin 2013) for matrix quadratic forms. This is because every summand in the decomposition (9) contains a left factor $\boldsymbol{Z}^\top$ and a right factor $\boldsymbol{Z}$ that aggregates the random variables according to their community memberships, while the Theorems in Lei and Lin (2023) work for the matrix quadratic form $\boldsymbol{F}\boldsymbol{G}\boldsymbol{F}^\top$ for a deterministic matrix $\boldsymbol{G}$ and a random matrix $\boldsymbol{F}$ that has independent entries or is symmetric with independent upper triangle entries. Apparently, the appearance of $\boldsymbol{Z}$ makes $\boldsymbol{Z}^\top[\boldsymbol{x}\boldsymbol{x}^\top * (\boldsymbol{A}-\mathbb{E}(\boldsymbol{A}))]\mathrm{diag}(\boldsymbol{Z}_{\cdot,k})$, the random matrix in $\boldsymbol{I}_1$ for example, neither symmetric nor have independent entries. Additionally, the Hadamard factor $\boldsymbol{x}\boldsymbol{x}^\top$ also adds an extra layer of difficulty to derive the probabilistic concentration bound. All of these require a subtle and careful analysis.

Remark 1 (Matrix quadratic form). In the simple scenario that $\boldsymbol{x} = \mathbf{1}_n$, if we upper bound $\|\boldsymbol{I}_1\|$ by $n_{\max}\|(\boldsymbol{A} - \mathbb{E}(\boldsymbol{A}))\mathrm{diag}(\boldsymbol{Z}_{\cdot,k})(\boldsymbol{A} - \mathbb{E}(\boldsymbol{A}))\|$ and employ Theorem 5 of Lei and Lin (2023) to upper bound $\|(\boldsymbol{A}-\mathbb{E}(\boldsymbol{A}))\mathrm{diag}(\boldsymbol{Z}_{\cdot,k})(\boldsymbol{A}-\mathbb{E}(\boldsymbol{A}))\| = \mathcal{O}_{\mathbb{P}}(ns_n \log n)$, it leads to $\|\boldsymbol{I}_1\| = \mathcal{O}_{\mathbb{P}}(n_{\max} n s_n \log n)$ with $n_{\max} = \max_{k' \in [K]} n_{k'}$, which fails to conclude concentration since $\lambda_{\min}(\mathbb{E}(\boldsymbol{H}_k)) = \Omega(n n_{\min} s_n)$. Therefore, it is vital to make full use of the aggregation structure in $\boldsymbol{I}_1$ while studying its concentration behavior.

3.3. Consistency

Under the random-design setting, the oracle coefficient $\boldsymbol{\beta}^*_{k,\cdot}$ is defined as the solution to the population-level normal equation. More specifically, by taking expectation on both sides of (7), we obtain

$$\boldsymbol{\beta}^*_{k,\cdot} = \mathbb{E}(\boldsymbol{H}_k)^{-1}\mathbb{E}(\boldsymbol{M}_k^\top \boldsymbol{y}),$$

where the expectation is taken with respect to both $\boldsymbol{A}$ and $\boldsymbol{\epsilon}$. Conditional on $\boldsymbol{x}$, $\boldsymbol{A}$ and the event that $\lambda_{\min}(\boldsymbol{H}_k)$ diverges, when $\mathrm{var}(\epsilon_i) = \sigma^2$, for all i in the kth community, Corollary 2 provides asymptotic normality and thus concludes the asymptotic consistency of $\widehat{\boldsymbol{\beta}}_{\cdot,k}$ without concerning the rate of convergence. For general random-design analysis, the next theorem shows that the estimator $\widehat{\boldsymbol{\beta}}_{k,\cdot}$ is an unbiased of $\boldsymbol{\beta}^*_{k,\cdot}$, and provides the rate of convergence.

Theorem 5 (Unbiasedness and Consistency). Under Assumption 4, it holds that $\mathbb{E}(\widehat{\boldsymbol{\beta}}_{k,\cdot}) = \boldsymbol{\beta}^*_{k,\cdot}$. Moreover, with probability at least $1 - (2K)/n^2$, we have

$$\|\widehat{\boldsymbol{\beta}}_{k,\cdot} - \boldsymbol{\beta}^*_{k,\cdot}\| \le 2\,(2K/\lambda_{\min}(\boldsymbol{H}_k))^{1/2}\,(\sigma_\epsilon + b_\epsilon)\log n.$$

Additionally under the assumptions in Theorem 3, we have

$$\|\widehat{\boldsymbol{\beta}}_{k,\cdot} - \boldsymbol{\beta}^*_{k,\cdot}\| \le 2\,(2K)^{1/2}\,(\sigma_\epsilon + b_\epsilon) \times \left\{\min_{k' \in [K]} B_{k',k}(1 - B_{k',k})(n_k - 1)\|\boldsymbol{x}^{(k')}\|^2 - r_k\right\}^{-1/2} \log n.$$

with probability at least $1 - 2K(C_1 n_k + 2C_1 + 3)/n^2$.

Clearly, the upper bound of $\|\widehat{\boldsymbol{\beta}}_{k,\cdot} - \boldsymbol{\beta}^*_{k,\cdot}\|$ comes from the lower bound of $\lambda_{\min}(\boldsymbol{H}_k)$ and the sub-exponential concentration behaviors on the ϵ_i's. The following corollary elaborates on this non-asymptotic upper bound in terms of asymptotic order.

Corollary 6. Under Assumptions 1–5, if additionally $B_{k',k} = \Omega(s_n)$, $\min_{k' \in [K]} \|\boldsymbol{x}^{(k')}\| = \Omega(n_{\min})$, and $n(\log n)^2 = O(n_k^{1/2} n_{\min})$, then there exists a constant C_3, such that

$$\|\widehat{\boldsymbol{\beta}}_{k,\cdot} - \boldsymbol{\beta}^*_{k,\cdot}\| \le C_3 \frac{K^{1/2}(\sigma_\epsilon + b_\epsilon)\log n}{(s_n n_k n_{\min})^{1/2}},$$

with probability at least $1 - 2K(C_1 n_k + 2C_1 + 3)/n^2 - \kappa_n$.

If $s_n \asymp 1$ and the community sizes are balanced such that $n = O(n_{\min})$, surprisingly, Corollary 6 suggests linear consistency, $\|\widehat{\boldsymbol{\beta}}_{k,\cdot} - \boldsymbol{\beta}^*_{k,\cdot}\| = \mathcal{O}(n^{-1})$, instead of the canonical root-n consistency for regression. This shows the blessing of incorporating network neighborhood information. Intuitively, after incorporating the network data, the effective sample size increases from n to $n + s_n n(n-1)/2$, which is of the order $s_n n^2$, while the number of parameters to be estimated is of constant order. This makes linear consistency possible as $(s_n n^2)^{1/2} = s_n^{1/2} n$ is linear in n if s_n is of constant order. This also suggests that the network sparsity s_n plays an important role as the variation of s_n smoothly transforms $\widehat{\boldsymbol{\beta}}_{k,\cdot}$ from canonical regime to blessing of neighborhood information regime.

We next investigate the influence of community detection error on the convergence rate of $\widehat{\boldsymbol{\beta}}_{k,\cdot}$ when the community membership is unknown. Let α_n be the number of mis-clustering nodes from community detection, which is potentially diverging. In most consistent community detection literature, $\alpha_n = \mathcal{O}_{\mathbb{P}}\left(\log n/s_n\right)$; see, for example, Zhen and Wang (2023). We have the following result.

Theorem 7 (Community membership misspecification). Denote $\widehat{\boldsymbol{\beta}}^{\mathrm{mis}}_{k,\cdot}$ be the community-wise least square estimator with α_n nodes misspecified. Under the condition of Corollary 6, there exists some positive constant C_4 and C_5, such that with probability at least $1 - C_4 K/n - \kappa_n$, we have

$$\|\widehat{\boldsymbol{\beta}}^{\mathrm{mis}}_{k,\cdot} - \boldsymbol{\beta}^*_{k,\cdot}\| \le C_5 \frac{K^{1/2}(\sigma_\epsilon + b_\epsilon + \gamma s_n\sqrt{n\alpha_n}\log n)\log n}{(s_n n_k n_{\min})^{1/2}}.$$

Clearly, misspecification of the community membership leads to an extra term $\gamma s_n\sqrt{n\alpha_n}\log n$ in the numerator, depending on the network sparsity. When $\alpha_n = \mathcal{O}(1)$, if the network is extremely sparse with $s_n = \mathcal{O}(\frac{\log n}{n})$ or relatively dense with $s_n = \mathcal{O}(1)$, Theorem 7 maintains the canonical $\sqrt{n}$-consistency, up to logarithm factor. For moderate sparse network, for example $s_n = \mathcal{O}(\frac{\log n}{\sqrt{n}})$, Theorem 7 improves the $\sqrt{n}$-consistency to $n^{3/4}$-consistency, up to logarithm factor. The intuition behind this is as follows. When the community membership of 1 node is misspecified, it leads to potential misspecification of n regression coefficients in $\boldsymbol{Z}^\top \boldsymbol{\beta} \boldsymbol{Z} * \boldsymbol{A}$. However, a sparser network helps to reduce the actual number of misspecified regression coefficients, though it also carries weaker signals. Before ending this subsection, in analogy to the Gauss-Markov Theorem for linear regression, we have the following theorem.

Theorem 8 (Community-wise best linear unbiased estimator). Under Assumptions 1–4, further assume that the variance of ϵ_i's are homogeneous within the kth community, then $\widehat{\boldsymbol{\beta}}_{k,\cdot}$ is the best linear unbiased estimator of $\boldsymbol{\beta}^*_{k,\cdot}$, that is, for any linear unbiased estimator $\bar{\boldsymbol{\beta}}_k$, we have $\mathrm{var}(\widehat{\boldsymbol{\beta}}_{k,\cdot}) \preceq \mathrm{var}(\bar{\boldsymbol{\beta}}_k)$.

Herein, a linear estimator means the estimator is linear in the response $\boldsymbol{y}$.

3.4. Minimax Optimality

To investigate the minimax optimality of the community-wise least square estimator $\widehat{\boldsymbol{\beta}}_{k,\cdot}$, we first introduce a class of data distributions:

$$\mathcal{P}_k(P_X, P_{\boldsymbol{A}}, \sigma) = \big\{P_{X,\boldsymbol{A},Y} : \boldsymbol{x} \sim P_X, \boldsymbol{A} \sim P_{\boldsymbol{A}}, \boldsymbol{\beta}^*_{k,\cdot} \in \mathbb{R}^K, \mathbb{E}(\boldsymbol{\epsilon}) = \boldsymbol{0}_n, \mathbb{E}(\epsilon_i^2) \le \sigma^2_{\max} \text{ for } i \in [n]\big\},$$

for $k \in [K]$, where $\sigma = (\sigma_1, \ldots, \sigma_n)^\top$ and $\sigma_{\max} = \max_{i\in[n]} \sigma_i$, for $\sigma_i = \mathbb{E}(\epsilon_i^2)$, $i \in [n]$. Define the following discrepancy between any estimator $\bar{\boldsymbol{\beta}}_{k,\cdot}$ and $\boldsymbol{\beta}^*_{k,\cdot}$

$$\mathcal{E}(\bar{\boldsymbol{\beta}}_{k,\cdot}) = \|\bar{\boldsymbol{\beta}}_{k,\cdot} - \boldsymbol{\beta}^*_{k,\cdot}\|^2_{\boldsymbol{\Sigma}},$$

for any positive definite matrix $\boldsymbol{\Sigma}$, where $\|\cdot\|_{\boldsymbol{\Sigma}}$ denotes the Mahalanobis distance of a vector to the origin. When $\boldsymbol{\Sigma} = \boldsymbol{I}_n$, $\mathcal{E}(\bar{\boldsymbol{\beta}}_{k,\cdot})$ characterizes the estimation error of $\bar{\boldsymbol{\beta}}_{k,\cdot}$. When $\boldsymbol{\Sigma} = \mathbb{E}(\boldsymbol{H}_k)$, $\mathcal{E}(\bar{\boldsymbol{\beta}}_{k,\cdot})$ captures the generalization performance of $\bar{\boldsymbol{\beta}}_{k,\cdot}$ in that $\mathcal{E}(\bar{\boldsymbol{\beta}}_{k,\cdot}) = \mathcal{R}_k(\bar{\boldsymbol{\beta}}_{k,\cdot}) - \mathcal{R}_k(\boldsymbol{\beta}^*_{k,\cdot})$ represents the excess risk, for the risk function $\mathcal{R}_k(\boldsymbol{\beta}_{k,\cdot}) = \mathbb{E}(\|\boldsymbol{Y} * \boldsymbol{Z}_{\cdot,k} - \boldsymbol{M}_{k,\cdot}\boldsymbol{\beta}_{k,\cdot}\|_2^2)$. The minimax expected discrepancy for the kth sub-problem is then defined as

$$\inf_{\bar{\boldsymbol{\beta}}_{k,\cdot}} \sup_{\boldsymbol{\beta}^*_{k,\cdot} \in \mathcal{P}_k(P_X, P_{\boldsymbol{A}}, \boldsymbol{\sigma}^2)} \mathbb{E}\left\{\mathcal{E}(\bar{\boldsymbol{\beta}}_{k,\cdot})\right\},$$

where the infimum is taken with respect to all estimators from the data. The next theorem provides an exact expression for the minimax expected discrepancy.

Theorem 9 (Minimax optimality). The exact minimax risk can be expressed as

$$\inf_{\widehat{\boldsymbol{\beta}}_{k,\cdot} \in \mathbb{R}^K} \sup_{\boldsymbol{\beta}^*_{k,\cdot} \in \mathcal{P}_k(P_X, P_{\boldsymbol{A}}, \boldsymbol{\sigma}^2)} \mathbb{E}\left(\mathcal{E}(\widehat{\boldsymbol{\beta}}_{k,\cdot})\right) = \sigma^2_{\max} \operatorname{tr}\left(\boldsymbol{\Sigma}\mathbb{E}(\boldsymbol{H}_k^{-1})\right).$$

Since $\operatorname{tr}(\boldsymbol{\Sigma}\boldsymbol{S}^{-1})$ is strictly convex with respect to $\boldsymbol{S}$ in the positive definite cone, Jensen's inequality yields that $\operatorname{tr}(\boldsymbol{\Sigma}\mathbb{E}(\boldsymbol{H}_k^{-1})) \ge \operatorname{tr}(\boldsymbol{\Sigma}(\mathbb{E}(\boldsymbol{H}_k))^{-1})$. Taking $\boldsymbol{\Sigma} = \boldsymbol{I}$, we have

$$\inf_{\widehat{\boldsymbol{\beta}}_{k,\cdot}} \sup_{\boldsymbol{\beta}^*_{k,\cdot} \in \mathcal{P}_k(P_X, P_{\boldsymbol{A}}, \boldsymbol{\sigma}^2)} \mathbb{E}\big(\|\widehat{\boldsymbol{\beta}}_{k,\cdot} - \boldsymbol{\beta}^*_{k,\cdot}\|^2\big) \ge \sum_{k'=1}^{K} \frac{\sigma^2_{\max}}{\lambda_{k'}(\mathbb{E}(\boldsymbol{H}_k))} \ge \frac{\sigma^2_{\max}}{\lambda_{\min}(\mathbb{E}(\boldsymbol{H}_k))},$$

which matches up with the probabilistic upper bound in Theorem 5. This indicates the non-asymptotic analysis in the previous sections is sharp.

4. Simulation

4.1. Impact of Network Neighborhood Information

The first simulation study analyzes how the network information helps in the estimation. Specifically, for any $n \in \{100, 200, \ldots, 1000\}$ and $K \in \{2, 3, 4\}$, we begin with randomly and uniformly sampling the community memberships for n genes, resulting in the membership matrix $\boldsymbol{Z} \in \{0, 1\}^{n\times K}$. Next, the network $\boldsymbol{A}$ is generated according to the stochastic block model stated in Assumption 1 with $B_{k_1,k_2} \sim \text{Uniform}(0, 0.5) + 0.5 * \mathbb{1}\{k_1 = k_2\}$ and $B_{k_2,k_1} = B_{k_1,k_2}$, for $1 \le k_1 < k_2 \le K$. We subsequently generate the response $\boldsymbol{y}$ according to model (3), where the entries of both the covariate $\boldsymbol{x}$ and the coefficient matrix $\boldsymbol{\beta} \in \mathbb{R}^{K\times K}$ are sampled independently from the standard normal distribution, and the additive noises are drawn from $\epsilon_1, \ldots, \epsilon_n \overset{\text{iid}}{\sim} \mathcal{N}(0, 0.5^2)$. We then evaluate 4 methods that use the network information in different ways: (i) CLSE that uses the network neighborhood information appropriately, (ii) CLSE with $\boldsymbol{A}$ replaced by $\boldsymbol{I}_n$ which completely ignores the interactions between vertices, (iii) CLSE with $\boldsymbol{A}$ replaced by $\boldsymbol{1}_n\boldsymbol{1}_n^\top$ which includes all potential interactions among vertices, and (iv) netcoh, the network cohesion estimator proposed by Li, Levina, and Zhu (2019), which minimizes the penalized loss function $\mathcal{L}(\boldsymbol{\alpha}, \beta) = \|\boldsymbol{y} - \boldsymbol{\alpha} - \beta\boldsymbol{x}\|^2 + \lambda\boldsymbol{\alpha}^\top\left[\operatorname{diag}(\boldsymbol{A}\boldsymbol{1}_n) - \boldsymbol{A}\right]\boldsymbol{\alpha}$.

Clearly, it equips each node with an individual intercept to adjust a linear regression model and penalizes these intercepts according to the graph Laplacian. In practice, we select the hyperparameter by cross-validation over a grid of 100 regularization parameters from 10^{-3} to 10 uniformly spaced in the $\log_{10}$-scale.

In all our experiments of Sections 4 and 5, we employ a variation of SCORE method (Jin 2015; Ke and Jin 2023) to estimate the community memberships for each node when it is not directly available. Precisely, we compute $\boldsymbol{U} \in \mathbb{R}^{n\times K}$ that contains the eigenvectors of $\boldsymbol{A}$ corresponding to the K leading singular values, subsequently normalize each row of $\boldsymbol{U}$ to have unit l_2-norm, and finally apply K-means algorithm to obtain the cluster assignments $\widehat{\boldsymbol{Z}}$. The number of communities is determined by identifying the elbow point in the singular value distribution of $\boldsymbol{A}$ as in (Ji and Jin 2016; Rohe, Qin, and Yu 2016). A scree plot of the leading singular values of the real-life autism gene co-expression network is provided in Figure 4(b). After obtaining any estimator $\widehat{\boldsymbol{\beta}}$ by community-wise least square estimation, the predicted values for $\boldsymbol{y}$ is given by $\widehat{\boldsymbol{y}} = (\widehat{\boldsymbol{Z}}\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{Z}}^\top * \boldsymbol{A})\boldsymbol{x}$. We inspect the estimation error and the prediction error, defined as

$$\begin{aligned} \text{Err}_{\text{est}} &= \tfrac{1}{K^2}\|\widehat{\boldsymbol{Q}}^\top\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{Q}} - \boldsymbol{\beta}^*\|_{\text{F}}^2 \text{ and} \\ \text{Err}_{\text{pred}} &= \tfrac{1}{n}\|\widehat{\boldsymbol{y}} - \boldsymbol{y}\|_2^2, \text{ where} \\ \widehat{\boldsymbol{Q}} &= \operatorname{argmin}_{\boldsymbol{Q}\in\mathbb{G}_K}\|\widehat{\boldsymbol{Z}}\boldsymbol{Q} - \boldsymbol{Z}\|_{\text{F}}^2. \end{aligned}$$

Herein, $\mathbb{G}_K$ denotes the set of $K \times K$ permutation matrices. Clearly, $\widehat{\boldsymbol{y}}$ is invariant to the permutation among the communities, while $\widehat{\boldsymbol{\beta}}$ is not. That is why we need to search for the best permutation that minimizes the Hamming distance of the community memberships encoded in $\widehat{\boldsymbol{Z}}$ and $\boldsymbol{Z}$, and then define the estimation error accordingly.

Results in the logarithm scale are shown in Figure 2, where the trends for small values have been zoomed in while those for large values have been zoomed out. As we can see, both the estimation and prediction errors of CLSE for different community sizes are decreasing in the sample size n and approaching zero quickly. On the other hand, without properly using the network information $\boldsymbol{A}$, both the estimation and prediction errors cannot be controlled even with large sample sizes. In summary, this simulation study showcases the asymptotic consistency of the CLSE estimator and suggests the necessity of using the network neighborhood information appropriately.

4.2. Impact of Community Structure

The second simulation study considers various regression coefficient models, including (i) full model: $\boldsymbol{\beta} \in \mathbb{R}^{K\times K}$, (ii) row

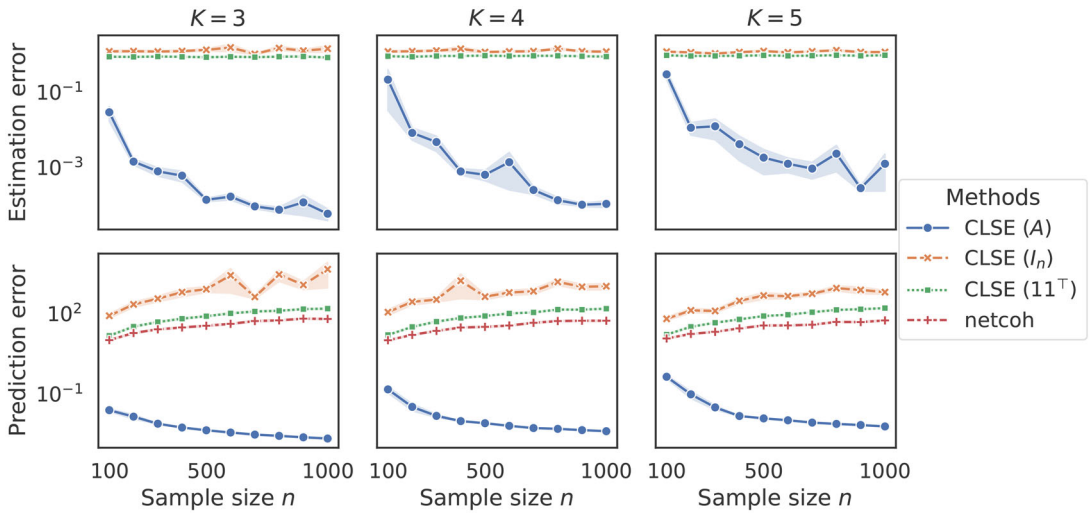

Figure 2. Estimation and prediction errors in the log scale of experiments in Section 4.1 with varying numbers of communities. The shaded regions represent the standard errors around the average values computed over 200 simulated datasets. For netcoh, only the prediction errors are available and presented.

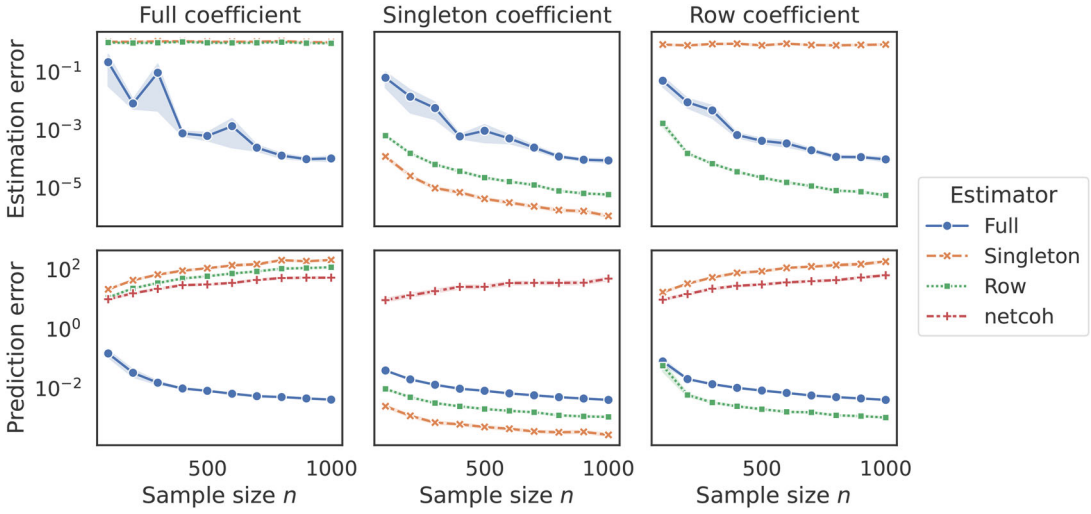

Figure 3. Estimation and prediction errors in the log scale of experiments in Section 4.2 with different coefficient structures. The shaded regions represent the standard errors around the average values computed over 200 simulated datasets.

model: $\boldsymbol{\beta} = \mathbf{1}_K \boldsymbol{\beta}_0^\top$ for $\boldsymbol{\beta}_0 \in \mathbb{R}^K$, and (iii) singleton model: $\boldsymbol{\beta} = \beta_0 \mathbf{1}_K \mathbf{1}_K^\top$ for $\beta_0 \in \mathbb{R}$. Denote by $\widehat{\boldsymbol{\beta}}^{\text{full}}$ the solution to (6) corresponding to the full model. The counterparts to the row and singleton models are derived as follows. Under the setting of row model, (6) reduces to a multiple linear regression problem:

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_0^{\text{row}} &= \underset{\boldsymbol{\beta}_0 \in \mathbb{R}^K}{\operatorname{argmin}} \frac{1}{2n} \|\boldsymbol{y} - (\boldsymbol{Z}\mathbf{1}_K \boldsymbol{\beta}_0^\top \boldsymbol{Z}^\top * \boldsymbol{A})\boldsymbol{x}\|_2^2 \\ &= (\boldsymbol{Z}^\top \operatorname{diag}(\boldsymbol{x}) \boldsymbol{A}^2 \operatorname{diag}(\boldsymbol{x}) \boldsymbol{Z})^{-1} \boldsymbol{Z}^\top \operatorname{diag}(\boldsymbol{x}) \boldsymbol{A} \boldsymbol{y},\end{aligned}$$

which yields the row estimator $\widehat{\boldsymbol{\beta}}^{\text{row}} = \mathbf{1}_K (\widehat{\boldsymbol{\beta}}_0^{\text{row}})^\top$. Under the singleton model setup, (6) reduces to a simple linear regression problem:

$$\widehat{\beta}_0^{\text{sgtn}} = \underset{\beta_0 \in \mathbb{R}}{\operatorname{argmin}} \frac{1}{2n} \|\boldsymbol{y} - (\boldsymbol{A}\boldsymbol{x})\beta_0\|_2^2 = \frac{\boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{y}}{\boldsymbol{x}^\top \boldsymbol{A}^2 \boldsymbol{x}}.$$

which yields the singleton estimator $\widehat{\boldsymbol{\beta}}^{\text{sgtn}} = \widehat{\beta}_0^{\text{sgtn}} \mathbf{1}_K \mathbf{1}_K^\top$. Different estimators use the community information at different levels.

We compare the estimation and prediction errors of the three estimators in all three generating schemes of regression coefficients with $K = 4$ and varying $n \in \{100, \ldots, 1000\}$. The results are shown in Figure 3. It is expected that each estimator works best under its own well-specified setting. However, both the singleton and row estimators are sensitive to model mis-specifications, while the full estimator adapts well due to its generality, with errors tending to zero relatively fast. These results suggest that the full estimator is capable of using the community information, and it does not suffer much when only a partial of this information is relevant. Note that netcoh fails to provide satisfying results in all scnarios, including in the singleton model setup. Though it penalizes the intercept vector such that nodes sharing similar linking pattern have similar intercept adjustments, it ignores the more subtle neighborhood information, and the predicted value of y_i given by netcoh is constructed from the estimate of α_i, β, and x_i only, without considering the covariates of the nodes in the neighborhood of node i.

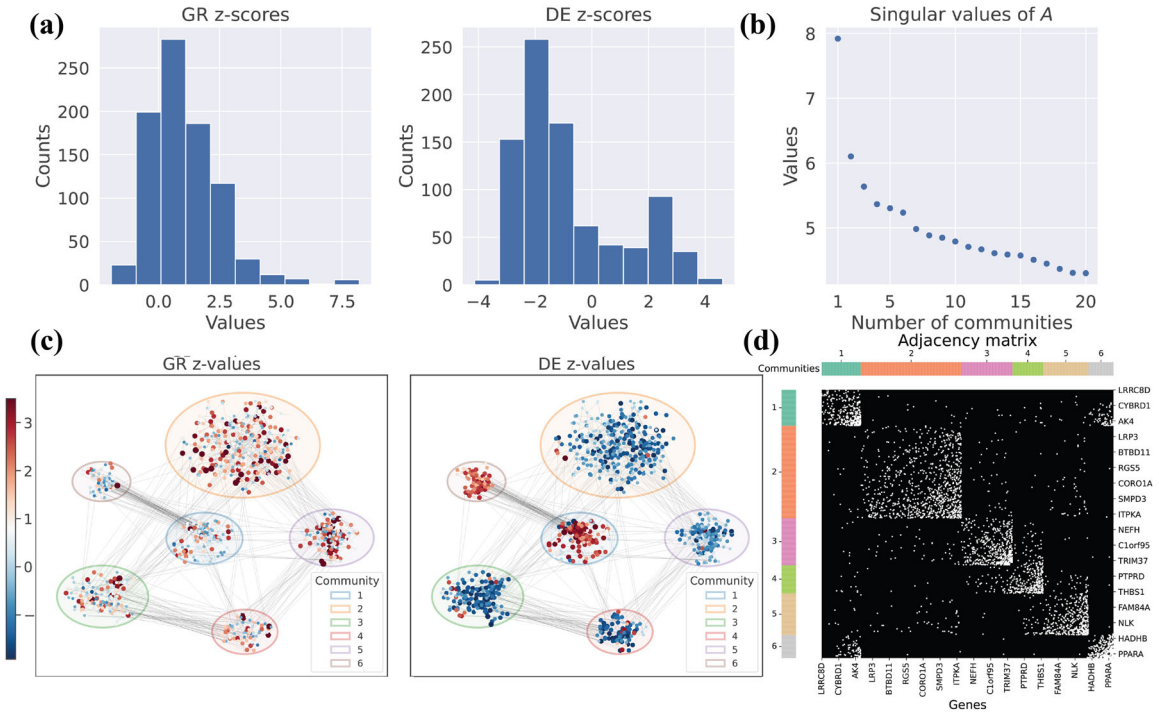

Figure 4. Visualization of ASD data and detected communities. (a) The histograms of GR one-sided z-scores ($\boldsymbol{x}$) and DE two-sided z-scores ($\boldsymbol{y}$) for 864 substantial autism genes. (b) The scree plot of the singular values of $\boldsymbol{A}$. (c) The gene co-expression network colored by z-values and grouped by estimated communities. (d) The adjacency matrix $\boldsymbol{A}$ colored by connectivity (white for 1 and black for 0) and ordered by estimated assignments.

5. Autism Spectrum Disorder Genetic Association

5.1. Background

Autism spectrum disorder (ASD) primarily stems from genetic variations, either inherited or arising spontaneously in individuals. This genetic diversity plays a crucial role in ASD's prevalence, with de novo exonic variations being particularly valuable for linking specific genes to the disorder (Fu et al. 2022). The Transmission and De Novo Association (TADA) method (He et al. 2013) has been pivotal in pinpointing genes susceptible to ASD by analyzing mutation frequencies in family trios, leading to the identification of numerous ASD-associated genes, yet many remain undiscovered. Despite the identification of thousands of genes with differential expression in ASD (Gandal et al. 2022), there is minimal overlap between DE genes and the GR genes deemed significant by TADA in the two studies. In this section, we attempt to using the gene co-expression networks to integrate these disparate data sources and disentangle the impact of the GR scores from TADA on the DE scores.

5.2. Data and Preprocessing

We use two types of genomics data: (a) The DE and GR test statistics are originally obtained from the differential expression analysis by Gandal et al. (2022) and the TADA analysis by Fu et al. (2022), respectively. We use the GR and DE z-values as the covariate $\boldsymbol{x} \in \mathbb{R}^n$ and the response $\boldsymbol{y} \in \mathbb{R}^n$, respectively. The histograms of the two scores are shown in Figure 4(a). (b) A whole cortex gene expression data (bulk RNA-sequencing data) on neurotypical individuals is also available from the previous study by Gandal et al. (2022). Based on gene expression data, Liu, Lei, and Roeder (2015) use the partial neighborhood selection algorithm to obtain a sparse network $\boldsymbol{A} \in \{0,1\}^{n\times n}$ of approximately scale-free form; though one can also use other networks, such as the protein-protein interaction networks. Finally, we restrict the analysis to a subset of 864 substantial autism genes, which is identified by using the partial neighborhood selection (PNS) algorithm of DAWN (Liu et al. 2014).

Based on the network $\boldsymbol{A}$, we first perform community detection to uncover the genes' community memberships. We first visualize the 20 leading singular values of $\boldsymbol{A}$ in Figure 4(b). It is clear that the singular values decay quickly and become smaller than 5 after the first 6 leading singular values, suggesting that there shall be 6 communities. Consequently, we select $K = 6$. We have then used the same community detection method as in Section 4 to obtain the estimated cluster membership $\widehat{\boldsymbol{Z}}$. Grouping the genes into the detected communities, we observe a clear block structure of the adjacency matrix, as shown in Figure 4(d).

The estimated cluster membership matrix allows us to visually compare the two sources of statistical evidence. As shown in Figure 4(c), for most of the communities, the GR and DE scores share similar patterns. The genes in a cluster with enriched GR scores in terms of absolute values, such as communities 1 and 2, typically also have large DE scores, while the genes in communities 5 and 6 have both small GR and DE scores simultaneously. These suggest that the genes in such clusters may be positively regulated by their genetic variations of the same gene module. On the other hand, genes in communities 3 and 4 have moderate GR scores while much larger DE ones. Given the scarce evidence of genetic variation, the genes in these two communities are likely to be regularized by other genetic modules. These observations motivate us to analyze the

Table 1. Estimation and inference results of $\widehat{\boldsymbol{\beta}}$.

Target Comm. (y)	Source Comm. (x)					
	1	2	3	4	5	6
1	0.325 ± 0.070 (∗∗∗) 0.0000	−0.037 ± 0.347 0.9155	−1.057 ± 1.026 0.3029	−0.171 ± 1.440 0.9057	−0.193 ± 0.572 0.7362	0.379 ± 0.146 (∗∗) 0.0092
2	−0.256 ± 0.270 0.3422	−0.197 ± 0.014 (∗∗∗) 0.0000	−0.537 ± 0.148 (∗∗∗) 0.0003	0.041 ± 1.096 0.9704	−0.222 ± 0.050 (∗∗∗) 0.0000	−2.067 ± 6.931 0.7655
3	−0.947 ± 0.650 0.1448	−0.179 ± 0.183 0.3262	−0.305 ± 0.027 (∗∗∗) 0.0000	−0.107 ± 0.107 0.3135	0.390 ± 16.968 0.9816	1.511 ± 2.472 0.5411
4	0.359 ± 0.428 0.4021	−0.724 ± 0.568 0.2023	−0.224 ± 0.098 (∗) 0.0221	−0.346 ± 0.043 (∗∗∗) 0.0000	−0.274 ± 0.098 (∗∗) 0.0051	−1.925 ± 0.362 (∗∗∗) 0.0000
5	0.054 ± 0.494 0.9123	−0.227 ± 0.046 (∗∗∗) 0.0000	−0.500 ± 2.795 0.8581	−0.239 ± 0.093 (∗) 0.0101	−0.154 ± 0.014 (∗∗∗) 0.0000	−0.520 ± 0.917 0.5707
6	0.407 ± 0.265 0.1239	0.356 ± 0.239 0.1361	2.164 ± 0.134 (∗∗∗) 0.0000	1.158 ± 4.500 0.7969	−1.277 ± 0.867 0.1406	0.490 ± 0.068 (∗∗∗) 0.0000

NOTE: The kth row of the table corresponds to $\widehat{\boldsymbol{\beta}}_{k,\cdot}$. Within each cell, the point estimate and the estimated standard deviation are given on top of the cell, while the significance level and the corresponding p-value are given at the bottom. For the significance levels, (∗∗∗), (∗∗), and (∗), indicate that the p-value locates in $[0, 0.001]$, $(0.001, 0.01]$, and $(0.01, 0.05]$, respectively. The significant positive and negative coefficients are highlighted in magenta and cyan, respectively.

interplay of different gene modules and understand how genetic variations affect gene expressions among different clusters by performing network-based neighborhood regression coupled with community-wise analysis.

5.3. Neighborhood Regression on ASD Genetic Association

Based on the evidence $(\boldsymbol{x}, \boldsymbol{y})$ and network $\boldsymbol{A}$, along with the estimated cluster membership $\widehat{\boldsymbol{Z}}$, we compute the estimated values of the CLSE and that given by netcoh. We find that the in-sample prediction error of CLSE is 0.5781, which is substantially smaller than that of netcoh (3.0719) in terms of their respective objective functions, showing that CLSE delivers more trustworthy result than netcoh. This also reflects that the genes actually co-express with each other, which can be efficiently characterized by the network-based neighborhood regression model, while netcoh fails to leverage the GR scores in the neighborhood to the DE score of any particular gene.

We next focus on CLSE only and perform individual hypothesis testing on $H_0 : \beta_{k_1,k_2} = 0$ versus $H_1 : \beta_{k_1,k_2} \neq 0$ for $k_1, k_2 \in [K]$. The heteroscedasticity-consistent (HC) standard errors (MacKinnon and White 1985) of the estimators are used to compute p-values. All of the results are summarized in Table 1. From Table 1, we observe that the GR scores have significant effects on the DE scores within the same community, which is expected because if a gene module is associated with genetic variations of ASD, then the gene expression levels of this module will also be affected. Though the intra-cluster interaction of the two modalities is important, we are more interested in the inter-cluster interaction, as it will shed new light on how one gene module regulates the others. By exploring the regulatory mechanisms and potential influences between distinct groups of genes, we can gain insights into the broader network dynamics at play. For this purpose, we further restrict our analysis to the two most significant inter-cluster coefficients $\beta_{4,6}$ and $\beta_{6,3}$, corresponding to directional effects from community 6 to community 4 and from community 3 to community 6, respectively.

By matching the gene modules identified in Gandal et al. (2022, Fig. 7), clusters 3 and 4 contain genes that are mostly enriched in three cell types: excitatory neurons, inhibitory neurons, and oligodendrocytes, and cluster 6 contains genes that are mostly enriched in astrocytes, which have a potential impact on neuronal function and connectivity and are critical in the pathology of ASD (Gandal et al. 2022).

From the gene ontology (GO) analysis result in Figure 5, community 6 is enriched for GO terms related to bone remodeling, ossification, biomineralization, and regulation of nucleotide metabolism, while community 4 is enriched for GO terms involved in synaptic vesicle cycling, neurotransmitter transport, and proton transport processes at the synapse. The estimated coefficient in Table 1 suggests a negative impact of community 6 on community 4, which could potentially be explained by the fact that excessive bone remodeling and mineralization processes (community 6) may disrupt normal synaptic functions (community 4) by altering the ionic balance or metabolic processes required for neurotransmission for ASD.

In addition, community 3, which is enriched for GO terms related to exocytosis, synaptic vesicle priming, nerve impulse transmission, and ion transport regulation, has a positive impact on community 6. This positive impact could be due to the fact that regulated exocytosis and ion transport processes (community 3) may facilitate the release of factors or signaling molecules that promote bone remodeling, ossification, and biomineralization (community 6). The interpretation is sorely based on the GO term enrichment of biological processes. The specific molecular mechanisms underlying the observed impacts would require further experimental validation and investigation.

5.4. A Novel Adjusted Coefficient of Determination

Since our primary goal is to understand the module-wise interactions, we employ β_{k_1,k_2} to capture the average causal effect from GR scores among genes in one community to the DE scores of those in another. Biologically, genes within functional modules or pathways often behave similarly and exhibit coordinated expression and regulation. Admittedly, not all genes within a module behave identically, while the assumption reasonably leverages genetic heterogeneity and allows for meaningful and broad biological insights.

To verify the validity of the assumption in a data-driven way, analogous to linear regression, we proposed the following adjusted coefficient of determination $R^{2,net}_{adj}$ for network-based neighborhood regression

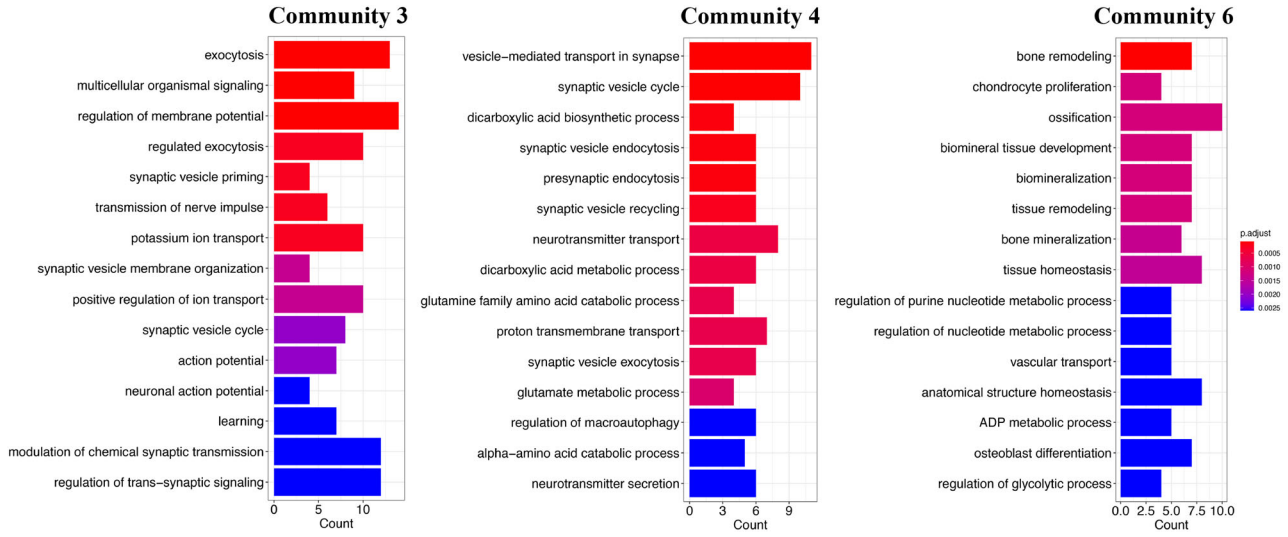

Figure 5. Top gene ontology terms for genes in community 3, 4, and 6.

$$R^{2,net}_{adj} = 1 - \frac{n-1}{n-K^2} \cdot \frac{\sum_{k=1}^{K} \|\boldsymbol{y}^{(k)} - \widehat{\boldsymbol{y}}^{(k)}\|^2}{\|\boldsymbol{y} - \widehat{\boldsymbol{y}}^{(ols)})\|^2},$$

where $\boldsymbol{y}^{(k)}, \widehat{\boldsymbol{y}}^{(k)}$, and $\widehat{\boldsymbol{y}}^{(k,ols)}$ are the actual response, the predicted response by the proposed method, and predicted response by ordinary least square in the kth community, respectively. Note that since the proposed network-based neighborhood regression does not contain an intercept, we chose the ordinary least square as the baseline for fair comparison. The value of $R^{2,net}_{adj}$ for the real data being studied is greater than 0.5867. This implies that when the block structure of the coefficient matrix is imposed, over 58.67% of the uncertainty of the actual response $\boldsymbol{y}$ can be explained by the neighborhood regression model, demonstrating the rationale behind the block-wise effects.

6. Discussion

This article incorporates network-based neighborhood information to bridge the predictor and response in the regression setup. Potential extensions of the current framework include allowing multivariate predictors and multivariate responses, extending the neighborhood regression to generalized neighborhood regression for binary or counting responses as in generalized linear models (Du, Wasserman, and Roeder 2023), and considering more general network structures, such as weighted, directed, multi-layer (Lei and Lin 2023), and hypergraph networks (Zhen and Wang 2023). In the supplement, we suggest a potential extension for multiple covariates, while its statistical property remains unclear. Moreover, it will be interesting to slightly relax the exact stochastic block model for the network data and the block structure for the coefficient matrix to allow more heterogeneity and flexibility among the entities, such as using the latent space model for network modeling. Besides, the proposed neighborhood regression framework is also closely related to the sum aggregator of graph neural network (GNN) (Xu et al. 2019), which is probably the most expressive among a number of classes of GNNs and is as powerful as the Weisfeiler-Lehman graph isomorphism test. Exploring connections between the proposed method and other aggregation operators in GNN with heterogeneous structures presents a promising avenue for future research.

Supplementary Materials

The supplementary material includes all technical proofs, necessary lemmas, and Python implementations of the paper's numerical experiments.

Acknowledgments

We thank the associate editor, three anonymous referees, and the reproducibility reviewer, whose constructive comments and suggestions have led to significant improvements of the article. We would also like to express our great gratitude to Kathryn Roeder, Bernie Devlin, Jinjin Tian, and Maya Shen for enlightening discussion and for sharing the processed data. This article was initiated when the authors met at the 2023 *IMS Young Mathematical Scientists Forum - Statistics and Data Science*, hosted by the National University of Singapore (NUS). We truly appreciate NUS for the warm invitation.

Disclosure Statement

The authors report there is no competing interest to declare.

ORCID

Yaoming Zhen http://orcid.org/0000-0002-3724-9200
Jin-Hong Du http://orcid.org/0000-0001-9683-4146

References

Abbe, E. (2018), "Community Detection and Stochastic Block Models: Recent Developments," *Journal of Machine Learning Research*, 18, 1–86. [2,5]

Adamson, B., Norman, T. M., Jost, M., Cho, M. Y., Nuñez, J. K., Chen, Y., Villalta, J. E., Gilbert, L. A., Horlbeck, M. A., Hein, M. Y., et al. (2016), "A Multiplexed Single-Cell Crispr Screening Platform Enables Systematic Dissection of the Unfolded Protein Response," *Cell*, 167, 1867–1882. [3]

Athreya, A., Fishkind, D. E., Tang, M., Priebe, C. E., Park, Y., Vogelstein, J. T., Levin, K., Lyzinski, V., Qin, Y., and Sussman, D. L. (2018), "Statistical Inference on Random Dot Product Graphs: A Survey," *Journal of Machine Learning Research*, 18, 1–92. [2]

Brohee, S., and Van Helden, J. (2006), "Evaluation of Clustering Algorithms for Protein-Protein Interaction Networks," *BMC Bioinformatics*, 7, 1–19. [1]

Celisse, A., Daudin, J.-J., and Pierre, L. (2012), "Consistency of Maximum-Likelihood and Variational Estimators in the Stochastic Block Model," *Electronic Journal of Statistics*, 6, 1847–1899. [2]

de la Pena, V. H., and Montgomery-Smith, S. J. (1995),"'Inequalities for the Tail Probabilities of Multivariate u-statistics," *The Annals of Probability*, 23, 806–816. [6]

Du, J.-H., Wasserman, L., and Roeder, K. (2023), "Simultaneous Inference for Generalized Linear Models with Unmeasured Confounders," ArXiv preprint arXiv:2309.07261. [12]

Eicker, F. (1963), "Asymptotic Normality and Consistency of the Least Squares Estimators for Families of Linear Regressions," *The Annals of Mathematical Statistics*, 34, 447–456. [5]

Fahrmeir, L., and Kaufmann, H. (1985), "Consistency and Asymptotic Normality of the Maximum Likelihood Estimator in Generalized Linear Models," *The Annals of Statistics*, 13, 342–368. [5]

Fu, J. M., Satterstrom, F. K., Peng, M., Brand, H., Collins, R. L., Dong, S., Wamsley, B., Klei, L., Wang, L., Hao, S. P., et al. (2022), "Rare Coding Variation Provides Insight into the Genetic Architecture and Phenotypic Context of Autism," *Nature Genetics*, 54, 1320–1331. [3,10]

Gandal, M. J., Haney, J. R., Wamsley, B., Yap, C. X., Parhami, S., Emani, P. S., Chang, N., Chen, G. T., Hoftman, G. D., de Alba, D., et al. (2022), "Broad Transcriptomic Dysregulation Occurs across the Cerebral Cortex in asd," *Nature*, 611, 532–539. [1,3,10,11]

Gao, C., and Ma, Z. (2021), "Minimax Rates in Network Analysis: Graphon Estimation, Community Detection and Hypothesis Testing," *Statistical Science*, 36, 16–33. [2]

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. (2022), "Surprises in High-Dimensional Ridgeless Least Squares Interpolation," *The Annals of Statistics*, 50, 949–986. [4]

He, X., Sanders, S. J., Liu, L., De Rubeis, S., Lim, E. T., Sutcliffe, J. S., Schellenberg, G. D., Gibbs, R. A., Daly, M. J., Buxbaum, J. D., et al. (2013), "Integrated Model of De Novo and Inherited Genetic Variants Yields Greater Power to Identify Risk Genes," *PLoS Genetics*, 9, e1003671. [10]

Hesamian, G., Johannssen, A., and Chukhrova, N. (2024), "An Explainable Fused Lasso Regression Model for Handling High-Dimensional Fuzzy Data," *Journal of Computational and Applied Mathematics*, 441, 115721. [3]

Hu, Y., and Wang, W. (2024), "Network-Adjusted Covariates for Community Detection," *Biometrika*, 111, 1221–1240. [2]

Ji, P., and Jin, J. (2016), "Coauthorship and Citation Networks for Statisticians," *The Annals of Applied Statistics*, 10, 1779–1812. [8]

Jin, J. (2015), "Fast Community Detection by SCORE," *The Annals of Statistics*, 43, 57–89. [2,8]

Jin, X., Simmons, S. K., Guo, A., Shetty, A. S., Ko, M., Nguyen, L., Jokhi, V., Robinson, E., Oyler, P., Curry, N., et al. (2020), "In Vivo Perturb-seq Reveals Neuronal and Glial Abnormalities Associated with Autism Risk Genes," *Science*, 370, eaaz6063. [3]

Ke, Z. T., and Jin, J. (2023), "Special Invited Paper: The Score Normalization, Especially for Heterogeneous Network and Text Data," *Stat*, 12, e545. [8]

Langfelder, P., and Horvath, S. (2008), "Wgcna: An r Package for Weighted Correlation Network Analysis," *BMC Bioinformatics*, 9, 1–13. [3]

Le, C. M., and Li, T. (2022), "Linear Regression and its Inference on Noisy Network-Linked Data," *Journal of the Royal Statistical Society*, Series B, 84, 1851–1885. [1,2]

Lee, C., and Wilkinson, D. J. (2019), "A Review of Stochastic Block Models and Extensions for Graph Clustering," *Applied Network Science*, 4, 1–50. [5]

Lei, J., Chen, K., and Lynch, B. (2020), "Consistent Community Detection in Multi-Layer Network Data," *Biometrika*, 107, 61–73. [5]

Lei, J., and Lin, K. Z. (2023), "Bias-tted Spectral Clustering in Mmulti-Layer Stochastic Block Models," *Journal of the American Statistical Association*, 118, 2433–2445. [6,7,12]

Lei, J., and Rinaldo, A. (2015), "Consistency of Spectral Clustering in Stochastic Block Models," *The Annals of Statistics*, 43, 215–237. [2]

Li, T., Levina, E., and Zhu, J. (2019), "Prediction Models for Network-Linked Data," *The Annals of Applied Statistics*, 13, 132–164. [1,2,8]

Liu, L., Lei, J., and Roeder, K. (2015), "Network Assisted Analysis to Reveal the Genetic Basis of Autism," *The Annals of Applied Statistics*, 9, 1571–1600. [1,3,5,10]

Liu, L., Lei, J., Sanders, S. J., Willsey, A. J., Kou, Y., Cicek, A. E., Klei, L., Lu, C., He, X., Li, M., et al. (2014), "Dawn: A Framework to Identify Autism Genes and Subnetworks Using Gene Expression and Genetics," *Molecular Autism*, 5, 1–18. [1,10]

Liu, L., Sabo, A., Neale, B. M., Nagaswamy, U., Stevens, C., Lim, E., Bodea, C. A., Muzny, D., Reid, J. G., Banks, E., et al. (2013), "Analysis of Rare, Exonic Variation Amongst Subjects with Autism Spectrum Disorders and Population Controls," *PLoS Genet.* 9, e1003443. [1]

MacKinnon, J. G., and White, H. (1985), "Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties," *Journal of Econometrics*, 29, 305–325. [11]

Merlevède, F., Peligrad, M., and Rio, E. (2011), "A Bernstein Type Inequality and Moderate Deviations for Weakly Dependent Sequences," *Probability Theory and Related Fields*, 151, 435–474. [5]

Newman, M. E., and Clauset, A. (2016), "Structure and Inference in Annotated Networks," *Nature Communications*, 7, 11863. [2]

Patil, P., Du, J.-H., and Tibshirani, R. (2024), "Optimal Ridge Regularization for Out-of-Distribution Prediction," in *Forty-first International Conference on Machine Learning*. [4]

Paul, S., and Chen, Y. (2020), "A Random Effects Stochastic Block Model for Joint Community Detection in Multiple Networks with Applications to Neuroimaging," *The Annals of Applied Statistics*, 14, 993–1029. [1]

Raftery, A. E., Niu, X., Hoff, P. D., and Yeung, K. Y. (2012), "Fast Inference for the Latent Space Network Model Using a Case-Control Approximate Likelihood," *Journal of Computational and Graphical Statistics*, 21, 901–919. [2]

Rohe, K., Qin, T., and Yu, B. (2016), "Co-Clustering Directed Graphs to Discover Asymmetries and Directional Communities," *Proceedings of the National Academy of Sciences*, 113, 12679–12684. [8]

Rudelson, M., and Vershynin, R. (2013), "Hanson-Wright Inequality and Sub-Gaussian Concentration," *Electronic Communications in Probability*, 18, 1–9. [6]

Shang, R., Bai, J., Jiao, L., and Jin, C. (2013), "Community Detection based on Modularity and an Improved Genetic Algorithm," *Physica A: Statistical Mechanics and its Applications*, 392, 1215–1231. [2]

Tropp, J. A. (2012), "User-Friendly Tail Bounds for Sums of Random Matrices," *Foundations of Computational Mathematics*, 12, 389–434. [6]

Wainwright, M. J. (2019), *High-Dimensional Statistics: A Non-Asymptotic Viewpoint* (Vol. 48), Cambridge: Cambridge University Press. [5]

Wang, P., Li, Q., Shen, D., and Liu, Y. (2023), "High-Dimensional Factor Regression for Heterogeneous Subpopulations," *Statistica Sinica*, 33, 1–27. [2]

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. (2019), "How Powerful Are Graph Neural Networks?" in *International Conference on Learning Representations*. [12]

Xu, S., Zhen, Y., and Wang, J. (2023), "Covariate-Assisted Community Detection in Multi-Layer Networks," *Journal of Business & Economic Statistics*, 41, 915–926. [2,5]

Yan, B., and Sarkar, P. (2021), "Covariate Regularized Community Detection in Sparse Graphs," *Journal of the American Statistical Association*, 116, 734–745. [2]

Yu, F., Yang, A. Y., and Zhang, T. (2025), "A Graph Decomposition-based Approach for the Graph-Fused Lasso," *Journal of Statistical Planning and Inference*, 235, 106221. [3]

Yuan, Y., and Qu, A. (2021), "Community Detection with Dependent Connectivity," *The Annals of Statistics*, 49, 2378–2428. [5]

Zhang, J., Sun, W. W., and Li, L. (2020), "Mixed-Effect Time-Varying Network Model and Application in Brain Connectivity Analysis,"

Journal of the American Statistical Association, 115, 2022–2036. [1,4]

Zhang, J., Sun, W. W., and Li, L. (2023), "Generalized Connectivity Matrix Response Regression with Applications in Brain Connectivity Studies," *Journal of Computational and Graphical Statistics*, 32, 252–262. [4]

Zhang, X., Xu, G., and Zhu, J. (2022), "Joint Latent Space Models for Network Data with High-Dimensional Node Variables," *Biometrika*, 109, 707–720. [2]

Zhang, Y., Levina, E., and Zhu, J. (2016), "Community Detection in Networks with Node Features," *Electronic Journal of Statistics*, 10, 3153–3178. [2]

——— (2017), "Estimating Network Edge Probabilities by Neighbourhood Smoothing," *Biometrika*, 104, 771–783. [1]

——— (2012), "Consistency of Community Detection in Networks Under Degree-Corrected Stochastic Block Models," *The Annals of Statistics*, 40, 2266–2292. [3]

Zhen, Y., and Wang, J. (2023), "Community Detection in General Hypergraph via Graph Embedding," *Journal of the American Statistical Association*, 118, 1620–1629. [5,7,12]

Zhou, L., Sun, S., Fu, H., and Song, P. X.-K. (2022), "Subgroup-Effects Models for the Analysis of Personal Treatment Effects," *The Annals of Applied Statistics*, 16, 80–103. [2]