National University of Singapore | Institute for Mathematical Sciences

SCIENTIFIC REPORTS

Frontiers of Statistical Network Analysis: Inference, Tensors and Beyond

12 May 2025–30 May 2025

Organizing Committee

Jialiang Li

National University of Singapore

Dong Xia

Hong Kong University of Science and Technology

Yuan Zhang

Ohio State University

CONTENTS PAGE

HIGHER-ORDER GRAPHON THEORY: FLUCTUATIONS, INFERENCE, AND DEGENERACIES

BHASWAR B. BHATTACHARYA

Classification AMS 2020: 05C80, 60F05, 05C60

Keywords: Inhomogeneous random graphs, network analysis, generalized U-statistics, subgraph counts.

Exchangeable random graphs, which include some of the most widely studied network models, play a central role in statistical network analysis. Graphons, which are the central objects in graph limit theory, provide a natural way to sample exchangeable random graphs. It is well known that network moments (motif/subgraph counts) identify a graphon (up to an isomorphism), hence, understanding the sampling distribution of subgraph counts in random graphs sampled from a graphon is pivotal for nonparametric network inference. Although there are a few results regarding the asymptotic normality of subgraph counts in graphon models, for many commonly appearing graphons this distribution is degenerate. This degeneracy phenomenon was overlooked until very recently and its consequences in network inference have remained unexplored. Towards this, in joint works with Chatterjee and Janson [1] and Chatterjee and Dan [2] we obtain the following results:

- We derive the joint asymptotic distribution of any finite collection of network moments in random graphs sampled from a graphon, that includes both the non-degenerate case (where the distribution is Gaussian) as well as the degenerate case (where the distribution has both Gaussian or non-Gaussian components). This provides the higher-order fluctuation theory for subgraph counts in the graphon model.
- We develop a novel multiplier bootstrap for graphons that consistently approximates the limiting distribution of the network moments (both in the Gaussian and non-Gaussian regimes). Using this and a procedure for testing degeneracy, we construct joint confidence sets for any finite collection of motif densities. This provides a general framework for statistical inference based on network moments in the graphon model.

We also discuss various structure theorems and open questions about degeneracies of the limiting distribution and connections to quasirandom graphs.

REFERENCES

[1] B. B. Bhattacharya, A. Chatterjee, and S. Janson, Fluctuations of subgraph counts in graphon based random graphs, *Combinatorics, Probability, and Computing*, Vol. 32 (3), 428–464, 2023.

[2] A. Chatterjee, S. Dan, B. B. Bhattacharya, Higher-order graphon theory: Fluctuations, degeneracies, and inference, *arXiv:2404.13822*, 2024.

UNIVERSITY OF PENNSYLVANIA
Email address: bhaswar@wharton.upenn.edu

AUTOREGRESSIVE NETWORKS WITH DEPENDENT EDGES

JINYUAN CHANG

Classification AMS 2020: 62F12; 62M10; 62A09

Keywords: conditional independence, dynamic networks, maximum likelihood estimation, stylized features of network data, transitivity.

Dynamic network modeling with dependent edges is practically important but challenging. In the absence of edge dependence, it becomes impossible to capture several stylized features commonly observed in real-world network data, such as transitivity, density dependence, and community structures. These features are crucial for accurately modeling the dynamics of networks in fields like social interactions, communication networks, and organizational behavior. However, including edge dependence complicates the dynamic structure of network processes and makes statistical analysis more challenging.

This talk introduces a novel autoregressive (AR) framework for modeling dynamic networks with dependent edges. Following [4, 5], we specify the transition probabilities of forming a new edge or dissolving an existing edge between each pair of nodes explicitly depending on its history and allow those probabilities depending on the histories of other edge processes. Similar to [3, 6, 7], we assume that the edges are conditionally independent given their joint histories. This makes both statistical inference and theoretical analysis more transparent. Consider a dynamic network process defined on p nodes denoted by $1, \ldots, p$. Let $\mathbf{X}_t \equiv \left(X_{i,j}^t\right)_{p \times p}$ be the adjacency matrix at time t, where $X_{i,j}^t = 1$ denotes the existence of an edge between nodes i and j at time t, and 0 otherwise. For simplicity, we only consider undirected networks without self-loops, i.e. $X_{i,i}^t \equiv 0$ and $X_{i,j}^t = X_{j,i}^t$. The AR networks with dependent edges is defined as follows.

Definition 1 (AR(m) networks)**.** *Conditionally on $\{\mathbf{X}_s\}_{s \leqslant t-1}$, the edges $\left\{X_{i,j}^t\right\}_{1 \leqslant i < j \leqslant p}$ are mutually independent with*

$$\begin{aligned}
\alpha_{i,j}^{t-1} &\equiv \mathbb{P}\left(X_{i,j}^t = 1 \mid X_{i,j}^{t-1} = 0, \mathbf{X}_{t-1} \backslash X_{i,j}^{t-1}, \mathbf{X}_{t-k} \text{ for } k \geqslant 2\right) \\
&= \mathbb{P}\left(X_{i,j}^t = 1 \mid X_{i,j}^{t-1} = 0, \mathbf{X}_{t-1} \backslash X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \ldots, \mathbf{X}_{t-m}\right) , \\
\beta_{i,j}^{t-1} &\equiv \mathbb{P}\left(X_{i,j}^t = 0 \mid X_{i,j}^{t-1} = 1, \mathbf{X}_{t-1} \backslash X_{i,j}^{t-1}, \mathbf{X}_{t-k} \text{ for } k \geqslant 2\right) \\
&= \mathbb{P}\left(X_{i,j}^t = 0 \mid X_{i,j}^{t-1} = 1, \mathbf{X}_{t-1} \backslash X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \ldots, \mathbf{X}_{t-m}\right) ,
\end{aligned}$$

where $m \geqslant 1$ is an integer.

An AR(m) network process defined above is a Markov chain with order m. Based on Definition 1, we have

$$\mathbb{P}\left(X_{i,j}^t = 1 \mid \mathbf{X}_{t-1}, \ldots, \mathbf{X}_{t-m}\right) = \alpha_{i,j}^{t-1} + X_{i,j}^{t-1}\left(1 - \alpha_{i,j}^{t-1} - \beta_{i,j}^{t-1}\right) \equiv \gamma_{i,j}^{t-1}.$$

Clearly edges $X_{i,j}^t$, for different (i,j), are not independent with each other. We may impose various forms for the conditional probabilities $\alpha_{i,j}^{t-1}$ and $\beta_{i,j}^{t-1}$ to reflect different stylized features of network data. For any $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)^\top \in \boldsymbol{\Theta}$, write

$$\begin{aligned} \alpha_{i,j}^{t-1}(\boldsymbol{\theta}) &= f_{i,j}\big(\mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \ldots, \mathbf{X}_{t-m}; \boldsymbol{\theta}\big)\,, \\ \beta_{i,j}^{t-1}(\boldsymbol{\theta}) &= g_{i,j}\big(\mathbf{X}_{t-1} \setminus X_{i,j}^{t-1}, \mathbf{X}_{t-2}, \ldots, \mathbf{X}_{t-m}; \boldsymbol{\theta}\big)\,, \\ \gamma_{i,j}^{t-1}(\boldsymbol{\theta}) &= \alpha_{i,j}^{t-1}(\boldsymbol{\theta}) + X_{i,j}^{t-1}\{1 - \alpha_{i,j}^{t-1}(\boldsymbol{\theta}) - \beta_{i,j}^{t-1}(\boldsymbol{\theta})\}\,, \end{aligned}$$

where $f_{i,j}$'s and $g_{i,j}$'s are known functions. Let $\boldsymbol{\theta}_0 = (\theta_{0,1}, \ldots, \theta_{0,q})^\top \in \boldsymbol{\Theta} \subset \mathbb{R}^q$ be a q-dimensional unknown true parameter vector. Then $\alpha_{i,j}^{t-1} = \alpha_{i,j}^{t-1}(\boldsymbol{\theta}_0)$, $\beta_{i,j}^{t-1} = \beta_{i,j}^{t-1}(\boldsymbol{\theta}_0)$ and $\gamma_{i,j}^{t-1} = \gamma_{i,j}^{t-1}(\boldsymbol{\theta}_0)$.

We use maximum likelihood estimation (MLE) to estimate the parameters of the autoregressive networks with dependent edges. For any $l \in [q]$, let

$$\mathcal{S}_l = \big\{(i,j) : 1 \le i < j \le p \text{ and } \gamma_{i,j}^{t-1}(\boldsymbol{\theta}) \text{ involves } \theta_l \text{ for any } t \in [n] \setminus [m]\big\}$$

and $\mathcal{G} = \big\{l \in [q] : \gamma_{i,j}^{t-1}(\boldsymbol{\theta}) \text{ involves } \theta_l \text{ for all } 1 \le i < j \le p \text{ and } t \in [n] \setminus [m]\big\}$. Define the following partial log-likelihood,

$$\hat{\ell}_{n,p}^{(l)}(\boldsymbol{\theta}) = \frac{1}{(n-m)|S_l|} \sum_{t=m+1}^{n} \sum_{(i,j) \in S_l} \log\Big[\big\{\gamma_{i,j}^{t-1}(\theta)\big\}^{X_{i,j}^t} \big\{1 - \gamma_{i,j}^{t-1}(\theta)\big\}^{1 - X_{i,j}^t}\Big]\,.$$

Letting $(\hat{\theta}_{*,1}^{(l)}, \ldots, \hat{\theta}_{*,q}^{(l)})^\top = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \hat{\ell}_{n,p}^{(l)}(\boldsymbol{\theta})$ for each $l \in [q]$, we define the initial estimator $\widetilde{\boldsymbol{\theta}} = (\widetilde{\boldsymbol{\theta}}_{\mathcal{G}}^\top, \widetilde{\boldsymbol{\theta}}_{\mathcal{G}^c}^\top)^\top$ for $\boldsymbol{\theta}_0$ as

$$\widetilde{\boldsymbol{\theta}}_{\mathcal{G}} = \Big(\hat{\theta}_{*,l}^{(l')}\Big)_{l \in \mathcal{G}} \quad \text{ and } \quad \widetilde{\boldsymbol{\theta}}_{\mathcal{G}^c} = \Big(\hat{\theta}_{*,l}^{(l)}\Big)_{l \in \mathcal{G}^c}$$

for some $l' \in \mathcal{G}$.

However, the above initial estimator $\widetilde{\boldsymbol{\theta}}$ may suffer from slow convergence rates due to the high dimensionality of $\boldsymbol{\theta}$. To overcome this, we improve the estimation for each component $\theta_{0,l}$ by projecting the score function onto certain direction. An improved estimator for $\theta_{0,l}$ is then obtained by solving the projected score function while letting $\boldsymbol{\theta}_{-l} = \widetilde{\boldsymbol{\theta}}_{-l}$. The projection mitigates the impact of $\widetilde{\boldsymbol{\theta}}_{-l}$ in the improved estimation for $\theta_{0,l}$. This strategy was initially proposed by [2, 1] for constructing the valid confidence regions of some low-dimensional subvector of the whole parameters in high-dimensional models with removing the impact of the high-dimensional nuisance parameter. We use the transitivity model as an example to illustrate the effectiveness of both the initial estimation and the improved estimation in simulations and real data analysis.

References

[1] Chang, J., Shi, Z. and Zhang, J. (2023). Culling the herd of moments with penalized empirical likelihood, *Journal of Business & Economic Statistics*, **41**, 791–805.

[2] Chang, J., Chen, S. X., Tang, C. Y. and Wu, T. (2021). High-dimensional empirical likelihood inference. *Biometrika*, **109**, 127–147.

[3] Hanneke, S., Fu, W. and Xing, E. (2010). Discrete temporal models of social networks. *Electronic Journal of Statistics*, **4**, 585–605.

[4] Jiang, B., Leng, C., Yan, T., Yao, Q. and Yu, X. (2023a). A two-way heterogeneity model for dynamic networks. *arXiv:2305.12643*.

[5] Jiang, B., Li, J. and Yao, Q. (2023b). Autoregressive networks, *Journal of Machine Learning Research*, **24**, 1–69.

[6] Leifeld, P., Cranmer, S. and Desmarais, B. (2018). Temporal exponential random graph models with btergm: Estimation and bootstrap confidence intervals. *Journal of Statistical Software*, **83**, 1–36.

[7] Süveges, M. and Olhede, S. C. (2023). Networks with correlated edge processes. *Journal of the Royal Statistical Society, Series A*, **186**, 441–462.

JOINT LABORATORY OF DATA SCIENCE AND BUSINESS INTELLIGENCE, SOUTHWESTERN UNIVERSITY OF FINANCE AND ECONOMICS, CHENGDU, SICHUAN 611130, CHINA

Email address: `changjinyuan@swufe.edu.cn`

LOW-DIMENSIONAL ADAPTATION OF DIFFUSION MODELS

YUXIN CHEN

Keywords: diffusion models, low-dimensional structure, acceleration

Motivated by the practical efficacy of diffusion models, the past few years have witnessed a flurry of activity towards establishing convergence theory for diffusion generative models, particularly the two mainstream algorithms: DDPM and DDIM. For a fairly general family of target distributions $\mathbb{P}_{\mathrm{data}}$ (without assuming smoothness and log-concavity), the state-of-the-art theory demonstrated that for both DDPM and DDIM, it takes at most the order of (modulo some log factor)

$$\frac{d}{\varepsilon} \text{ iterations} \tag{0.1}$$

to yield a sample whose distribution is ε-close in total variation (TV) distance to the target distribution, provided that perfect score function estimates are available.

Nevertheless, even linear scaling in the ambient dimension d can still be prohibitively expensive for many contemporary applications. Take the ImageNet dataset for instance: each image might contain 150,528 pixels, while its intrinsic dimension is estimated to be 43 or less. As a result, applying the state-of-the-art theory (0.1) could suggest an iteration complexity that exceeds one million, even though practical implementations of DDIM and DDPM often produce high-quality samples in just a few hundred (or even a few ten) iterations. The discrepancy between theory and practice suggests that worst-case bounds, such as (0.1), may be overly conservative. To reconcile this discrepancy, it is crucial to bear in mind the intrinsic dimension of the target data distribution and explore whether and how diffusion models can harness this potentially low-dimensional structure.

Motivated by this, in this talk we would like to explore how diffusion models leverage low-dimensional structure to speed up the sampling process. Focusing on two mainstream samplers — the denoising diffusion implicit model (DDIM) and the denoising diffusion probabilistic model (DDPM) — and assuming accurate score estimates, we prove that their iteration complexities are no greater than the order of k/ε (up to some log factor), where ε is the precision in total variation distance and k is some intrinsic dimension of the target distribution. Our results are applicable to a broad family of target distributions without requiring smoothness or log-concavity assumptions. Further, we develop a lower bound that suggests the (near) necessity of the coefficients introduced by Ho et al. 2020 and Song et al. 2020 in facilitating low-dimensional adaptation. Our findings provide the first rigorous evidence for the adaptivity of the DDIM-type samplers to unknown low-dimensional structure, and improve over the state-of-the-art DDPM theory regarding total variation convergence.

DEPARTMENT OF STATISTICS AND DATA SCIENCE, THE WHARTON SCHOOL, UNIVERSITY OF PENNSYLVANIA, USA

Email address: yuxinc@wharton.upenn.edu

AGNOSTIC CHARACTERIZATION OF INTERFERENCE IN RANDOMIZED EXPERIMENTS (EXTENDED ABSTRACT)

DAVID CHOI

Classification AMS 2020: 62D99

Keywords: causal inference, spillovers, social networks, interference

In randomized experiments, it may be possible for the participants to affect each other, by mechanisms such as transmission of disease, sharing of information, peer influence, or economic competition. Such phenomena (termed "interference between units") violates assumptions that are commonly used for statistical inference.

Mechanisms for interference often play fundamental roles in our understanding of social outcomes. For this reason, the empirical characterization of interference (such as its nature, prevalence, or strength) may be of scientific interest. For experiments with binary-valued outcomes, we give a general approach for characterizing the prevalence of interference, which can be used to explore questions such as

Q1. How many units are affected by any treatment (including their own)?
Q2. How many units are affected by the treatment of others? of distant others?
Q3. How many units are affected by the treatment of others, provided that their own treatment satisfies some condition?

For each of these questions, our approach gives conservative point estimates and one-sided confidence intervals, which both lower bound the true value. Under reasonable experiment designs, the point estimate will be consistent for a lower bound on the true value, while the one-sided interval will cover the true value at the desired level. These consistency and coverage properties hold without any additional assumptions or restrictions on the nature of the interference, requiring only a randomized experiment whose design is known. As a result, our estimates remain valid even if they use an observed social network that is only a crude proxy for the actual social mechanisms.

A previous attempt to answer such questions relied on inversion of a test statistic, and produced quite conservative (though valid) lower bounds. Our new approach is significantly tighter, and may be more practical as a result. Our point estimates are asymptotically equal to Hajek-normalized contrasts, such as comparisons of treated versus untreated, or comparisons of different levels of indirect exposure, or comparisons that combine measures of direct and indirect treatment. Under stronger assumptions, such contrasts arise naturally as estimates of treatment effects. Our results indicate that without assumptions on interference, these contrasts may be interpreted more weakly as lower bounds on the number of units who are affected by the treatments. We also find empirically that our interval estimates have efficiency (i.e., interval widths) which is competitive with, and often better than, that of the expected average treatment effect (EATE), an assumption-lean treatment effect.

0.1. **Idea of Method.** Consider an experiment on N units, with $X = (X_1, \ldots, X_N)$ denoting the binary-valued treatment of each unit, and $Y = (Y_1, \ldots, Y_N)$ denoting their binary outcomes. We allow for arbitrary interference, so that the outcome Y_i of unit i may potentially depend on all N treatments,

$$Y_i = f_i(X_1, \ldots, X_N), \qquad i \in [N] \tag{0.1}$$

where the potential outcome mappings $\{f_i\}_{i=1}^N$ may be arbitrary and unknown.

Suppose that we wish to estimate τ^{basic}, the number of units who are affected by any treatment, including their own treatment or the treatment of others. To define this estimand, let $\mathcal{I} \subset [N]$ denote the subset of units who are unaffected by treatment and have constant outcome mappings,

$$\mathcal{I} = \{i : f_i(X) \text{ is constant in } X\}$$

so that $\tau^{\text{basic}} = N - |\mathcal{I}|$.

Our high-level approach to estimating τ^{basic} is the following:

(1) Propose idealized estimators $\hat{\tau}_1$ and $\hat{\tau}_2$ which will have good statistical properties, such as consistency and asymptotic normality, but require knowledge of $\mathcal{I}$

(2) Show that Δ, the difference in average outcomes between treated and control (which can be computed without knowledge of $\mathcal{I}$) converges to a lower bound for $\max(\hat{\tau}_1, \hat{\tau}_2)$, so that if τ_1 and τ_2 are both consistent for τ^{basic}, then Δ is an asymptotic lower bound.

(3) Lower bound the boundary of the lower 1-sided confidence intervals induced by $\hat{\tau}_1$ and $\hat{\tau}_2$ and their variance estimates, by minimizing the tighter of the two boundaries over all hypotheses for the unknown subset $\mathcal{I}$.

To define $\hat{\tau}_1$ and $\hat{\tau}_2$, let S_i denote the indicator of whether unit i's treatment and outcome have the same binary value,

$$S_i = 1\{(X_i, Y_i) = (1, 1) \text{ or } (0, 0)\}, \tag{0.2}$$

and let $\hat{\tau}_1$ and $\hat{\tau}_2$ denote sampling-based estimators of $\tau^{\text{basic}} = N - |\mathcal{I}|$, in which the unknown cardinality of $\mathcal{I}$ is unbiasedly estimated by a probability-weighted (i.e., Horvitz-Thompson) sample:

$$\hat{\tau}_1 = N - \sum_{i \in \mathcal{I}} \frac{1\{S_i = 1\}}{P(S_i = 1)} \qquad \text{and} \qquad \hat{\tau}_2 = N - \sum_{i \in \mathcal{I}} \frac{1\{S_i = 0\}}{P(S_i = 0)} \tag{0.3}$$

Because $\hat{\tau}_1$ and $\hat{\tau}_2$ involve only units in $\mathcal{I}$ whose outcomes are unaffected by treatment and hence are constant, they often will exhibit simple statistical behavior, even if strong interference exists between units who are not in $\mathcal{I}$. For example, if treatment is assigned by independent Bernoulli randomization, then $\hat{\tau}_1$ and $\hat{\tau}_2$ are sums of independent variables. Similarly, if the dependencies between the unit treatments are bounded, then $\hat{\tau}_1$ and $\hat{\tau}_2$ are sums of variables whose dependencies will be similarly bounded. For this reason, under a variety of designs we may expect the values of $\hat{\tau}_1$ and $\hat{\tau}_2$, while unknown due to $\mathcal{I}$ being unknown, to concentrate at their expectation (which equals τ^{basic}) and to be asymptotically normal.

Our motivation for constructing $\hat{\tau}_1$ and $\hat{\tau}_2$ is the following: under mild conditions on the experiment design, the maximum of $\hat{\tau}_1$ and $\hat{\tau}_2$ is lower bounded by the magnitude of the propensity-weighted difference in outcomes between treated and control, given by

$$\Delta = \sum_{i=1}^{N} \left(\frac{X_i}{P(X_i = 1)} - \frac{1 - X_i}{P(X_i = 0)} \right) Y_i,$$

as stated by Theorem 0.1 below:

Theorem 0.1. *Let the total weights of the treated and control converge to their expectations, so that*

$$\sum_{i=1}^{N} \frac{X_i}{P(X_i = 1)} = N + O_P(N^{1/2}) \quad \text{and} \quad \sum_{i=1}^{N} \frac{1 - X_i}{P(X_i = 0)} = N + O_P(N^{1/2}) \tag{0.4}$$

Then it holds that

$$\left| \sum_{i=1}^{N} \left(\frac{X_i}{P(X_i = 1)} - \frac{1 - X_i}{P(X_i = 0)} \right) Y_i \right| \leq \max(\hat{\tau}_1, \hat{\tau}_2) + O_P(N^{1/2}) \tag{0.5}$$

If $\hat{\tau}_1^{\text{Haj}}$ and $\hat{\tau}_2^{\text{Haj}}$ are asymptotically normal, with consistent variance estimators denoted by $\hat{V}_1$ and $\hat{V}_2$, then by combining 1-sided confidence intervals it holds with probability converging to at least $1 - \alpha$ that

$$\tau^{\text{basic}} \geq \max \left\{ \hat{\tau}_1^{\text{Haj}} - z_{1-\frac{\alpha}{2}} \sqrt{\hat{V}_1},\ \hat{\tau}_2^{\text{Haj}} - z_{1-\frac{\alpha}{2}} \sqrt{\hat{V}_2} \right\}. \tag{0.6}$$

As the right hand side of (0.6) requires knowledge of $\mathcal{I}$, it cannot be computed.

To construct a computable one-sided confidence interval for τ^{basic}, we will lower bound (0.6) by minimizing over all possible hypotheses for the unknown $\mathcal{I}$. Doing so results in the confidence statement that with probability at least $1 - \alpha$,

$$\tau^{\text{basic}} \geq \max \left(\min_{\phi \in \{0,1\}^N} \hat{\tau}_1^{\text{Haj}}(\phi) - z_{1-\frac{\alpha}{2}} \sqrt{\hat{V}_1(\phi)},\ \min_{\phi \in \{0,1\}^N} \hat{\tau}_2^{\text{Haj}}(\phi) - z_{1-\frac{\alpha}{2}} \sqrt{\hat{V}_2(\phi)} \right), \tag{0.7}$$

where $\hat{\tau}_k^{\text{Haj}}(\phi)$ and $\hat{V}_k(\phi)$ denote $\hat{\tau}_k^{\text{Haj}}$ and $\hat{V}_k$ evaluated under the hypothesis that $\mathcal{I} = \{i : \phi_i = 1\}$ for $\phi \in \{0, 1\}^N$. (See paper for further details, such as the form of the variance estimators $\hat{V}_1$ and $\hat{V}_2$.)

0.2. **Illustrative Example.** In an experiment described in [Cai et al., 2015], rural farmers in China were randomly assigned to information sessions where they would be given the opportunity to purchase weather insurance. The sessions were randomized to give either high or low levels of information about the insurance product. First round sessions were held three days before second round sessions, so that first round attendees would have opportunity for informal conversations with their second round friends, in which they might share their opinions about the insurance product. Social network information was elicited, with the farmers instructed to list 5 close friends with whom they specifically discussed rice production or financial issues.

The goal of the experiment was to broadly demonstrate the importance of information sharing, by measuring its effects in a randomized setting. One of the conclusions of [Cai et al., 2015] was that the decision to purchase insurance was affected not only by a farmer's own treatment assignment, but also by that of their friends; specifically, farmers assigned to a second round low-information session were more likely to purchase insurance if more of their listed friends in the first round were assigned to a high-information session.

For this experiment, our point estimate is that at least 23% of second round farmers, if assigned to a low information session, would be affected by information given to the first round farmers (1-sided 95% CI: at least 9%). This point estimate of 23% is asymptotically equal to a Hajek-normalized comparison of second round units who received low information directly but had many first round friends with high information, versus those in the second round who received low information directly and had few or no first round friends with high information :

$$\text{point estimate} \approx \sum_{i=1}^{N} \left(\frac{1}{\hat{N}_1} \frac{1\{X_i = 0, W_i = 1\}}{P(X_i = 0, W_i = 1)} - \frac{1}{\hat{N}_0} \frac{1\{X_i = 0, W_i = 0\}}{P(X_i = 0, W_i = 0)} \right) Y_i$$

Here $i \in [N]$ enumerates the second round units, $X_i = 0$ if unit i was assigned to a low information session, $W_i = 1$ if all of unit i's first round friends received high information, Y_i denotes unit i's decision of whether or not to purchase insurance, and $\hat{N}_1$ and $\hat{N}_0$ denote the Hajek normalization factors, where $\hat{N}_a = \sum_{i=1}^{N} (P(X_i = 0, W_i = a))^{-1}$ for $a = 0, 1$.

For comparison, we consider an EATE-type treatment effect that considers the relative effects of receiving (X_i, W_i) equal to $(0, 1)$ versus $(0, 0)$:

$$\text{treatment effect} = \frac{1}{N} \sum_{i=1}^{N} \left(\mathbb{E}[Y_i | X_i = 0, W_i = 1] - \mathbb{E}[Y_i | X_i = 0, W_i = 0] \right),$$

where the expectation is taken over the randomization of treatment under the experiment design. For this target parameter, the method of [Gao and Ding, 2023] gives a Hajek-normalized point estimate of 23%, and 95% CI of $[2\%, 45\%]$. This confidence interval requires an assumption of "approximate neighborhood interference", in which the interference between farmers in different villages is asymptotically negligible. Such an assumption might be debatable, as farmers listed cross-village friendships. In contrast, no assumptions on interference are required for our estimand. Thus for the purposes of demonstrating the presence of social influence (as opposed to policy recommendation), our estimand may be an appropriate target parameter, and has tighter, less questionable CIs when compared to an analogous treatment effect.

References

[Cai et al., 2015] Cai, J., De Janvry, A., and Sadoulet, E. (2015). Social networks and the decision to insure. *American Economic Journal: Applied Economics*, 7(2):81–108.

[Gao and Ding, 2023] Gao, M. and Ding, P. (2023). Causal inference in network experiments: regression-based analysis and design-based properties. *arXiv preprint arXiv:2309.07476*.

Heinz College of Public Policy and Information Systems, Carnegie Mellon University
E-mail address: davidch@andrew.cmu.edu

HIGH-DIMENSIONAL NETWORK CAUSAL INFERENCE

YINGYING FAN

Classification AMS 2020: 62E20; 91D30

Keywords: Network interference, Nonparametric interference effects, Average direct treatment effect on the treated, Confidence intervals, Neighborhood size confidence set

We propose a new method of high-dimensional network causal inference (HNCI) that provides both valid confidence intervals for the average direct treatment effect on the treated (ADET) and valid confidence sets for the neighborhood size affecting the interference effect. Consider a sample of n units indexed by $i \in [n] := \{1, 2, \dots, n\}$, connected through an interference network G, where each unit is randomly assigned a binary treatment $Z_i \sim$ Bernoulli(p_i) for some $p_i \in (0,1)$. Let $\mathbf{z} = (z_1, z_2, \cdots, z_n)^T \in \{0,1\}^n$ denote the treatment assignments, which serves as a realization of the random vector $\mathbf{Z} = (Z_1, Z_2, \cdots, Z_n)^T$. For example, $\mathbf{z}$ could indicate that a tax incentive is offered to a specific subset of businesses in a region. In the network setting, the units are referred to as nodes in G, which are rarely independent of each other. Hence, the effect of a tax incentive on a specific company may depend on whether its collaborators or competitors also receive the tax incentive. For the n nodes connected through G, the potential outcome of the ith node is defined as $\widetilde{Y}_i(\mathbf{z}) = \widetilde{Y}_i(z_i, \mathbf{z}_{-i})$, where $\widetilde{Y}_i(\cdot) : \{0,1\}^n \to \mathbb{R}$, and z_i and $\mathbf{z}_{-i}$ are the treatment assignments for the ith node and the remaining nodes, respectively. In practice, we may observe node covariates $\{\mathbf{C}_i\}_{i\in[n]}$.

We exploit the following potential outcome model framework introduced in [1], where the potential outcome of the ith node is defined as

$$\widetilde{Y}_i(z_i, \mathbf{z}_{-i}) = z_i\tau_i + f\big(\gamma_0(G_i^{\mathbf{z}}(k_0))\big) + \epsilon_i. \tag{0.1}$$

Here, $\tau_i := \mathbb{E}\{\widetilde{Y}_i(1, \mathbf{0}_{-i}) - \widetilde{Y}_i(0, \mathbf{0}_{-i}) | \mathbf{C}_i\}$ is the average direct effect of the treatment on the ith node, $i \in [n]$, $\gamma_0(\cdot)$ is a known mapping satisfying the *nested matching* property that $\gamma_0(G_i^{\mathbf{z}}(k)) = \gamma_0(G_j^{\mathbf{z}}(k))$ implies $\gamma_0(G_i^{\mathbf{z}}(k')) = \gamma_0(G_j^{\mathbf{z}}(k'))$ for all $k' \in [k]$, $f(\cdot)$ is an unknown interference function, k_0 is the smallest neighborhood size that satisfies (0.1), and ϵ_i's are independent errors with $\mathbb{E}(\epsilon_i) = 0$, $\mathrm{Var}(\epsilon_i) = \sigma_0^2$.

We work under model (0.1) to estimate and infer the average direct treatment effect on the treated (ADET)

$$\tau := \frac{1}{\sum_{i=1}^n Z_i} \sum_{i=1}^n Z_i\tau_i, \tag{0.2}$$

which represents the average incremental response of treated units to their own treatments. We are also interested in estimating the neighborhood size k_0 with statistical uncertainty guarantee.

For untreated nodes $z_i = 0$, we have

$$\widetilde{Y}_i(0, \mathbf{z}_{-i}) = f\big(\gamma_0(G_i^{\mathbf{z}}(k_0))\big) + \epsilon_i. \tag{0.3}$$

Thanks to the nested matching property, for each pre-specified neighborhood size $k \geq k_0$, the set of untreated nodes can be partitioned into $d(k)$ disjoint subsets, denoted as $S_k = \{S_{k,1}, S_{k,2}, \ldots, S_{k,d(k)}\}$, where each subset $S_{k,j}$ contains nodes with the same interference function value $\gamma_0(G_i^{\mathbf{z}}(k))$ for all $i \in S_{k,j}$. Define the vector of true interference function values over the node partition S_k as

$$\boldsymbol{\beta}_k^0 = (\beta_{k,1}^0, \beta_{k,2}^0, \cdots, \beta_{k,d(k)}^0)^T. \tag{0.4}$$

Based on this property, the response vector $\mathbf{y}_{obs} \in \mathbb{R}^{n_0}$ of untreated nodes can be rewritten in the form of a linear regression model

$$\mathbf{y}_{obs} = \mathbf{X}_k \boldsymbol{\beta}_k^0 + \boldsymbol{\varepsilon}_0, \tag{0.5}$$

where $\mathbf{X}_k \in \{0,1\}^{n_0 \times d(k)}$ is the design matrix with each row indicating the corresponding unit's membership in S_k, and $\boldsymbol{\varepsilon}_0$ is the error term. Since k can be larger than k_0 and the function f can be many-to-one, there exisits unknown homogeneity in the regression coefficient vector $\boldsymbol{\beta}_k^0$, and the true interference function values $\{f(\gamma_0(G_i^{\mathbf{z}}(k))) : z_i = 0, i = 1, \cdots, n\}$ can be estimated by estimating the regression coefficients $\boldsymbol{\beta}_k^0$ in (0.5).

By considering this linear representation, we reformulate the original nonparametric model into a linear regression model where the regression coefficients, corresponding to the underlying true interference function values of nodes, exhibit a latent homogeneous structure. This formulation enables us to leverage existing literature on homogeneity pursuit [3] to conduct valid statistical inferences with theoretical guarantees for estimating the unknown $\boldsymbol{\beta}_k^0$. This gives us the estimates of the set of interference function values $\{f(\gamma_0(G_i^{\mathbf{z}}(k))) : z_i = 0, i = 1, \cdots, n\}$ and the confidence interval for these estimates.

By using the matching technique, the estimates of ADET can also be constructed and the confidence interval can be calculated. We theoretically justify the inference for the ADET through establishing asymptotic normality with estimable variances. By employing the repro samples approach [4], we further provide the confidence set for the interference of neighborhood size k_0 with theoretical guarantees. The practical utility of the newly suggested methods is demonstrated through simulations and real data examples.

References

[1] Alexandre Belloni, Fei Fang, and Alexander Volfovsky. Neighborhood adaptive estimators for causal inference under network interference. *arXiv preprint arXiv:2212.03683*, 2022.

[2] Wenqin Du, Rundong Ding, Yinging Fan, and Jinchi Lv. HNCI: High-Dimensional Network Causal Inference *arXiv preprint arXiv:2412.18568*, 2024.

[3] Xiaotong Shen and Hsin-Cheng Huang. Grouping pursuit through a regularization solution surface. *Journal of the American Statistical Association*, 105, 727–739, 2010.

[4] Peng Wang, Min-Ge Xie, and Linjun Zhang. Finite- and large-sample inference for model and coefficients in high- dimensional linear regression with repro samples. *arXiv preprint arXiv:2209.09299*, 2022.

University of Southern California
Email address: fanyingy@usc.edu

SEMIPARAMETRIC MODELING AND ANALYSIS FOR LONGITUDINAL NETWORK DATA

BY YINQIU HE[1,a], JIAJIN SUN[2,b], YUANG TIAN[3,c],
ZHILIANG YING[4,d], AND YANG FENG[5,e *]

[1]*Department of Statistics, University of Wisconsin-Madison,* [a]*yinqiu.he@wisc.edu*

[2]*Department of Statistics, Florida State University,* [b]*jsun5@fsu.edu*

[3]*Shanghai Center for Mathematical Sciences, Fudan University,* [c]*yatian20@fudan.edu.cn*

[4]*Department of Statistics, Columbia University,* [d]*zying@stat.columbia.edu*

[5]*Department of Biostatistics, School of Global Public Health, New York University,* [e]*yang.feng@nyu.edu*

We introduce a semiparametric latent space model for analyzing longitudinal network data. The model consists of a static latent space component and a time-varying node-specific baseline component. We develop a semiparametric efficient score equation for the latent space parameter by adjusting for the baseline nuisance component. Estimation is accomplished through a one-step update estimator and an appropriately penalized maximum likelihood estimator. We derive oracle error bounds for the two estimators and address identifiability concerns from a quotient manifold perspective. Our approach is demonstrated using the New York Citi Bike Dataset.

1. Semiparametric Poisson Latent Space Model. We consider longitudinal pairwise interaction counts of n subjects (nodes) over T discrete time points. Specifically, for a time point $t \in \{1, \dots, T\}$ and nodes $i, j \in \{1, \dots, n\}$, $A_{t,ij}$ denotes the number of i-j interactions at the time point t. We propose a Poisson-based latent space model

$$(1) \qquad A_{t,ij} = A_{t,ji} \sim \text{Poisson}\{\mathbb{E}(A_{t,ij} \mid z, \alpha)\}, \quad \text{independently with}$$
$$\mathbb{E}(A_{t,ij} \mid z, \alpha) = \exp(\alpha_{it} + \alpha_{jt} + \langle z_i, z_j \rangle),$$

which naturally adopts the exponential link function $\exp(\cdot)$ to model the event counts. For any two nodes i and j, their interaction effect is modeled through the inner product of two corresponding latent vectors $\langle z_i, z_j \rangle = z_i^\top z_j$, similarly to the inner product model of a single network (Ma, Ma and Yuan, 2020). The latent vectors z_i's do not change with respect to the time point t and represent the shared latent structures across T heterogeneous networks. For example, z_i's can encode the time-invariant geographic information in multiple transportation networks. At a given time point t, when α_{it} increases and all the other parameters are fixed, edges connecting the node i tend to have higher numbers of counts at the time point t, indicating higher baseline activity levels. Therefore, α_{it}'s model the degree heterogeneity across different nodes $i \in \{1, \dots, n\}$ and time points $t \in \{1, \dots, T\}$, and are called baseline degree heterogeneity parameters of nodes and time. In the hourly bike-sharing networks, α_{it}'s can represent distinct baseline activity levels across different stations and hours.

Model specification (1) may be expressed in vector-matrix notation as

$$(2) \qquad \mathbb{E}(\mathbf{A}_t \mid Z, \alpha) = \exp(\alpha_t 1_n^\top + 1_n \alpha_t^\top + ZZ^\top),$$

[*] He, Sun, and Tian contribute equally to this work. Corresponding Author: Yang Feng.

MSC2020 subject classifications: Primary 62H12, 05C82; secondary 91D30, 62F12.

Keywords and phrases: network, counting process Poisson model, latent space model, low rank, semiparametric, heterogeneity.

where $\alpha_t = (\alpha_{1t}, \ldots, \alpha_{nt})^\top \in \mathbb{R}^{n\times 1}$, $Z = (z_1, \ldots, z_n)^\top \in \mathbb{R}^{n\times k}$, $1_n = (1, \ldots, 1)^\top \in \mathbb{R}^{n\times 1}$, and $\exp(\cdot)$ is the elementwise exponential operation. Throughout this paper, we consider the asymptotic regime in which the number of nodes n and the number of time periods T increase to infinity while the dimension of the latent space k is fixed. Thus, $\alpha_t 1_n^\top + 1_n \alpha_t^\top + ZZ^\top$ is a low-rank matrix. To ensure identifiability, we assume that column means of Z are zero, i.e., $1_n^\top Z/n = 0$. This centering assumption is analogous to the classical two-way analysis of variance (ANOVA) modeling with interaction (Scheffe, 1999). Additionally, since $ZZ^\top = ZQQ^\top Z^\top$ for any $Q \in \mathcal{O}(k)$, where $\mathcal{O}(k) = \{Q \in \mathbb{R}^{k\times k} : QQ^\top = \mathrm{I}_k\}$, Z is identifiable up to a common orthogonal transformation of its rows.

2. Generalized Semiparametric One-Step Estimator. In this section, we introduce our generalized semiparametric one-step estimator of Z and provide theoretical guarantees. We first introduce some notation. Model (1) leads to the following form for the log-likelihood function

(3)

$$L(Z, \alpha) = L(Z_v, \alpha_v) = \sum_{t=1}^{T} \sum_{1\leqslant i\leqslant j\leqslant n} \{A_{t,ij}(\alpha_{it} + \alpha_{jt} + \langle z_i, z_j\rangle) - \exp(\alpha_{it} + \alpha_{jt} + \langle z_i, z_j\rangle)\}$$

where, for notational convenience in the differentiation of the likelihood, we use Z_v and α_v to denote vectorizations of Z and α, respectively, i.e., $Z_v = (z_1^\top, \ldots, z_n^\top)^\top \in \mathbb{R}^{nk\times 1}$ and $\alpha_v = (\alpha_1^\top, \ldots, \alpha_T^\top)^\top \in \mathbb{R}^{nT\times 1}$. Then we let $\dot{L}_Z(Z, \alpha)$ and $\dot{L}_\alpha(Z, \alpha)$ denote the partial derivatives of $L(Z, \alpha)$ with respect to vectors Z_v and α_v, respectively.

With the above preparations, we construct our generalized semiparametric one-step estimator as

$$\hat{Z}_v = \check{Z}_v + \left\{I_{eff}(\check{Z}, \check{\alpha})\right\}^+ S_{eff}(\check{Z}, \check{\alpha}), \tag{4}$$

where $(\check{Z}, \check{\alpha})$ denotes an initial estimate, and B^+ represents the Moore-Penrose inverse of a matrix B, which is uniquely defined and also named pseudo inverse (Ben-Israel and Greville, 2003).

2.1. *Theory.* Throughout the sequel, we use $(Z^\star, \alpha^\star)$ to denote the true value of (Z, α). In other words, our observed data follow the model (1) with $(Z, \alpha) = (Z^\star, \alpha^\star)$. Besides, we denote $\Theta_{t,ij} = \alpha_{it} + \alpha_{jt} + \langle z_i, z_j\rangle$. We define the estimation error from the i-th row of $\check{Z}$ as $\mathrm{dist}_i(\check{z}_i, z_i^\star) = \|\check{z}_i - \check{Q}^\top z_i^\star\|_2$, where

$$\check{Q} = \underset{Q\in\mathcal{O}(k)}{\arg\min} \|\check{Z} - Z^\star Q\|_{\mathrm{F}}, \tag{5}$$

so that $\mathrm{dist}^2(\check{Z}, Z^\star) = \sum_{i=1}^n \mathrm{dist}_i^2(\check{z}_i, z_i^\star)$. The goal in this subsection is to establish an error bound for the proposed one-step estimator (4) in terms of the distance defined above.

THEOREM 1. *Let $\hat{Z}$ be the generalized one-step estimator defined as in (4). Let $\varsigma = \max\{\epsilon, 1/2\}$. For any constant $s > 0$, there exists a constant $C_s > 0$ such that when $n/\log^{2\varsigma}(T)$ is sufficiently large,*

$$\Pr\left\{\mathrm{dist}^2(\hat{Z}, Z^\star) > \frac{1}{T} \times C_s r_{n,T}\right\} = O(n^{-s}),$$

where $r_{n,T} = \max\left\{1, \frac{T}{n}\right\} \log^{4\varsigma}(nT)$.

3. Discussion. In this work, we propose a longitudinal latent space model tailored for recurrent interaction events. We develop two novel semiparametric estimation techniques, i.e., the generalized semiparametric one-step updating and the penalized maximum likelihood estimation, and show that the resulting estimators attain the oracle estimation error rate for the shared latent structure. The first approach utilizes the semiparametric efficient score equation to construct a second-order updating estimator. We show that the estimator possesses a geometric interpretation on the quotient manifold, which automatically overcomes the non-uniqueness issue due to overparametrization. The second approach corresponds to a convex relaxation of the low-rank static latent space component.

By separating the (primary) parameters of interest associated with the static latent space from the dynamic nuisance parameters, a strategy commonly found in semiparametrically efficient parameter estimation, we are able to delineate the oracle rates of convergence for the primary and the nuisance parameters according to their dimensions. This strategy also helps us to untangle the static and time-heterogeneous components inherent in the network model and construct the oracle estimators.

There are a few other interesting future works. First, the ability to accurately estimate latent structures could enable important downstream analysis such as prediction, hypothesis testing, and change-point detection. For instance, it may be useful to ascertain a change point in the structure of the latent space (Bhattacharjee, Banerjee and Michailidis, 2020; Enikeeva and Klopp, 2021; Padilla, Yu and Priebe, 2022). The achievement of oracle estimation error rates could facilitate the quantification of uncertainty in estimators, which in turn lays a strong foundation for conducting reliable statistical inference.

Second, this paper focuses on the variance in estimation error rates as a function of n and T, while treating the latent dimension k and network sparsity level as fixed. Extending the current methodology and theory to the cases when k grows (Choi, Wolfe and Airoldi, 2012) as well as sparse networks (Qin and Rohe, 2013; Le, Levina and Vershynin, 2017) are important topics.

Third, this work aims to unveil the fundamental relationship between the estimation errors and the degree of baseline heterogeneity. We focus on the most challenging scenario where the degree of baseline heterogeneity increases linearly with respect to n and T. It is possible to impose additional structures to reduce baseline heterogeneity, such as assuming $\{\alpha_{i1}, \ldots, \alpha_{iT}\}$ to be piecewise constants. Nevertheless, as T increases, the intrinsic number of parameters would eventually become large to keep up with the increasing data complexity. The developed results would also provide us with techniques for investigating such scenarios. Which structural assumptions are appropriate may vary across different applications and require case-by-case analyses in future research.

Fourth, the proposed model has the potential for further extensions to capture more complex network structures. Currently, heterogeneity across networks is only characterized at the first-order baseline levels α_{it}'s, while the second-order interaction terms are modeled by the time-invariant components z_i's. We find that this model adequately describes the analyzed dataset. But more generally, it may also be of interest to incorporate time-varying interaction terms, which would further increase the model complexity and pose new theoretical challenges. Moreover, the proposed model adopts the Euclidean inner product to describe the interactions between nodes, which can be limited to capturing homophilic network structures. Recently, researchers proposed to use indefinite inner products to capture heterophilic structures (Rubin-Delanchy et al., 2022; Lei, 2021; MacDonald, Levina and Zhu, 2022).

Finally, it would be worthwhile to generalize our results to other models, including various distributions for weighted edges, continuous time stamps, or additional covariates influencing the network structure (Hoff, Raftery and Handcock, 2002; Vu et al., 2011; Perry and Wolfe, 2013; Hoff, 2015; Kim et al., 2018; Sit, Ying and Yu, 2021; Weng and Feng, 2021; Huang, Sun and Feng, 2023). We believe that the proposed semiparametric analysis framework can function as a valuable building block for establishing sharp estimation error rates under those models.

REFERENCES

Ben-Israel, A. and Greville, T. N. (2003). *Generalized Inverses: Theory and Applications* **15**. Springer Science & Business Media.

Bhattacharjee, M., Banerjee, M. and Michailidis, G. (2020). Change point estimation in a dynamic stochastic block model. *Journal of Machine Learning Research* **21** 4330–4388.

Choi, D. S., Wolfe, P. J. and Airoldi, E. M. (2012). Stochastic blockmodels with a growing number of classes. *Biometrika* **99** 273–284.

Enikeeva, F. and Klopp, O. (2021). Change-point detection in dynamic networks with missing links. *arXiv preprint arXiv:2106.14470*.

Hoff, P. D. (2015). Multilinear tensor regression for longitudinal relational data. *The Annals of Applied Statistics* **9** 1169–1193.

Hoff, P. D., Raftery, A. E. and Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association* **97** 1090–1098.

Huang, S., Sun, J. and Feng, Y. (2023). PCABM: Pairwise covariates-adjusted block model for community detection. *Journal of the American Statistical Association*.

Kim, B., Niu, X., Hunter, D. R. and Cao, X. (2018). A dynamic additive and multiplicative effects model with application to the United Nations voting behaviors. *arXiv preprint arXiv:1803.06711*.

Le, C. M., Levina, E. and Vershynin, R. (2017). Concentration and regularization of random graphs. *Random Structures & Algorithms* **51** 538–561.

Lei, J. (2021). Network representation using graph root distributions. *The Annals of Statistics* **49** 745 – 768.

Ma, Z., Ma, Z. and Yuan, H. (2020). Universal latent space model fitting for large networks with edge covariates. *Journal of Machine Learning Research* **21** 1–67.

MacDonald, P. W., Levina, E. and Zhu, J. (2022). Latent space models for multiplex networks with shared structure. *Biometrika* **109** 683–706.

Padilla, O. H. M., Yu, Y. and Priebe, C. E. (2022). Change point localization in dependent dynamic nonparametric random dot product graphs. *Journal of Machine Learning Research* **23** 10661–10719.

Perry, P. O. and Wolfe, P. J. (2013). Point process modelling for directed interaction networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75** 821–849.

Qin, T. and Rohe, K. (2013). Regularized spectral clustering under the degree-corrected stochastic blockmodel. *Advances in neural information processing systems* **26**.

Rubin-Delanchy, P., Cape, J., Tang, M. and Priebe, C. E. (2022). A statistical interpretation of spectral embedding: The generalised random dot product graph. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **84** 1446-1473.

Scheffe, H. (1999). *The Analysis of Variance* **72**. John Wiley & Sons.

Sit, T., Ying, Z. and Yu, Y. (2021). Event history analysis of dynamic communication networks. *Biometrika* **108** 223âĂŞ230.

Vu, D., Hunter, D., Smyth, P. and Asuncion, A. (2011). Continuous-time regression models for longitudinal networks. In *Advances in Neural Information Processing Systems* 2492–2500.

Weng, H. and Feng, Y. (2021). Community detection with nodal information: likelihood and its variational approximation. *Stat* e428.

LEARNING UNDER LATENT GROUP SPARSITY VIA DIFFUSION ON NETWORKS

SUBHRO GHOSH

Classification AMS 2020:

Keywords: Latent group sparsity |Networks |Laplacian geometry |Diffusion |Gaussian Free Field |Stochastic Block Model

Group or cluster structure on explanatory variables in machine learning problems is a very general phenomenon, which has attracted broad interest from practitioners and theoreticians alike. In this work, joint with Soumendu Mukherjee (Indian Statistical Institute), we contribute an approach to sparse learning under such group structure, that does not require prior information on the group identities. Our paradigm is motivated by the Laplacian geometry of an underlying network with a related community structure, and proceeds by directly incorporating this into a penalty that is effectively computed via a heat-flow-based local network dynamics. The proposed penalty interpolates between the lasso and the group lasso penalties, the runtime of the heat-flow dynamics being the interpolating parameter. As such it can automatically default to lasso when the group structure reflected in the Laplacian is weak. In fact, we demonstrate a data-driven procedure to construct such a network based on the available data. Notably, we dispense with computationally intensive pre-processing involving clustering of variables, spectral or otherwise. Our technique is underpinned by rigorous theorems that guarantee its effective performance and provide bounds on its sample complexity. In particular, in a wide range of settings, it provably suffices to run the diffusion for time that is only logarithmic in the problem dimensions. We explore in detail the interfaces of our approach with key statistical physics models in network science, such as the Gaussian Free Field and the Stochastic Block Model. We validate our approach by successful applications to real-world data from a wide array of application domains, including computer science, genetics, climatology and economics. Our work raises the possibility of applying similar diffusion-based techniques to classical learning tasks, exploiting the interplay between geometric, dynamical and stochastic structures underlying the data.

References

[1] Ghosh, S. and Mukherjee, S.S. Learning under Latent Group Sparsity via Diffusion on Networks. arXiv preprint arXiv:2507.15097

Dept of Math, National University of Singapore
Email address: subhrowork@gmail.com

LEARNING LATENT FEATURES FROM NETWORK DATA

CHRISTOPHE GIRAUD

Classification AMS 2020: 62F07

Keywords: seriation, latent model, recursive trees, Jordan ordering

1. Latent variable model

We describe a graph by its adjacency matrix $X \in \{0,1\}^{n\times n}$.

Definition 1.1. Latent variable model. *The graph is generated randomly as follows:*

- *each vertex i is characterized by a latent variable $z_i \in \mathcal{Z}$*
- *conditionally on z, the X_{ij} are independent, with*

$$\mathbb{P}\left[X_{ij}=1|z\right]=\mathbb{E}\left[X_{ij}=1|z\right]=f(z_i,z_j)$$

where $f:\mathcal{Z}\times\mathcal{Z}\to[0,1]$.

This model encompass the graphon model, random geometric graphs where $f(z_i,z_j)=g(d(z_i,z_j))$ with d a distance on $\mathcal{Z}$, the Robinson model where f decreases when moving away from the diagonal, the stochastic block model, ranking models and so on.

An ideal objective is to recover the latent values $z_1,\dots,z_n$ from the observation of X. Yet, it is an ill-posed problem, due to the lack of identifiability: while it is a minor issue in parametric models (estimation up to some “invariant” transformation), it is a much more severe issue in non-parametric models. For simplicity, we focus on the case where

- f belongs to some non-parametric class, with smoothness or shape assumptions;
- the latent positions are $z_i=\pi^*(i)$ for π^* a permutation of $\{1,\dots,n\}$.

The goal is to recover π^* from $X=\pi^*F\pi^{*T}+E$, with $F_{ij}=f(i,j)$ unknown, π^* an unknown permutation matrix and E_{ij} independent sub-Gaussian random variables. What is the rate of estimation without computational constraints? What is the rate of estimation with poly-time algorithms? Is there a gap between the two?

Statistical-computational gaps exist in latent variable model [2], for example both in the geometric seriation model $F_{ij}=\lambda\,\mathbf{1}_{|i-j|\le\sqrt{n}}$, with $\lambda>0$, and in the Hölder graphon model $f:[0,1]\times[0,1]\to[0,1]$ with α-Hölder regularity, $0<\alpha<1$. We outline below two settings where estimation can be performed in poly-time at the optimal statistical rate.

2. Bi-Lipschitz shape constraint

Definition 2.1. Bi-Lipschitz $\mathcal{BL}(\alpha,\beta)$. *Assume that $F\in[0,1]^{n\times n}$ is symmetric and*

- *$F_{ik}-F_{jk}\ge\alpha\frac{|i-j|}{n}$ for $k<i<j$, and $F_{jk}-F_{ik}\ge\alpha\frac{|i-j|}{n}$ for $i<j<k$;*
- *$|F_{ik}-F_{jk}|\le\beta\frac{|i-j|}{n}$.*

The parameter α drives the signal strength. The parameter β is a smoothness parameter.

Example 2.2. Toeplitz matrix. *The simple Toeplitz matrix $F_{ij} = 1 - \alpha\frac{|i-j|}{n}$ belongs to $\mathcal{BL}(\alpha, \alpha)$.*

We define the max-error loss as

$$\ell_\infty(\hat{\pi}, \pi^*) = \min_{\pi^* \, admissible} \frac{1}{n} \max_{i\in[n]} |\hat{\pi}_i - \pi_i^*|$$

Theorem 2.3. [3] *There exist some poly-time estimators $\hat{\pi}^{poly}$ such that, for any $F \in \mathcal{BL}(\alpha, \beta)$, any $n \geq C_{\alpha,\beta}$, and for some numerical constant $c > 0$, with probability at least $1 - n^{-2}$*

$$\ell_\infty(\hat{\pi}^{poly}, \pi^*) \leq \frac{c}{\alpha}\sqrt{\frac{\log n}{n}} \, .$$

Furthermore, this rate is optimal for the Toeplitz matrix, up to a possible log factor.

This result proved in [3] ensures that the optimal statistical rate $n^{-1/2}$ for estimating Bi-Lipstchitz permuted matrices can be achieved by poly-time algorithms. Hence, there is no statistical-computational gap in this case. The optimal algorithm essentially

(1) first estimates the neighborhood distance

$$D_{ij}^* = \sqrt{n \sum_k (F_{ik}^{\pi^*} - F_{jk}^{\pi^*})^2}$$

at optimal rate $|\hat{D} - D^*|_\infty = O(n^{3/4})$;

(2) then perform a first partial ordering of points separated by $O(n^{3/4})$ based on this estimation;

(3) then refine this partial ordering by comparing partial sums, providing a reliable partial ordering of points separated by $O(n^{1/2})$.

Furthermore, Theorem 2.3 remains valid for weak local Bi-Lipschitz functions.

3. Recursive trees

Going beyond conditional i.i.d. graphs, we can consider recursive trees such as

Definition 3.1. Random Recursive Tree (RRT) *Build a tree recursively by connecting each new node to existing nodes uniformly at random*

Definition 3.2. Preferential Attachement Tree (PA) *Build a tree recursively by connecting each new node to existing nodes with a probability proportional to their degree*

Let us define a Jordan centroid $\hat{\sigma}_J$ as a vertex such that, when removing it, it splits the tree into components of size smaller than $\lfloor n/2 \rfloor$. At least one, maximum two such vertices exist. For a node u, we can define $\hat{d}(u)$ as the number of descendants of u in the tree rooted at $\hat{\sigma}_J$. Then, we can order the nodes according to the number of descendant $\hat{d}(u)$, ties being broken randomly. We call Jordan ordering this ordering $\hat{\sigma}_J(u)$, which can be computed in $O(n \log n)$ time.

Theorem 3.3. [1] *For any $\alpha \geq 1$, in the RTT model*

$$R_\alpha(\hat{\sigma}_J) := \sum_{u\in V} \frac{|\hat{\sigma}_J(u) - \sigma(u)|}{\sigma(u)^\alpha} \leq \kappa(\alpha)\, n^{2-\alpha} + C \log^4 n$$

with

$$\kappa(\alpha) = \frac{2}{2-\alpha} + \frac{2e^2}{(2-\alpha)^2} + \frac{2}{(2-\alpha)^3}.$$

Furthermore, the rate $n^{2-\alpha}$ *is optimal for* $\alpha \in [1,2)$.

Most of the story is the same for the PA model, except that Descendant ordering is optimal up to constant only for $1 \leq \alpha < 5/4$.

References

[1] Simon Briend, Gabor Lugosi, Christophe Giraud, and Deborah Sulem. Estimating the history of a random recursive tree. *Bernoulli*. To appear

[2] Bertrand Even, Christophe Giraud, and Nicolas Verzelen. Computational lower bounds in latent models: clustering, sparse-clustering, biclustering *arXiv*, `https://arxiv.org/abs/2506.13647`

[3] Yann Issartel, Christophe Giraud, Nicolas Verzelen. Minimax optimal seriation in polynomial time *arXiv*, `https://arxiv.org/abs/2405.08747`

Institut de Mathématiques d'Orsay, Université Paris-Saclay
E-mail address: `christophe.giraud@universite-paris-saclay.fr`

IDENTIFICATION AND INFERENCE FROM CROSS-SECTIONAL DATA VIA HIGHER ORDER CUMULANTS

NIELS RICHARD HANSEN

Classification AMS 2020: 62R01, 62A09

Keywords: cumulants, identification, Lyapunov equation, operator selfdecomposable distributions, steady-state distributions

1. Introduction

We are interested in identification of and inference about properties of dynamical systems from cross-sectional data, that is, from data obtained at a single timepoint from a multivariate stationary process. In this report we present recent results for a particular class of distributions that appear as steady-state distributions of a Markov process solving a linear stochastic differential equation (SDE) driven by a Lévy process.

Definition 1.1. *Let M be a $p \times p$ stable matrix and let $(Z_t)_{t \geq 0}$ denote a p-dimensional Lévy process with $\mathbf{E}(\log(1 + \|Z_1\|)) < \infty$. The distribution of X is M-selfdecomposable if*

$$X = \int_0^\infty e^{sM} \mathrm{d}Z_s. \tag{1.1}$$

Recall that a matrix M is stable if all eigenvalues of M have negative real part, which guarantees that the integral in (1.1) is well-defined. Distributions defined by (1.1) for some stable matrix M and Lévy process Z are called operator-selfdecomposable (OSD) distributions. Our interest in these distributions follows from Theorem 4.1 in [5], which implies that an M-selfdecomposable distribution is the unique steady-state distribution of the stationary Markov process solving the linear SDE

$$\mathrm{d}X_t = MX_t\mathrm{d}t + \mathrm{d}Z_t. \tag{1.2}$$

We are particularly interested in identification and estimation of the matrix M from cross-sectional data. Our main result is Theorem 3.1, which shows that for OSD distributions, M is generically identified from second and third order cumulants – provided that Z_1 has a non-degenerate third order cumulant (and thus is non-Gaussian). For results on identification of and inference about M from second order cumulants only (the covariance matrix), see [1] and [7].

One of the main reasons for our interest in identification and estimation of M is that the SDE (1.2) entails a collection of interventional distributions, and thus has a causal interpretation. This will not be discussed futher in this report, but see [2], [3] and [6].

Date: 13-August-2025.

2. Cumulants

The k-th order cumulant tensor of $X \in \mathbb{R}^p$ is

$$\mathrm{cum}_k(X)_{i_1,\dots,i_k} = \mathrm{cum}(X_{i_1}, \dots, X_{i_k}).$$

With $\times_r$ denoting the r-mode product, the proposition below, shown in [3], characterizes the cumulant tensors of an OSD distribution.

Proposition 2.1. *If the Lévy process Z_t has finite k-th moment, the k-th order cumulant tensor $K = \mathrm{cum}_k(X)$ of the OSD distribution given by* (1.1) *solves*

$$K \times_1 M + \ldots + K \times_k M + \mathcal{C}_k = 0 \tag{2.1}$$

where $\mathcal{C}_k = \mathrm{cum}_k(Z_1)$ is the k-th order cumulant tensor of Z_1.

The equation (2.1) shows for $k = 1$ that

$$\mathbf{E}(X) = -M^{-1}\mathbf{E}(Z_1),$$

and for $k = 2$ we find the equation

$$M\Sigma + \Sigma M^T + \mathcal{C}_2 = 0 \tag{2.2}$$

for the covariance matrix $\Sigma = \mathrm{Var}(X)$. The equation (2.2) is the well-known Lyapunov equation, and (2.1) can therefore be viewed as a generalization of the Lyapunov equation to higher order cumulants.

3. Identification

In the following, the cumulants $\mathcal{C}_2$ and $\mathcal{C}_3$ of Z_1 are assumed to be diagonal (which is a weak form of independence between the coordinates of Z_1), and we identify $\mathcal{C}_2$ with its diagonal as an element in $\mathbb{R}^p_+$ and $\mathcal{C}_3$ with its diagonal as an element in $(\mathbb{R}\backslash\{0\})^p$. Note that when $\mathcal{C}_3 \in (\mathbb{R}\backslash\{0\})^p$ all diagonal entries are non-zero, which is a necessary non-degeneracy condition for our identification result. It is a strong form of a non-Gaussianity condition.

We will parametrize M in terms of a digraph $G = ([p], E)$ with p nodes and edgeset E, which will restrict the non-zero entries of M. In terms of G we define

$$\mathbb{R}^E_{\text{stab}} = \{M \in \mathbb{R}^{p\times p} \mid M_{ij} = 0 \text{ if } (j,i) \notin E,\ M \text{ is stable}\},$$

and we let

$$\Theta_G = \mathbb{R}^E_{\text{stab}} \times \mathbb{R}^p_+ \times (\mathbb{R}\backslash\{0\})^p$$

be the parameter set of $\theta = (M, \mathcal{C}_2, \mathcal{C}_3)$. Furthermore, we let

$$\varphi_G : \Theta_G \to \mathrm{PD}_p \times \mathrm{Sym}^3(\mathbb{R}^p)$$

denote the parametrization of the second and third order cumulant tensors, $(\Sigma, K) = \varphi_G(M, \mathcal{C}_2, \mathcal{C}_3)$, in terms of M and the diagonal cumulants $\mathcal{C}_2$ and $\mathcal{C}_3$ of Z_1. The map φ_G is defined via the solutions of (2.1) for $k = 2$ and $k = 3$.

With these definitions we can state our main identification result, shown in [4].

Theorem 3.1. *Suppose G is connected and has all self-loops then there exists a proper algebraic set $\mathcal{N}_G$ such that for $\theta \in \Theta_G\backslash\mathcal{N}_G$,*

$$\varphi_G^{-1}(\varphi_G(\theta)) = \{c\theta \mid c > 0\}.$$

The interpretation of Theorem 3.1 is that the parameter $\theta = (M, \mathcal{C}_2, \mathcal{C}_3)$ is generically identifiable, up to a global scaling factor $c > 0$, from the second and third order cumulant tensors of the steady-state distribution. Here ”generic” means ”outside the proper algebraic set $\mathcal{N}_G$”.

Based on empirical estimates $\hat{\Sigma}$ and $\hat{K}$ of the second and third order cumulant tensors, we can estimate M (we well as the nuisance parameters $\mathcal{C}_2$ and $\mathcal{C}_3$) via standard estimating equations. For $\theta = (M, \mathcal{C}_2, \mathcal{C}_3) \in \Theta_G \backslash \mathcal{N}_G$ we obtain a consistent and asymptotically normal estimator under appropriate moment conditions on the Lévy process Z_t. We refer to [4] for details on the estimation procedure.

4. Discussion

Since $\varphi_G(\theta) = \varphi_G(c\theta)$ for all $c > 0$, which follows directly from (2.1), it will at best be possible to identify M up to a global scaling factor. Such a scaling factor determines how quickly the Markov process solving (1.2) converges to its steady-state distribution, and it makes sense that we cannot identify the scaling factor from cross-sectional data only. The conclusion in Theorem 3.1 is therefore as good as it gets.

If G is not connected, the identification result holds for each connectivity component of G separately – with independent scaling factors for each component – and it holds that for all $\theta \in \Theta_G$,

$$\dim(\varphi_G^{-1}(\varphi_G(\theta))) \geq \sharp \text{ connec. comp. in } G.$$

We conjecture that

$$\mathcal{N}_G = \{(M, \mathcal{C}_2, \mathcal{C}_3) \in \Theta_G \mid \text{graph}(M) \text{ not connec.}\},$$

but we have not been able to prove this or find a counterexample. In our proof in [4], $\mathcal{N}_G$ is defined more implicitly and it could potentially be larger than conjectured.

References

[1] P. Dettling, R. Homs, C. Améndola, M. Drton, N. R. Hansen. Identifiability in Continuous Lyapunov Models. *SIAM J. Matrix Anal. Appl.*, 44(4), 1799–1821, 2023

[2] S. Lauritzen. T. Richardson. Chain graph models and their causal interpretations. *J. R. Statist. Soc. B*, 64(3), 321–361, 2002.

[3] A. Markham, C. O. Recke, J. Adams, N. R. Hansen. Linear non-Gaussian steady-state models. *Working paper*, 2025+.

[4] C. O. Recke, N. R. Hansen. Identification and estimation in continuous Lyapunov models. *Working paper*, 2025+.

[5] K. Sato, M. Yamazato. Operator-selfdecomposable distributions as limit distributions of processes of Ornstein-Uhlenbeck type. *Stochastic Processes and their Applications*, 17(1), 73-100, 1984.

[6] A. Sokol, N. R. Hansen. Causal interpretation of stochastic differential equations. *Electron. J. Probab.*, 19(100), 1–24, 2014.

[7] G. Varando, N. R. Hansen. Graphical continuous Lyapunov models. *Proceedings of UAI, PMLR*, 124, 989–998, 2020

University of Copenhagen, Department of Mathematical Sciences, Universitetsparken 5, DK-2100 Copenhagen, Denmark

Email address: niels.hansen@math.ku.dk

LIMIT LAWS FOR GROMOV-WASSERSTEIN ALIGNMENT WITH APPLICATIONS TO TESTING GRAPH ISOMORPHISMS

KENGO KATO

The Gromov-Wasserstein (GW) distance enables comparing metric measure spaces based solely on their internal structure, making it invariant to isomorphic transformations. This property is particularly useful for comparing datasets that naturally admit isomorphic representations, such as unlabelled graphs or objects embedded in space. However, apart from the recently derived empirical convergence rates for the quadratic GW problem, a statistical theory for valid estimation and inference remains largely obscure. Pushing the frontier of statistical GW further, this work derives the first limit laws for the empirical GW distance across several settings of interest: (i) discrete, (ii) semi-discrete, and (iii) general distributions under moment constraints under the entropically regularized GW distance. The derivations rely on a novel stability analysis of the GW functional in the marginal distributions. The limit laws then follow by an adaptation of the functional delta method. As asymptotic normality fails to hold in most cases, we establish the consistency of an efficient estimation procedure for the limiting law in the discrete case, bypassing the need for computationally intensive resampling methods. We apply these findings to testing whether collections of unlabelled graphs are generated from distributions that are isomorphic to each other.

CORNELL UNIVERSITY, USA
Email address: kk976@cornell.edu

TRANS-GLASSO: A TRANSFER LEARNING APPROACH TO PRECISION MATRIX ESTIMATION

MLADEN KOLAR

Classification AMS 2020: 62H22, 62J07, 68T05

Keywords: transfer learning, precision matrix estimation, minimax rates, high-dimensional inference

Many real-world systems—ranging from gene regulatory interactions in biology to financial asset dependencies—can be represented by networks, whose edges correspond to conditional relationships among variables. These relationships are succinctly captured by the precision matrix of a multivariate distribution. Estimating the precision matrix is thus fundamental to uncovering the underlying network structure. However, this task can be challenging when the available data for the target domain are limited, undermining accurate inference.

In this talk, I will present Trans-Glasso, a novel two-step transfer learning framework for precision matrix estimation that leverages data from source studies to improve estimates in the target study. First, Trans-Glasso identifies shared and unique features across studies via a multi-task learning objective. Then, it refines these initial estimates through differential network estimation to account for structural differences between the target and source precision matrices. Assuming that most entries of the target precision matrix are shared with at least one source matrix, we derive non-asymptotic error bounds and show that Trans-Glasso achieves minimax optimality under certain conditions.

Through extensive simulations, Trans-Glasso demonstrates improved performance over standard methods, especially in small-sample settings. Applications to gene regulatory networks across multiple brain tissues and protein networks in various cancer subtypes confirm its practical effectiveness in biological contexts, where understanding network structures can provide insights into disease mechanisms and potential interventions. Beyond biology, these techniques are broadly applicable wherever precision matrix estimation and network inference play a crucial role, including neuroscience, finance, and social science.

This is a joint work with Boxin Zhao and Cong Ma [1].

References

[1] Boxin Zhao, Cong Ma, and Mladen Kolar. Trans-Glasso: A Transfer Learning Approach to Precision Matrix Estimation. *arXiv preprint arXiv:2411.15624*, 2024.

Department of Data Sciences and Operations, University of Southern California Marshall School of Business, and Department of Statistics and Data Science, Mohamed bin Zayed University of Artificial Intelligence

Email address: mkolar@usc.edu; mladen.kolar@mbzuai.ac.ae

STATISTICAL ANALYSIS OF RECIPROCITY

CHENLEI LENG

Classification AMS 2020: 62F10, 62F12, 62C20, 62R99

Keywords: Asymptotic normality; Effective sample size; Maximum likelihood estimator; Reciprocity; Sparse networks.

Consider a directed network with n nodes, denoted by $G_n = (V, E)$, where $V = \{1, \ldots, n\}$ is the set of nodes and $E \subseteq V \times V$ represents the edge set. We focus on simple graphs, so no self-loops are allowed, i.e., $(j, j) \notin E$ for any $j \in V$. Let $A_{ij} \in \{0, 1\}$ denote the random variable indicating the presence of a directed link from node i to node j. Assuming that dyads (A_{ij}, A_{ji}) and (A_{kl}, A_{lk}) are independent whenever $\{i, j\} \cap \{k, l\} = \emptyset$, the Bernoulli model with reciprocity (BR) specifies multinomial probabilities for each dyad as follows (Krivitsky and Kolaczyk, 2015):

(0.1) **BR model:** $p_{ij}(0,0) \propto 1,\ p_{ij}(1,0) = p_{ij}(0,1) \propto \exp(\mu_n),\ p_{ij}(1,1) \propto \exp(2\mu_n + \rho_n),$

where $p_{ij}(a, b) = p(A_{ij} = a, A_{ji} = b)$. In this model, μ_n represents the baseline tendency of nodes i and j to connect, while ρ_n captures *reciprocity*, the propensity for pairs of nodes to form mutual links. BR model serves as a natural extension of the Erdős–Rényi model (Erdős and Rényi, 1959, 1960) for undirected graphs, adapted to incorporate reciprocity for the analysis of directed networks. This model raises a fundamental question:

Question 1: What is the effective sample size for the statistical inference of μ_n and ρ_n?

This question would be straightforward if μ_n and ρ_n were fixed, as it would fall under standard maximum likelihood estimation. However, when μ_n and ρ_n depend on n–the regime where the network is sparse–the inference of these parameters has been only partially explored in Krivitsky and Kolaczyk (2015). That work assumes that the effective sample sizes for μ_n and ρ_n are of the same order. Extending the analysis to allow different sparsity levels for μ_n and ρ_n provides a more comprehensive solution to Question 1, offering deeper insights into the effective sample sizes required for a broader range of network structures. Related, Chen et al. (2021) examines the effective sample size in the context of the Erdős–Rényi model under arbitrary sparsity, focusing on a single density parameter similar in spirit to μ_n. The examination of the interplay between the two parameters, μ_n and ρ_n , under differing sparsity regimes represents a new and more nuanced perspective, offering insights beyond those provided by models with a single density parameter.

More importantly, a complete answer to this question will pave the way for developing new models. As an example, we extend the BR model to the following:

$$\begin{aligned} \mathbf{p_{1.5}\ model}: \quad & p_{ij}(0,0) \propto 1, \quad p_{ij}(1,0) \propto \exp\big(\mu_n + X_i^T\gamma_1 + Y_j^T\gamma_2\big), \\ & p_{ij}(0,1) \propto \exp\big(\mu_n + X_j^T\gamma_1 + Y_i^T\gamma_2\big), \\ (0.2) \qquad & p_{ij}(1,1) \propto \exp\big(2\mu_n + \big(X_i^T + X_j^T\big)\gamma_1 + \big(Y_i^T + Y_j^T\big)\gamma_2 + \rho_n + V_{ij}^T\delta\big), \end{aligned}$$

with additional parameters γ_1, γ_2, and δ, where $X_i \in \mathbb{R}^{d_1}$ represents covariates related to node i's outgoingness, $Y_i \in \mathbb{R}^{d_2}$ relates to its incomingness, and $V_{ij} \in \mathbb{R}^{d_3}$ governs the reciprocity between nodes i and j. The model in (0.2) allows for node-specific heterogeneity via $X_i^T\gamma_1$ for outgoingness and $Y_j^T\gamma_2$ for incomingness, and $V_{ij}^T\delta$ to model heterogeneity in reciprocal relationships. Assuming that the parameters associated with the covariates are fixed, we further pose the following question:

Question 2: What are the effective sample sizes for the statistical inference of γ_1, γ_2, and δ?

The model in (0.2) has a close relationship with the p_1 model introduced by Holland and Leinhardt (1981), where the p_1 model employs node-specific fixed effects without explicitly accounting for link-specific reciprocity. Our model in (0.2) parametrizes these fixed effects through covariates, achieving a more parsimonious structure. Although it may lack some of the flexibility of the p_1 model, this approach offers certain advantages, such as enabling link prediction for new nodes not used in model fitting. Additionally, a key advantage of the model in (0.2) lies in its suitability for sparser networks. We show that inference is feasible as long as the number of links diverges. In contrast, the p_1 model, with its large number of parameters, typically requires much denser networks to ensure the existence and asymptotic normality of its estimators, though no formal inference procedures are currently available for these estimators (see literature review below). Additionally, the model in (0.2) shares features with the p_2 model (Van Duijn et al., 2004), which also includes random effects for outgoingness and incomingness. As our model conceptually bridges the p_1 and p_2 models, we refer to it as the $p_{1.5}$ model.

1. The BR Model

We begin by examining the effective sample sizes for the BR model as specified in (0.1). For the sake of theoretical analysis and notational convenience, it is beneficial to work with the parameters (μ_n, τ_n), where $\tau_n = 2\mu_n + \rho_n$. The negative log-likelihood function with respect to (μ_n, τ_n) can be expressed as:

$$\ell_n^{(1)}(\mu_n, \tau_n) = \sum_{i<j} \log(k_{n,ij}) - \mu_n \sum_{i<j} (A_{ij}(1 - A_{ji}) + A_{ji}(1 - A_{ij})) - \tau_n \sum_{i<j} A_{ij}A_{ji},$$

where $k_{n,ij} = 1 + 2\exp(\mu_n) + \exp(\tau_n)$ serves as the normalizing constant. It is important to note that the likelihood functions defined in terms of (μ_n, ρ_n) and (μ_n, τ_n) are equivalent, as are their corresponding maximum likelihood estimators. This leads us to the following lemma:

Lemma 1.1. *Suppose $(\hat{\mu}_n, \hat{\tau}_n) = \text{argmin}_{(\mu_n,\tau_n)\in\mathbb{R}^2}\, \ell_n^{(1)}(\mu_n, \tau_n)$. Then, it follows that $(\hat{\mu}_n, \hat{\tau}_n - 2\hat{\mu}_n) = \text{argmin}_{(\mu_n,\rho_n)\in\mathbb{R}^2}\, \ell_n^{(2)}(\mu_n, \rho_n)$, where $\ell_n^{(2)}(\mu_n, \rho_n)$ denotes the negative log-likelihood function parametrized by μ_n and ρ_n. The reverse direction also holds.*

Given this equivalence, we focus on estimating μ_n and τ_n. Inspired by the role of $-\log n$ in the Erdős–Rényi model for sparse networks, we define

$$\mu_n = -a\log n + \mu, \quad \tau_n = -b\log n + \tau,$$

where $\mu \in [-M_\mu, M_\mu]$, $\tau \in [-M_\tau, M_\tau]$, and $a, b > 0$. The constant a preceding $\log n$ directly reflects network sparsity, though similar asymptotic normality results may arise from other scaling factors beyond $\log n$. From $\ell_n^{(1)}(\mu_n, \tau_n)$, we interpret a as the sparsity index for non-reciprocal links and b for reciprocal links.

This transformation clarifies the dependence of sparsity on n while allowing for intuitive statistical inference on the fixed parameters μ and τ. For further discussions on this topic, we refer to Krivitsky and Kolaczyk (2015) and Chen et al. (2021). It is important to note that the constants a, μ, b, and τ are not identifiable or estimable. To address these challenges, we will later develop a straightforward inference procedure for μ_n and τ_n.

Under the given scaling, we find that the expected number of non-reciprocal links is $E\big(\sum_{i,j=1}^n A_{ij} - \sum_{i<j} A_{ij}A_{ji}\big) \asymp n^{2-a}$, while the expected number of reciprocated links is $E\big(\sum_{i<j} A_{ij}A_{ji}\big) \asymp n^{2-b}$. Consequently, the total expected number of links is of order n^{2-a} if $a \leq b$, or n^{2-b} if $a > b$. This scaling choice highlights that the two quantities can indeed differ in magnitude. Notably, Krivitsky and Kolaczyk (2015) examined a special case of our framework when $a = b = 1$, leading to comparable expected numbers of non-reciprocal and reciprocated links. For sparse networks, the sufficient statistics $(\sum_{i<j} A_{ij} + A_{ji}, \sum_{i<j} A_{ij}A_{ji})$ in the BR model can be efficiently computed using a sparse adjacency matrix. As a result, the time complexity for computing the maximum likelihood estimator is $O(n^{2-\min\{a,b\}})$, which is lower than $O(n^2)$ when $\min\{a,b\} > 0$.

We now derive the effective sample sizes for μ and τ, assuming that a and b are known. We begin by expressing the negative log-likelihood function as follows:

$$\ell_n(\mu,\tau) = \sum_{i<j} \log(k_{ij}) - \mu \sum_{i<j} \left(A_{ij}(1-A_{ji}) + A_{ji}(1-A_{ij})\right) - \tau \sum_{i<j} A_{ij}A_{ji}, \tag{1.1}$$

where $k_{ij} = 1 + 2n^{-a}\exp(\mu) + n^{-b}\exp(\tau)$ serves as the normalizing constant. Our maximum likelihood estimator is defined as

$$(\hat{\mu},\hat{\tau}) = \operatorname{argmin}_{(\mu,\tau)\in\Omega_1} \frac{1}{\binom{n}{2}} \ell_n(\mu,\tau),$$

with $\Omega_1 = [-M_\mu, M_\mu] \times [-M_\tau, M_\tau]$. To derive the asymptotic results, we make the following assumptions:

Assumption 1.2. *(Sparse network) Assume $0 < a, b < 2$. The true values (μ_0,τ_0) lie within the interior of Ω_1.*

The conditions $a > 0$ and $b > 0$ ensure that the resulting graph is sparse, while $a < 2$ and $b < 2$ are necessary to guarantee that the total numbers of reciprocal and non-reciprocal links approach infinity. Without these conditions, consistent estimation would not be achievable. We now present the following result regarding the maximum likelihood estimator (MLE). All our results hold under Assumption 1.2, meaning they apply to arbitrarily sparse networks.

Proposition 1.3. *(Asymptotic normality of the MLE in BR model) Under Assumption 1.2, as n approaches infinity, the MLE $(\hat{\mu},\hat{\tau})$ is consistent and asymptotically normal, specifically:*

$$\left(\sqrt{n^{2-a}}(\hat{\mu}-\mu_0), \quad \sqrt{n^{2-b}}(\hat{\tau}-\tau_0)\right)^T \rightsquigarrow N(0,\Sigma^{-1}),$$

where

$$\Sigma = \begin{pmatrix} \exp(\mu_0) & 0 \\ 0 & \exp(\tau_0)/2 \end{pmatrix}.$$

Following the reasoning in Krivitsky and Kolaczyk (2015) and Chen et al. (2021), we can interpret n^{2-a} and n^{-b} as the effective sample sizes for μ and τ, respectively. This interpretation is intuitive, as from equation (1.1), μ can be seen as the density parameter for the configuration $(1, 0)$ and $(0, 1)$, while τ represents the density parameter for the configuration $(1, 1)$.

References

Chen, M., Kato, K., and Leng, C. (2021). Analysis of networks via the sparse β-model. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(5):887–910.

Erdős, P. and Rényi, A. (1959). On random graph. *Publicationes Mathematicate*, 6:290–297.

Erdős, P. and Rényi, A. (1960). On the evolution of random graphs. *Publ. math. inst. hung. acad. sci*, 5(1):17–60.

Holland, P. W. and Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. *Journal of the American Statistical association*, 76(373):33–50.

Krivitsky, P. N. and Kolaczyk, E. D. (2015). On the question of effective sample size in network modeling: an asymptotic inquiry. *Statistical Science*, 30(2):184–198.

Van Duijn, M. A., Snijders, T. A., and Zijlstra, B. J. (2004). p2: a random effects model with covariates for directed graphs. *Statistica Neerlandica*, 58(2):234–254.

Department of Statistics, University of Warwick
Email address: c.leng@warwick.ac.uk

TENSOR DATA ANALYSIS AND SOME APPLICATIONS IN NEUROSCIENCE

LEXIN LI

Classification AMS 2020:

Keywords:

Multidimensional arrays, or tensors, are becoming increasingly prevalent in a wide range of scientific applications. In this talk, I will present two case studies from neuroscience, where tensor decomposition proves particularly useful.

The first study is a cross-area neuronal spike trains analysis, which we formulate as the problem of regressing a multivariate point process on another multivariate point process. We develop a new point process regression, and model the predictor effects through the conditional intensities using a set of basis transferring functions in a convolutional fashion. We organize the corresponding transferring coefficients in the form of a three-way tensor, then impose the low-rank, sparsity, and subgroup structures on this coefficient tensor. These structures help reduce the dimensionality, integrate information across different individual processes, and facilitate the interpretation. We develop a highly scalable optimization algorithm for parameter estimation. We derive the large sample error bound for the recovered coefficient tensor, and establish the subgroup identification consistency, while allowing the dimension of the multivariate point process to diverge. We demonstrate the efficacy of our method through both simulations and a cross-area neuronal spike trains analysis in a sensory cortex study.

The second study is a multimodal neuroimaging analysis for Alzheimer's disease, which we formulate as the problem of modeling the correlations of two sets of variables conditioning on the third set of variables. We propose a generalized liquid association analysis method, which offers a new and unique angle to the problem of studying three-way associations. We extend the notion of liquid association of Li (2002) from the univariate setting to the sparse, multivariate, and high-dimensional setting. We establish a population dimension reduction model, transform the problem to sparse Tucker decomposition of a three-way tensor, and develop a higher-order orthogonal iteration algorithm for parameter estimation. We derive the non-asymptotic error bound and asymptotic consistency of the proposed estimator, while allowing the variable dimensions to be larger than and diverge with the sample size. We demonstrate the efficacy of the method through both simulations and a multimodal neuroimaging application for Alzheimer's disease research.

References

[1] Tang, X. and Li, L. Multivariate temporal point process regression. *Journal of the American Statistical Association*, 118, 830-845, 2023.

[2] Li, L., Zeng, J., and Zhang, X. Generalized liquid association analysis for multimodal neuroimaging. *Journal of the American Statistical Association*, 118, 1984-1996, 2023.

Department of Biostatistics and Epidemiology & Helen Wills Neuroscience Institute. University of California, Berkeley. 2121 Berkeley Way, #5302, Berkeley, CA 94720-7360

Email address: lexinli@berkeley.edu

A REGRESSION TREE APPROACH TO MISSING DATA

WEI-YIN LOH

Classification AMS 2020: 68T09

Keywords: imputation, incomplete data, prediction, recursive partitioning

The standard method to fitting a prediction model to incomplete data that have missing values in the predictor variables is to first complete the data by imputing (i.e., estimating) the missing values. This approach may not be logical if the "missing" values are non-existent instead of missing due to non-response. One example is the variable "age of spouse" for people who are single. Another common example occurs in so-called "skip questions", where variable $x_1 = 1$ if a person has a credit card and $x_1 = 0$ otherwise, and x_2 is the credit card balance. Here, x_2 would be reported as missing for people who do not have credit cards.

This talk introduces a new approach to missing values that makes missing-value imputation unnecessary. It accomplishes this by means of the GUIDE regression tree algorithm [4, 5], which fits a binary decision tree model to the incomplete data. A major strength of GUIDE is that it treats missing values as observed qualitative information and sends them to the left or right subnode at each split according to the values of the outcome (y) variable relative to those with non-missing values. In particular, it allows for splits that send missing values and only missing values to one subnode [6]. Other regression tree algorithms either impute the missing values before splitting the node [2], or send observations with missing values randomly to the left or right subnode [3]. The method is demonstrated on a dataset to predict death or intubation in patients hospitalized for Covid-19 [1].

REFERENCES

[1] Baker, T. B., Loh, W.-Y., Piasecki, T. M., Bolt, D. M., Smith, S. S., Slutske, W. S., Conner, K. L., Bernstein, S. L., and Fiore, M. C. (2023). A machine learning analysis of correlates of mortality among patients hospitalized with COVID-19. *Scientific Reports*, 13(4080).

[2] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Chapman & Hall/CRC.

[3] Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased recursive partitioning: a conditional inference framework. *Journal of Computational and Graphical Statistics*, 15:651–674.

[4] Loh, W.-Y. (2002). Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica*, 12:361–386.

[5] Loh, W.-Y. (2009). Improving the precision of classification trees. *Annals of Applied Statistics*, 3:1710–1737.

[6] Loh, W.-Y., Eltinge, J., Cho, M. J., and Li, Y. (2019). Classification and regression trees and forests for incomplete data from sample surveys. *Statistica Sinica*, 29:431–453.

DEPARTMENT OF STATISTICS, UNIVERSITY OF WISCONSIN, MADISON, USA
Email address: loh@stat.wisc.edu

ASYMPTOTIC THEORY OF EIGENVECTORS FOR LATENT EMBEDDINGS WITH GENERALIZED LAPLACIAN MATRICES

JINCHI LV

Classification AMS 2020: 60F05, 62E20, 94A15

Keywords: Graph and manifold embeddings; Asymptotic distributions; Eigenvectors and eigenvalues; Local laws; RMT under dependency; High dimensionality

Laplacian matrices are commonly employed in many real applications, encoding the underlying latent structural information such as graphs and manifolds. The use of the normalization terms naturally gives rise to random matrices with dependency. It is well-known that dependency is a major bottleneck of new random matrix theory (RMT) developments. To this end, in this paper, we formally introduce a class of generalized (and regularized) Laplacian matrices, which contains the Laplacian matrix and the random adjacency matrix as a specific case, and suggest the new framework of the asymptotic theory of eigenvectors for latent embeddings with generalized Laplacian matrices (ATE-GL). Our new theory is empowered by the tool of generalized quadratic vector equation for dealing with RMT under dependency, and delicate high-order asymptotic expansions of the empirical spiked eigenvectors and eigenvalues based on local laws. The asymptotic normalities established for both spiked eigenvectors and eigenvalues will enable us to conduct precise inference and uncertainty quantification for applications involving the generalized Laplacian matrices with flexibility. We discuss some applications of the suggested ATE-GL framework and showcase its validity through some numerical examples. *This is a joint work with Jianqing Fan, Yingying Fan, Fan Yang and Diwen Yu*.

Graphs and manifolds are commonly associated with sequence data such as texts. To enable text modeling and token generation, one may first construct Word2Vec embeddings of individual words and then build a graph of short sequences, where each short sequence can be viewed as a node of the graph and also be viewed as a point in a latent low-dimensional manifold. The link strengths between each pair of nodes can be calculated using a certain similarity measure of the embedding vectors, giving rise to a high-dimensional random matrix representing the graph data. For network applications, an important question is how to uncover the latent structural information underlying the graphs via low-dimensional manifold representations, often much lower than the ambient embedding dimensionality of each node. The Laplacian matrices for network data have been widely used to construct latent embeddings of graphs, where the nodes of the graph are represented in a latent subspace spanned by the corresponding leading eigenvectors of the Laplacian matrix. A natural question is how to characterize the asymptotic distributions of the leading eigenvectors and eigenvalues of the Laplacian matrix. The existing results in random matrix theory (RMT) have focused almost always on the setting of independent entries modulo symmetry, which is a major bottleneck of new RMT developments. Due to the use of the normalization terms, the Laplacian matrix is an example of a random matrix with dependency. To enable more flexible latent embeddings of graphs, we will extend the concept of the Laplacian matrix to that of the generalized (regularized) Laplacian matrix with index $\alpha \in [0, \infty)$. A key question we aim to address in this paper is how to characterize the asymptotic distributions of the leading eigenvectors and eigenvalues of the generalized (regularized) Laplacian matrices, a *new* class of high-dimensional random matrices with *dependency* representing the network data.

The primary objective of this paper is to investigate the asymptotic behaviors of the empirical spiked eigenvalues and eigenvectors of the generalized (regularized) Laplacian matrix (with some commonly used regularization terms) for the signal-plus-noise model when the signals are above a certain threshold. In particular, we will derive both the law of large numbers (LLN) and central limit theorems (CLTs) for the spiked sample eigenvalues and eigenvector components. Our results extend significantly the previous works [Fan, Fan, Han, and LvFan et al.2022a, Fan, Fan, Lv, and YangFan et al.2024] to the context of the generalized Laplacian matrix framework. These prior studies established the LLN and CLTs for spiked sample eigenvalues and eigenvector components of the adjacency matrices of large networks, which can be viewed as a special case of our results when $\alpha = 0$. Our results also compensate for the results of a recent work [Ke and WangKe and Wang2024], where entrywise large-deviation bounds for the eigenvectors associated with the largest eigenvalues of the Laplacian matrix for the DCMM model were established through the leave-one-out strategy. Additionally, in [Tang and PriebeTang and Priebe2018], the CLTs for the components of eigenvectors pertaining to the adjacency matrix and the Laplacian matrix of a random dot product graph were established, under the assumption of a prior distribution on the mean adjacency matrix.

Our results can be of independent theoretical interest due to the important role played by Laplacian matrices in the spectral graph theory. On the other hand, they can also serve as crucial ingredients for statistical inference concerning large networks and more general models. For example, they may enhance the characterization of the community membership probability matrix $\mathbf{\Pi}$ through spectral clustering methods for community detection, a widely used and scalable tool in the literature, as demonstrated in [Von LuxburgVon Luxburg2007, AbbeAbbe2017, JinJin2015, Le, Levina, and VershyninLe et al.2016, Lei and RinaldoLei and Rinaldo2015, Rohe, Chatterjee, and YuRohe et al.2011], or may enable hypothesis testing with network data, a prevalent technique utilized in various contexts such as [Arias-Castro and VerzelenArias-Castro and Verzelen2014, Verzelen and Arias-CastroVerzelen and Arias-Castro2015, Bickel and SarkarBickel and Sarkar2016, LeiLei2016, Wang and BickelWang and Bickel2017, Fan, Fan, Han, and LvFan et al.2022b, Fan, Fan, Lv, and YangFan et al.2024]. Due to the length constraint, we leave the investigation of various important applications of our theoretical results obtained in this paper to future work.

References

[AbbeAbbe2017] Abbe, E. (2017). Community detection and stochastic block models: recent developments. *Journal of Machine Learning Research 18*(1), 6446–6531.

[Arias-Castro and VerzelenArias-Castro and Verzelen2014] Arias-Castro, E. and N. Verzelen (2014). Community detection in dense random networks. *The Annals of Statistics 42*(3), 940–969.

[Bickel and SarkarBickel and Sarkar2016] Bickel, P. J. and P. Sarkar (2016). Hypothesis testing for automated community detection in networks. *Journal of the Royal Statistical Society Series B 78*(1), 253–273.

[Fan, Fan, Han, and LvFan et al.2022a] Fan, J., Y. Fan, X. Han, and J. Lv (2022a). Asymptotic theory of eigenvectors for random matrices with diverging spikes. *Journal of the American Statistical Association 117*, 996–1009.

[Fan, Fan, Han, and LvFan et al.2022b] Fan, J., Y. Fan, X. Han, and J. Lv (2022b). SIMPLE: statistical inference on membership profiles in large networks. *Journal of the Royal Statistical Society Series B 84*, 630–653.

[Fan, Fan, Lv, and YangFan et al.2024] Fan, J., Y. Fan, J. Lv, and F. Yang (2024). SIMPLE-RC: group network inference with non-sharp nulls and weak signals. *arXiv preprint arXiv:2211.00128*.

[JinJin2015] Jin, J. (2015). Fast community detection by SCORE. *The Annals of Statistics 43*(1), 57–89.

[Ke and WangKe and Wang2024] Ke, Z. T. and J. Wang (2024). Optimal network membership estimation under severe degree heterogeneity. *Journal of the American Statistical Association* (just-accepted), 1–28.

[Le, Levina, and VershyninLe et al.2016] Le, C. M., E. Levina, and R. Vershynin (2016). Optimization via low-rank approximation for community detection in networks. *The Annals of Statistics 44*(1), 373–400.

[LeiLei2016] Lei, J. (2016). A goodness-of-fit test for stochastic block models. *The Annals of Statistics 44*(1), 401–424.

[Lei and RinaldoLei and Rinaldo2015] Lei, J. and A. Rinaldo (2015). Consistency of spectral clustering in stochastic block models. *The Annals of Statistics 43*(1), 215–237.

[Rohe, Chatterjee, and YuRohe et al.2011] Rohe, K., S. Chatterjee, and B. Yu (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics 39*(4), 1878–1915.

[Tang and PriebeTang and Priebe2018] Tang, M. and C. E. Priebe (2018). Limit theorems for eigenvectors of the normalized Laplacian for random graphs. *The Annals of Statistics 46*(5), 2360–2415.

[Verzelen and Arias-CastroVerzelen and Arias-Castro2015] Verzelen, N. and E. Arias-Castro (2015). Community detection in sparse random networks. *The Annals of Applied Probability 25*(6), 3465–3510.

[Von LuxburgVon Luxburg2007] Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing 17*, 395–416.

[Wang and BickelWang and Bickel2017] Wang, Y. R. and P. J. Bickel (2017). Likelihood-based model selection for stochastic block models. *The Annals of Statistics 45*(2), 500–528.

Data Sciences and Operations Department, Marshall School of Business, University of Southern California, Los Angeles, CA 90089, USA

Email address: jinchilv@marshall.usc.edu

GRAPH MATRICES AND TENSOR NETWORKS

AARON POTECHIN

Classification AMS 2020: 15A60

Keywords: graph matrices, tensor networks, tensors

Graph matrices are a type of matrix which is a powerful tool for analyzing problems on random inputs. Graph matrices have been used extensively for sum of squares lower bounds on average case problems [5, 11, 10, 19, 15, 18, 16, 24, 17, 21] and have also recently been used to analyze power-sum decompositions of polynomials [3], to analyze the ellipsoid fitting conjecture [20], [13], and to analyze a class of first-order iterative algorithms including belief propagation and approximate message passing [14]. That said, we only have a partial understanding of graph matrices. We currently know the following about graph matrices:

(1) We have general norm bounds for graph matrices [2, 15, 22, 4, 17, 23, 24].
(2) The limiting distribution of the singular values as $n \to \infty$ has been determined for a family of graph matrices called multi-Z-shaped graph matrices [7, 8].
(3) A certain family of graph matrices behaves like Hermite polynomials of Gassian random variables [14].

When the random input is $G(n, \frac{1}{2})$, graph matrices are defined as follows:

Definition 0.1 (Fourier characters over $G(n, \frac{1}{2})$)**.** *Given a set of potential edges E, we define $\chi_E(G) = (-1)^{|E \setminus E(G)|} = \prod_{e \in E} \chi_{\{e\}}(G)$ where $\chi_{\{e\}}(G) = 1$ if $e \in E(G)$ and -1 if $e \notin E(G)$.*

Proposition 0.2. *$E_{G \sim G(n,\frac{1}{2})}[\chi_E(G)\chi_{E'}(G)] = 1$ if $E' = E$ and 0 if $E' \neq E$.*

Definition 0.3 (Shapes)**.** *A shape α consists of a graph with vertices $V(\alpha)$ and edges $E(\alpha)$ together with two distinguished tuples of vertices U_α and V_α which are subsets of $V(\alpha)$.*

Definition 0.4 (Graph matrices)**.** *Given a shape α, we define the graph matrix M_α to be the $\frac{n!}{(n-|U_\alpha|)!} \times \frac{n!}{(n-|V_\alpha|)!}$ matrix whose rows and columns are indexed by tuples of size $|U_\alpha|$ and $|V_\alpha|$ with entries*

$$M_\alpha(A, B) = \sum_{\pi: V(\alpha) \to V(G): \pi \text{ is injective}, \pi(U_\alpha) = A, \pi(V_\alpha) = B} \chi_{\pi(E(\alpha))(G)}$$

Definition 0.5. *A vertex separator of a shape α is a set of vertices $S \subseteq V(\alpha)$ such that every path from U_α to V_α (including paths of length 0) must contain a vertex in S.*

Theorem 0.6 (AMP20)**.** *For all shapes α which have no isolated vertices outside of U_α and V_α, with high probability, $||M_\alpha||$ is $\tilde{O}(n^{\frac{|V(\alpha)| - s_\alpha}{2}})$ where s_α is the minimum size of a vertex separator of α and the $\tilde{O}$ contains factors depending on the size of α and logarithmic factors.*

In my talk, I started by describing tensor networks (using the paper "Hand-waving and Interpretive Dance: An Introductory Course on Tensor Networks" [6] as a guide). I then described graph matrices, norm bounds on graph matrices, and the close connection between tensor networks and graph matrices. In particular, tensor networks which are flattened into matrices can be transformed into graph matrices by replacing the indices with vertices and replacing the matrix/tensor entries with edges/hyperedges. Finally, I illustrated the power of graph matrices by showing how they can be used to easily rederive part of the analysis for tensor PCA, the faster tensor PCA algorithm in [12], and the tensor decomposition algorithm in [9].

References

[1] Kwangjun Ahn, Dhruv Medarametla, and Aaron Potechin. Graph Matrices: Norm Bounds and Applications. arXiv 1604.03423, 2020.

[2] Mitali Bafna, Jun-Ting Hsieh, Pravesh Kothari, and Jeff Xu. Polynomial-Time Power-Sum Decomposition of Polynomials. FOCS 2022.

[3] Afonso Bandeira, Kevin Lucca, Petar Nizić-Nikolac, and Ramon van Handel. Matrix Chaos Inequalities and Chaos of Combinatorial Type. STOC 2025.

[4] Boaz Barak, Samuel Hopkins, Jonathan Kelner, Pravesh Kothari, Ankur Moitra, and Aaron Potechin. A Nearly Tight Sum-of-Squares Lower Bound for the Planted Clique Problem. SIAM Journal on Computing Vol. 48, Issue 2, p.687-735, 2019.

[5] Jacob Bridgeman and Christopher Chubb. Hand-waving and Interpretive Dance: An Introductory Course on Tensor Networks. arXiv:1603.03039, 2017.

[6] Wenjun Cai and Aaaron Potechin. The Spectrum of the Singular Values of Z-Shaped Graph Matrices. arXiv 2006.14144, 2020.

[7] Wenjun Cai and Aaaron Potechin. On Mixing Distributions Via Random Orthogonal Matrices and the Spectrum of the Singular Values of Multi-Z Shaped Graph Matrices. arXiv 2206.02224, 2022.

[8] Rong Ge and Tengyu Ma. Decomposing overcomplete 3rd order tensors using sum-of-squares algorithms. APPROX/RANDOM 2015.

[9] Mrinalkanti Ghosh, Fernando Granha Jeronimo, Chris Jones, Aaron Potechin, and Goutham Rajendran. Sum-of-Squares Lower Bounds for Sherrington-Kirkpatrick via Planted Affine Planes. FOCS 2020.

[10] Samuel Hopkins, Pravesh Kothari, Aaron Potechin. Prasad Raghavendra, Tselil Schramm, and David Steurer. The Power of Sum-of-Squares for Detecting Hidden Structures. FOCS 2017.

[11] Samuel Hopkins, Tselil Schramm, Jonathan Shi, and David Steurer. Fast spectral algorithms from sum-of-squares proofs: tensor decomposition and planted sparse vectors. STOC 2016.

[12] Jun-Ting Hsieh, Pravesh Kothari, Aaron Potechin, and Jeff Xu. Ellipsoid Fitting Up to a Constant. ICALP 2023.

[13] Chris Jones and L. Pesenti. Fourier Analysis of Iterative Algorithms. arXiv 2404.07881, 2024.

[14] Chris Jones, Aaron Potechin, Goutham Rajendran, Madhur Tulsiani, and Jeff Xu. Sum-of-Squares Lower Bounds for Sparse Independent Set. FOCS 2021.

[15] Chris Jones, Aaron Potechin, Goutham Rajendran, and Jeff Xu. Sum-of-Squares Lower Bounds for Densest k-Subgraph. STOC 2023.

[16] Pravesh Kothari, Aaron Potechin, and Jeff Xu. Sum-of-Squares Lower Bounds for Independent Set on Ultra-Sparse Random Graphs. STOC 2024.

[17] Shuo Pang. SOS lower bound for exact planted clique. CCC 2021.

[18] Aaron Potechin and Goutham Rajendran. Machinery for Proving Sum-of-Squares Lower Bounds on Certification Problems. arXiv 2011.04253, 2020.

[19] Aaron Potechin, Paxton Turner, Prayaag Venkat, and Alex Wein. Near-optimal fitting of ellipsoids to random points. COLT 2023.

[20] Aaron Potechin and Jeff Xu. Sum-of-Squares Lower Bounds for Coloring Random Graphs. STOC 2025.

[21] Goutham Rajendran and Madhur Tulsiani. Concentration of polynomial random matrices via Efron-Stein inequalities. SODA 2023.

[22] Madhur Tulsiani and June Wu. Simple Norm Bounds for Polynomial Random Matrices via Decoupling. ITCS 2025.

[23] Jeff Xu. Switching Graph Matrix Norm Bounds: from i.i.d. to Random Regular Graphs. CCC 2025.

University of Chicago
Email address: `potechin@uchicago.edu`

REPRESENTATION RETRIEVAL LEARNING FOR HETEROGENEOUS DATA INTEGRATION

ANNIE QU

Classification AMS 2020: 68T07

Keywords: Block-missing Data; Multi-task Learning; Multi-modality Data; Representation learning.

In the era of big data, large-scale, multi-modal datasets are increasingly ubiquitous, offering unprecedented opportunities for predictive modeling and scientific discovery. However, these datasets often exhibit complex heterogeneity—such as covariate shift, posterior drift, and missing modalities—that can hinder the accuracy of existing prediction algorithms. To address these challenges, we propose a novel Representation Retrieval (R^2) framework, which integrates a representation learning module (the *representer*) with a sparsity-induced machine learning model (the *learner*). Moreover, we introduce the notion of "integrativeness" for representers, characterized by the effective data sources used in learning representers, and propose a *Selective Integration Penalty* (SIP) to explicitly improve the property. Theoretically, we demonstrate that the R^2 framework relaxes the conventional full-sharing assumption in multi-task learning, allowing for partially shared structures, and that SIP can improve the convergence rate of the excess risk bound. Extensive simulation studies validate the empirical performance of our framework, and applications to two real-world datasets further confirm its superiority over existing approaches.

Large-scale data integration has made transformative contributions across numerous fields, including computer vision, natural language processing, biomedicine, genomics and healthcare. For example, in biomedicine, integrating randomized clinical trials and observational studies is of great interest, as it leverages the benefits of both data sources [19, 10, 2]. In genomics, multi-modality and multi-batch assays enable the discovery of cellular heterogeneity and development [5, 3]. In healthcare, multiple types of time-series measurements, such as cardiovascular, physical activities, and sleep data are integrated to improve real-time health and well-being monitoring [21, 13, 9]. However, integration of large-scale data effectively remains challenging, particularly when data are collected from diverse sources or populations, and across various collections of variables and modalities.

In particular, integrating large-scale data is challenging primarily due to various types of heterogeneity. First, the marginal distribution of the same covariate is often heterogeneous across different sources or populations, a phenomenon called "distribution heterogeneity", or "covariate shift" in the literature [11]. Second, in the context of supervised learning, the conditional distribution of responses given covariates could be heterogeneous, which is named "posterior heterogeneity", or "posterior drift" [16]. Third, observed covariates or modalities are often not uniformly measured: some covariates are observed across all data sources, while others are observed in only partial data sources. We refer to this as “observation heterogeneity” or “block missing”, which is considered in existing works [20, 18, 1].

In the current literature, various problem setups related to integrative supervised learning have been studied, while most of them only concern one or two types of the aforementioned heterogeneity. In particular, distribution heterogeneity, posterior heterogeneity, or both are considered in multi-task learning or transfer learning [16, 6, 14, 15]. Observation heterogeneity has been studied in multi-source data integration [20, 18, 17]. Recent work [1] has addressed all three types of heterogeneity in the transfer learning problem; however, their distributional and linear model assumption restrict their applicability for more general contexts. Sui et al. [12] propose a deep learning-based method to handle all three types of heterogeneity, where one modality is required to be observed among all data sources, which is restrictive in practice. In this work, we target the integrative supervised learning problem and aim to improve the predictive performance for all data sources. Our framework can accommodate all three types of heterogeneity and allow for nonparametric modeling for complex association between covariates and responses. Indeed, incorporating all three types of heterogeneity within a unified framework would allow for the integration of much broader datasets, thereby enhancing prediction performance by leveraging substantially more information.

Contributions: Our method offers several significant contributions. Methodologically, we introduce the Representation Retrieval (R^2) framework, which constructs a dictionary of representers—such as neural networks [7], kernels [4], or smoothing function bases [8]—to capture the complex distribution across multiple data sources. For each data source, a sparse learner built upon this dictionary selectively retrieves the most informative representers for prediction. Notably, the R^2 framework flexibly accommodates partially shared structures among data sources via a sparsity-inducing penalty.

Moreover, we introduce the concept of “integrativeness” of representers, defined as the effective data sources utilized for learning representers. To directly encourage the integrativeness of representers, we propose an innovative Selective Integration Penalty (SIP). Theoretically, we derive an excess risk bound for the R^2 framework, explicitly controlled by the integrativeness of representers, thereby demonstrating that SIP effectively enhances the model’s generalization performance. Computationally, we develop an efficient alternating minimization algorithm to iteratively update both the representer dictionary and the sparse learners. Extensive simulation studies and real-world applications further support the superior performance of our proposed method.

References

[1] Chang, J. H., Russo, M., and Paul, S. (2024). Heterogeneous transfer learning for high dimensional regression with feature mismatch. *arXiv preprint arXiv:2412.18081*.

[2] Colnet, B., Mayer, I., Chen, G., Dieng, A., Li, R., Varoquaux, G., Vert, J.-P., Josse, J., and Yang, S. (2024). Causal inference methods for combining randomized trials and observational studies: a review. *Statistical science*, 39(1):165–191.

[3] Du, J.-H., Cai, Z., and Roeder, K. (2022). Robust probabilistic modeling for single-cell multimodal mosaic integration and imputation via scvaeit. *Proceedings of the National Academy of Sciences*, 119(49):e2214414119.

[4] Koltchinskii, V. and Yuan, M. (2010). Sparsity in multiple kernel learning. *The Annals of Statistics*, 38(6):3660–3695.

[5] Kriebel, A. R. and Welch, J. D. (2022). Uinmf performs mosaic integration of single-cell multi-omic datasets using nonnegative matrix factorization. *Nature communications*, 13(1):780.

[6] Li, S., Cai, T. T., and Li, H. (2022). Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):149–173.

[7] Ma, J., Zhao, Z., Yi, X., Chen, J., Hong, L., and Chi, E. H. (2018). Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1930–1939.

[8] Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71(5):1009–1030.

[9] Rim, J., Xu, Q., Tang, X., Guo, Y., and Qu, A. (2025). Individualized time-varying nonparametric model with an application in mobile health. *Statistics in Medicine*, 44(5):e70005.

[10] Shi, X., Pan, Z., and Miao, W. (2023). Data integration in causal inference. *Wiley Interdisciplinary Reviews: Computational Statistics*, 15(1):e1581.

[11] Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244.

[12] Sui, Y., Xu, Q., Bai, Y., and Qu, A. (2025). Multi-task learning for heterogeneous multi-source block-wise missing data. *openreview.net*.

[13] Sun, X., Zhao, B., and Xue, F. (2025). Generalized heterogeneous functional model with applications to large-scale mobile health data. *arXiv preprint arXiv:2501.01135*.

[14] Tian, Y. and Feng, Y. (2023). Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association*, 118(544):2684–2697.

[15] Tian, Y., Gu, Y., and Feng, Y. (2023). Learning from similar linear representations: Adaptivity, minimaxity, and robustness. *arXiv preprint arXiv:2303.17765*.

[16] Widmer, G. and Kubat, M. (1996). Learning in the presence of concept drift and hidden contexts. *Machine learning*, 23:69–101.

[17] Xue, F., Ma, R., and Li, H. (2025). Statistical inference for high-dimensional linear regression with blockwise missing data. *Statistica Sinica*, 35:431–456.

[18] Xue, F. and Qu, A. (2021). Integrating multisource block-wise missing data in model selection. *Journal of the American Statistical Association*, 116(536):1914–1927.

[19] Yang, S. and Ding, P. (2020). Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association*, 115(531):1540–1554.

[20] Yu, G., Li, Q., Shen, D., and Liu, Y. (2020). Optimal sparse linear prediction for block-missing multi-modality data without imputation. *Journal of the American Statistical Association*, 115(531):1406–1419.

[21] Zhang, J., Xue, F., Xu, Q., Lee, J., and Qu, A. (2024). Individualized dynamic latent factor model for multi-resolutional data with application to mobile health. *Biometrika*, 111(4):1257–1275.

DEPARTMENT OF STATISTICS AND APPLIED PROBABILITY, UC SANTA BARBARA
Email address: `aqu2@ucsb.edu`

RELIABLE AND SCALABLE VARIABLE IMPORTANCE ESTIMATION

GARVESH RASKUTTI

Classification AMS 2020: 62C99, 62E17

Keywords: variable importance; neural tangent kernel; early stopping

As opaque black-box predictive models become more prevalent, the need to develop interpretations for these models is of great interest. The concept of *variable importance* and *Shapley values* are interpretability measures that applies to any predictive model and assesses how much a variable or set of variables improves prediction performance. Such approaches are also critical in *network estimation* when comparing different methods or using model-agnostic approaches. When the number of variables is large, estimating variable importance presents a significant computational challenge because re-training neural networks or other black-box algorithms requires significant additional computation. In this talk, we address this challenge for algorithms using gradient descent and gradient boosting (e.g. neural networks, gradient-boosted decision trees). By using the ideas of early stopping of gradient-based methods in combination with warm-start using the *dropout* method, we develop a scalable method to estimate variable importance for any algorithm that can be expressed as an *iterative kernel update equation*. Importantly, we provide theoretical guarantees by using the theory for early stopping of kernel-based methods for neural networks with sufficiently large (but not necessarily infinite) width and gradient-boosting decision trees that use symmetric trees as a weaker learner. We also demonstrate the efficacy of our methods through simulations and a real data example which illustrates the computational benefit of early stopping rather than fully re-training the model as well as the increased accuracy of our approach. This work is based on joint work with PhD students Zexuan Sun. This work is also related to prior work using a penalized apporach in (2).

0.1. **Our Contributions.** The main contributions of the talk are as follows:

- We propose a general scalable framework with supporting theoretical guarantees to estimate VI efficiently for any iterative algorithm that can be expressed as an *iterative kernel update equation*. Importantly we provide theoretical guarantees for this method by leveraging theory for the early stopping of kernel-based methods.
- Utilizing the *neural tangent kernel* for neural networks with sufficiently large (but not infinite) width we apply our general theoretical bound to feed-forward neural networks. Further, we use a well-defined kernel to also adapt our bounds to gradient boosted decision trees. Each of these theoretical results is of independent interest. Moreover, if the VI estimator is constructed using neural network, the asymptotically normality holds and we can build Wald-type confidence interval accordingly.

- As an interesting side product of our theoretical results, we find that under warm-start initialization, the global convergence of Neural tangent kernel still holds and the network behave approximately as linear model under mean square error loss .
- The theoretical bounds are supported in a simulation. We then demonstrate the computational advantages over re-training and the the accuracy advantages over dropout for estimating both variable importance.

1. Our algorithm

First we introduce our overall algorithm based on warm-start involving the full pre-trained model and then applying the early stopping strategy.

Algorithm 1 Early stopping training for VI_I

```
Input Data $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^N$, training size $N_1 = qN$; $N_2 = (1-q)N$; Kernel based model: $f_\theta(\cdot)$; Drop features set $I \subseteq \{1, \ldots, p\}$; Patience $P$;
Train full model $f_{N_1}^c$ using training data $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^{N_1}$;
Replace feature $j \in I$ with its empirical mean to get $\boldsymbol{X}_i^{(I)}$;
Split $\left\{(\boldsymbol{X}_i^{(I)}, Y_i)\right\}_{i=1}^{N_1}$ into a training set $\mathcal{D}_1$ of sample size $qN_1$ and a validation set $\mathcal{D}_2$ of sample size $(1-q)N_1$;
Initialize model with $f_{N_1}^c$, for each epoch $\tau$, train on $\mathcal{D}_1$, evaluate on $\mathcal{D}_2$;
If the loss evaluated on $\mathcal{D}_2$ at epoch $\widehat{T}$ has no improvement after $P$ epochs, stop and return $f_{\widehat{T}}$ as the estimator for $f_{0,-I}$;
Use remaining $N_2$ instances to construct $\widehat{VI}_I$ and plug in estimate of $\tau_{N,I}$
    $\widehat{\mathrm{VI}}_I = \frac{1}{N_2} \sum_{i=1}^{N_2} \left[Y_i - f_{\widehat{T}}(\mathbf{X}_i^{(I)})\right]^2 - \left[Y_i - f_{N_1}^c(\mathbf{X}_i)\right]^2$
    $t_{i,I} = \left(Y_i - f_{\widehat{T}}\left(\mathbf{X}_i^{(I)}\right)\right)^2 - \left(Y_i - f_{N_1}^c\left(\mathbf{X}_i\right)\right)^2$
    $\hat{\tau}_{N,I} = \frac{1}{N_2} \sum_{i=1}^{N_2} \left(t_{i,I} - \bar{t}_I\right)^2 / N_2$
Construct $\alpha$-level Wald-type CI as $\widehat{\mathrm{VI}}_I \pm z_{\frac{\alpha}{2}} \cdot \hat{\tau}_{N,I}$.
```

2. Theoretical bounds

Our theoretical bounds provide a general result for early stopping based on the algorithm above for any gradient-based approach.

Theorem 1 (General convergence bound under fixed design). *Under suitable regularity assumptions, consider our above algorithm with full model f_N^c, and stop the update early at iteration $\widehat{T}_{op}$, with high probability, the following bound holds*

$$\|f_{\widehat{T}_{op}} - f_{0,-I}\|_N^2 \leq \mathcal{O}\left(\frac{1}{N^{\frac{1}{2}}}\right).$$

Using additional techniques we are able to provide a Wald-type confidence interval for feed-forward neural networks with "large width."

Corollary 1. *Applying to feed-forward network with ReLu activation and no bias can accurately predict the reduced model, i.e.,*

$$\left\|f_{\widehat{T}_{op}} - f_{0,-I}\right\|_2 = O_p\left(N^{-1/4}\right). \tag{2.1}$$

Then if $\mathrm{VI}_I \neq 0$*, our variable importance estimator* $\widehat{\mathrm{VI}}_I$ *is asymptotically normal and has an error rate* $O_p\left(N^{-1/2}\right)$*:*

$$\widehat{\mathrm{VI}}_I - \mathrm{VI}_I = \Delta_{N,I} + O_p\left(N^{-1/2}\right) \tag{2.2}$$

where

$$\Delta_{N,I} \to_d \mathcal{N}\left(0, \tau^2_{N,j}\right) \tag{2.3}$$

here the variance is $\tau^2_{N,j} = \mathrm{Var}\left(w^{(I)^2} - w^2\right)/N$*, where* w *and* $w^{(I)}$ *are the population version of the residuals.*

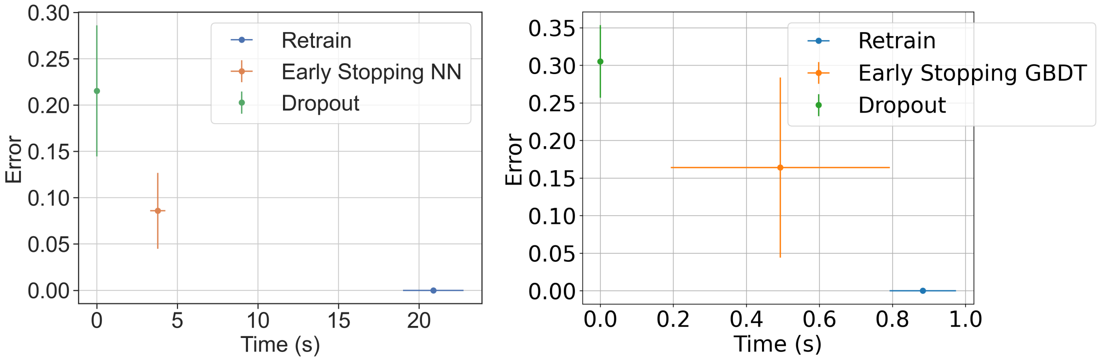
A key component of the proof is the use of the so-called *neural tangent kernel* first introduced in (3). the Neural Tangent Kernel (NTK) provides a theoretical tool to study the neural network in the RKHS regime. Denote a neural network by $f(\theta, x)$,

$$\left\langle \nabla_\theta f(\theta, x), \nabla_\theta f(\theta, x')\right\rangle. \tag{2.4}$$

By defining this kernel and adapting techniques based on early stopping applied to kernel ridge regression (1) allows us to prove the main result and corollary.

3. Simulation

Our simulation study reveals that as expected, early stopping after a small number of iterations (1-5) achieves performance close to full re-training while accounting for significant savings in computation.



(A) Neural Networks

(B) GBDT

FIGURE 1. Distribution of computation time vs. normalized estimation error relative to retrain for the VI of X_1.

References

[1] Garvesh Raskutti and Martin J. Wainwright and Bin Yu, Early Stopping and Non-parametric Regression: An Optimal Data-dependent Stopping Rule, Journal of Machine Learning Research, 15, p. 335–366, 2014

[2] Yue Gao and Abby Stevens and Garvesh Raskutti and Rebecca Willett, Lazy Estimation of VI for Large NNs Proceedings of the 39th International Conference on Machine Learning, 2022, Baltimore, Maryland, USA

[3] Arthur Jacot and Franck Gabriel and Clément Hongler, Neural Tangent Kernel: Convergence and Generalization in Neural Networks 32nd Conference on Neural Information Processing Systems, 2018,Montréal , Canada

Email address: `raskutti@stat.wisc.edu`

REGRESSION UNDER NETWORK INTERFERENCE

MICHAEL SCHWEINBERGER

Classification AMS 2020: 62A09, 62J02, 68P99

Keywords: Dependent Data, Generalized Linear Models, Minorization-Maximization, Pseudo-Likelihood

This extended abstract is based on Fritz, Schweinberger, Bhadra, and Hunter (2024) and Stewart and Schweinberger (2025).

1. Network interference

In connected populations, the treatments and outcomes of units can affect the outcomes of other units, which implies that the outcomes of units are interdependent. To study causal and non-causal relationships among attributes under network interference, a comprehensive regression framework for dependent predictors $\boldsymbol{X}$, outcomes $\boldsymbol{Y}$, and connections $\boldsymbol{Z}$ is needed.

2. Regression under network interference

We introduce a comprehensive regression framework for dependent predictors $\boldsymbol{X}$, outcomes $\boldsymbol{Y}$, and connections $\boldsymbol{Z}$ (Fritz et al., 2024). The regression framework can be used for studying non-causal and causal relationships among attributes $(\mathbf{X}, \mathbf{Y})$ of connected units and captures attribute-attribute, attribute-connection, and connection-connection dependencies, while retaining the advantages of linear regression, logistic regression, and other regression models by being interpretable and widely applicable. Scalable statistical computing is based on convex optimization of pseudo-likelihoods using minorization-maximization algorithms. An application to hate speech on social media demonstrates the advantages of the regression framework.

3. Theoretical guarantees

Theoretical guarantees for regression under network interference are non-trivial, because the outcomes and connections $(\boldsymbol{Y}, \boldsymbol{Z}) \mid \boldsymbol{X} = \boldsymbol{x}$ are dependent. We provide theoretical guarantees by generalizing results of Stewart and Schweinberger (2025) for dependent connections $\boldsymbol{Z}$ to dependent outcomes and connections $(\boldsymbol{Y}, \boldsymbol{Z}) \mid \boldsymbol{X} = \boldsymbol{x}$.

Lemma 1 of Stewart and Schweinberger (2025). *Let $g : \mathbb{R}^p \mapsto \mathbb{R}^p$ ($p \geq 1$) be a homeomorphism and $\| \cdot \|$ be a vector norm with induced matrix norm $\||\cdot\||$. Consider any $\boldsymbol{\theta}^\star \in \mathbb{R}^p$ and any $\epsilon > 0$, and define*

$$\delta(\epsilon) \;\; := \inf_{\boldsymbol{\theta} \,\in\, \mathrm{bd}\, \mathscr{B}(\boldsymbol{\theta}^\star,\, \epsilon)} \|g(\boldsymbol{\theta}) - g(\boldsymbol{\theta}^\star)\|,$$

where $\mathscr{B}(\boldsymbol{c}, \rho) := \{\boldsymbol{a} \in \mathbb{R}^p : \|\boldsymbol{a} - \boldsymbol{c}\| < \rho\}$ is a ball with center $\boldsymbol{c} \in \mathbb{R}^p$ and radius $\rho > 0$ and $\mathrm{bd}\, \mathscr{B}(\boldsymbol{c}, \rho)$ is the boundary of $\mathscr{B}(\boldsymbol{c}, \rho)$. If $g(\boldsymbol{\theta})$ is continuously differentiable and

$\mathcal{I}(\boldsymbol{\theta}) := \nabla_{\boldsymbol{\theta}}\, g(\boldsymbol{\theta})$ is invertible for all $\boldsymbol{\theta} \in \mathscr{B}(\boldsymbol{\theta}^\star,\, \epsilon)$, then

$$\frac{\epsilon}{\sup_{\boldsymbol{\theta} \,\in\, \mathscr{B}(\boldsymbol{\theta}^\star,\, \epsilon)}\, |||\mathcal{I}(\boldsymbol{\theta})^{-1}|||} \;\; \leq \;\; \delta(\epsilon). \quad \square$$

Lemma 1 helps "transport" concentration-of-measure between homeomorphic spaces, facilitating rates of convergence. To demonstrate, consider regression models with exponential-family densities of the form $f_{\boldsymbol{\theta}^\star}(\boldsymbol{t}) \propto e^{\langle \boldsymbol{\theta}^\star,\, \boldsymbol{t}\rangle}$, where $\boldsymbol{\theta}^\star \in \mathbb{R}^p$ and $\boldsymbol{\mu}(\boldsymbol{\theta}^\star) := \mathbb{E}_{\boldsymbol{\theta}^\star}\, \boldsymbol{T} \in \mathbb{R}^p$ are the data-generating natural and mean-value parameters of the exponential family, and $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\mu}(\widehat{\boldsymbol{\theta}}) = \boldsymbol{T}$ are the maximum likelihood estimators of $\boldsymbol{\theta}^\star$ and $\boldsymbol{\mu}(\boldsymbol{\theta}^\star) = \mathbb{E}_{\boldsymbol{\theta}^\star}\, \boldsymbol{T}$, respectively. Since the natural and mean-value parameter spaces of exponential families are homeomorphic, Lemma 1 implies that

$$\begin{array}{rcll} \mathbb{P}(\widehat{\boldsymbol{\theta}} \in \mathscr{B}(\boldsymbol{\theta}^\star,\, \epsilon)) & = & \mathbb{P}(\boldsymbol{T} \in \boldsymbol{\mu}(\mathscr{B}(\boldsymbol{\theta}^\star,\, \epsilon))) & \textit{because } \boldsymbol{\mu} \textit{ is a homeomorphism} \\[2ex] & \geq & \mathbb{P}(\boldsymbol{T} \in \mathscr{B}(\boldsymbol{\mu}(\boldsymbol{\theta}^\star),\, \delta(\epsilon))) & \textit{by definition of } \delta(\epsilon) \\[2ex] & \geq & 1 - \alpha(\delta(\epsilon)) & \textit{by concentration of } \boldsymbol{T} \\[2ex] & \geq & 1 - \alpha\left(\dfrac{\epsilon}{\sup_{\boldsymbol{\theta} \,\in\, \mathscr{B}(\boldsymbol{\theta}^\star,\, \epsilon)}\, |||\mathcal{I}(\boldsymbol{\theta})^{-1}|||}\right) & \textit{by Lemma 1 applied to } \boldsymbol{\mu}, \end{array}$$

where $\alpha(.)$ is a non-increasing function that quantifies the strength of concentration of $\boldsymbol{T}$ around $\boldsymbol{\mu}(\boldsymbol{\theta}^\star) = \mathbb{E}_{\boldsymbol{\theta}^\star}\, \boldsymbol{T}$. In other words: If the probability mass of $\boldsymbol{\mu}(\widehat{\boldsymbol{\theta}}) = \boldsymbol{T}$ concentrates around $\boldsymbol{\mu}(\boldsymbol{\theta}^\star) = \mathbb{E}_{\boldsymbol{\theta}^\star}\, \boldsymbol{T}$, then the probability mass of $\widehat{\boldsymbol{\theta}}$ concentrates around $\boldsymbol{\theta}^\star$, paving the way for convergence rates for $\widehat{\boldsymbol{\theta}}$ based on $\boldsymbol{\mu}(\widehat{\boldsymbol{\theta}}) = \boldsymbol{T}$ (compare Theorems 1 and 2 of Stewart and Schweinberger, 2025). While specific convergence rates depend on additional properties of the data-generating model, the above argument suggests that the convergence rate of maximum likelihood estimators $\widehat{\boldsymbol{\theta}}$ depends on

- the precision in a neighborhood of $\boldsymbol{\theta}^\star$ as quantified by $\sup_{\boldsymbol{\theta} \,\in\, \mathscr{B}(\boldsymbol{\theta}^\star,\, \epsilon)}\, |||\mathcal{I}(\boldsymbol{\theta})^{-1}|||$;
- the strength of concentration of $\boldsymbol{T}$ as quantified by $\alpha(.)$, which depends on the tails of the distribution of $\boldsymbol{T}$ and the dependence induced by the model.

The above argument applies to all exponential families (e.g., generalized linear models, graphical models, and Gaussian and non-Gaussian Markov random fields), and helps establish theoretical guarantees for regression based on independent or dependent observations, including regression under network interference (Fritz et al., 2024).

References

Fritz, C., M. Schweinberger, S. Bhadra, and D. R. Hunter (2024). A regression framework for studying relationships among attributes under network interference. *Available from: arXiv:2410.07555*.

Stewart, J. R. and M. Schweinberger (2025). Pseudo-likelihood-based M-estimators for random graphs with dependent edges and parameter vectors of increasing dimension. *The Annals of Statistics*. To appear.

Department of Statistics, The Pennsylvania State University, 326 Thomas Building, University Park, PA 16802, United States of America

Email address: michael.schweinberger@psu.edu

A MULTILAYER PROBIT NETWORK MODEL FOR COMMUNITY DETECTION WITH DEPENDENT LAYERS

DAPENG SHI

Classification AMS 2020: 62A09, 91D30

Keywords: inter-layer dependence; sparse networks; multivariate probit model; stochastic block model.

Multilayer networks often exhibit various dependence structures between network layers. Various inter-layer dependence modeling highlights the importance of incorporating such dependencies for more accurate and efficient network analysis. For example, [3] introduced the autoregressive stochastic block model (SBM) to capture inter-layer dependence with a time series structure; [7] proposed the multilayer Ising model to capture the full inter-layer dependence. However, it remains unclear how to extend [3] to accommodate more general dependence structures, whereas the method in [7] appears to have difficulty in estimating connection probabilities due to the intractable computation cost of the partition function [5]. Moreover, very little has been done in the literature to theoretically investigate the impact of dependence structures on the community detection accuracy.

In this work, we introduce a novel multilayer probit network model that integrates the classical multilayer SBM [4, 2] with the multivariate probit model [1]. It incorporates diverse inter-layer dependence structures between layers into network modeling so as to achieve better estimation of the homogeneous community structure.

Let $\mathcal{G}$ denote a multilayer network comprising M network layers on N common nodes, where each network layer can be represented via its adjacency matrix $\boldsymbol{A}^{(b)} = (A_{ij}^{(b)})_{N\times N} \in \{0,1\}^{N\times N}$ for $b \in [M]$. Here, $A_{ij}^{(b)} = A_{ji}^{(b)} = 1$ if an edge exists between nodes i and j in the b-th layer, and $A_{ij}^{(b)} = A_{ji}^{(b)} = 0$ otherwise. We consider the following multilayer probit network model,

$$\begin{aligned} &A_{ij}^{(b)} = \mathbb{I}\big\{\mu_{e_ie_j}^{(b)} + \varepsilon_{ij}^{(b)} > 0\big\}, \text{ for any } b \in [M],\\ &\big(\varepsilon_{ij}^{(1)}, \cdots, \varepsilon_{ij}^{(M)}\big)^\top \sim N\big(0, \boldsymbol{\Sigma}_{e_ie_j}\big), \text{ for any } i \neq j, \end{aligned}$$

where $\mathbb{I}(\cdot)$ is the indicator function, $e_i \in [K]$ denotes the homogeneous community membership of node i across M layers, $\boldsymbol{\mu}^{(b)} \in \mathbb{R}^{K\times K}$ denotes the mean matrix for each network layer, and $\boldsymbol{\Sigma}_{kl} \in \mathbb{R}^{M\times M}$ is positive definite for any $k,l \in [K]$. Note that $P(A_{ij}^{(b)} = 1) = P(\varepsilon_{ij}^{(b)} > -\mu_{e_ie_j}^{(b)}) = \Phi(\mu_{e_ie_j}^{(b)})$, where $\Phi(\cdot)$ is the cumulative distribution function of $N(0,1)$.

Let $\boldsymbol{\mu} = (\mu_{kl}^{(b)})_{k,l\in[K],b\in[M]}$ and $\boldsymbol{\mu}^{(b)} = (\mu_{kl}^{(b)})_{k,l\in[K]}$ for each $b \in [M]$. Further, let $\boldsymbol{\Sigma} = (\Sigma_{kl}^{(bd)})_{k,l\in[K],b,d\in[M]}$ and $\boldsymbol{\Sigma}_{kl} = (\Sigma_{kl}^{(bd)})_{b,d\in[M]}$ for any $k,l \in [K]$. Define $\boldsymbol{\Theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Theta}_{kl}^{(bd)} = (\mu_{kl}^{(b)}, \mu_{kl}^{(d)}, \Sigma_{kl}^{(bd)})$. Denote $\boldsymbol{Z} = (Z_{ik})_{i\in[N];k\in[K]}$ as the homogeneous community membership matrix, where $Z_{ik} = 1$ if $e_i = k$, and $Z_{ik} = 0$ otherwise. Since the full likelihood of the multilayer network is computationally inefficient, we consider a pairwise likelihood function as an alternative, which largely facilitates the computation without compromising estimation accuracy [6]. Specifically, we replace the full likelihood $\mathbb{P}\big(\boldsymbol{A}_{ij}; \{\mu_{kl}^{(b)}\}_{b=1}^{M}, \boldsymbol{\Sigma}_{kl}\big)$ with

$$\prod_{1\le b<d\le M} \mathbb{P}\big(A_{ij}^{(b)}, A_{ij}^{(d)}; \boldsymbol{\Theta}_{kl}^{(bd)}\big),$$

where

$$\begin{aligned}\mathbb{P}\big(A_{ij}^{(b)}, A_{ij}^{(d)}; \boldsymbol{\Theta}_{kl}^{(bd)}\big) = \alpha_1 \left(\boldsymbol{\Theta}_{kl}^{(bd)}\right)^{A_{ij}^{(b)}A_{ij}^{(d)}} \times \alpha_2 \left(\boldsymbol{\Theta}_{kl}^{(bd)}\right)^{A_{ij}^{(b)}(1-A_{ij}^{(d)})} \\ \times \alpha_3 \left(\boldsymbol{\Theta}_{kl}^{(bd)}\right)^{(1-A_{ij}^{(b)})A_{ij}^{(d)}} \times \alpha_4 \left(\boldsymbol{\Theta}_{kl}^{(bd)}\right)^{(1-A_{ij}^{(b)})(1-A_{ij}^{(d)})}.\end{aligned}$$

The terms $\alpha_1, \alpha_2, \alpha_3$ and α_4 are defined as
(0.1)

$$\begin{aligned}\alpha_1\big(\boldsymbol{\Theta}_{kl}^{(bd)}\big) &= \mathbb{P}\big(A_{ij}^{(b)} = 1, A_{ij}^{(d)} = 1; \boldsymbol{\Theta}_{kl}^{(bd)}\big) = \Phi_2\big(\mu_{kl}^{(b)}, \mu_{kl}^{(d)}, \Sigma_{kl}^{(bd)}\big), \\ \alpha_2\big(\boldsymbol{\Theta}_{kl}^{(bd)}\big) &= \mathbb{P}\big(A_{ij}^{(b)} = 1, A_{ij}^{(d)} = 0; \boldsymbol{\Theta}_{kl}^{(bd)}\big) = \Phi\big(\mu_{kl}^{(b)}\big) - \Phi_2\big(\mu_{kl}^{(b)}, \mu_{kl}^{(d)}, \Sigma_{kl}^{(bd)}\big), \\ \alpha_3\big(\boldsymbol{\Theta}_{kl}^{(bd)}\big) &= \mathbb{P}\big(A_{ij}^{(b)} = 0, A_{ij}^{(d)} = 1; \boldsymbol{\Theta}_{kl}^{(bd)}\big) = \Phi\big(\mu_{kl}^{(d)}\big) - \Phi_2\big(\mu_{kl}^{(b)}, \mu_{kl}^{(d)}, \Sigma_{kl}^{(bd)}\big), \\ \alpha_4\big(\boldsymbol{\Theta}_{kl}^{(bd)}\big) &= \mathbb{P}\big(A_{ij}^{(b)} = 0, A_{ij}^{(d)} = 0; \boldsymbol{\Theta}_{kl}^{(bd)}\big) = 1 - \Phi\big(\mu_{kl}^{(b)}\big) - \Phi\big(\mu_{kl}^{(b)}\big) + \Phi_2\big(\mu_{kl}^{(b)}, \mu_{kl}^{(d)}, \Sigma_{kl}^{(bd)}\big),\end{aligned}$$

where $\Phi(\cdot)$ is the cumulative distribution function of $N(0,1)$, and $\Phi_2(\cdot,\cdot,\sigma)$ is the cumulative distribution function of $N_2\big(\left(\begin{smallmatrix}0\\0\end{smallmatrix}\right), \left(\begin{smallmatrix}1 & \sigma\\ \sigma & 1\end{smallmatrix}\right)\big)$. The pairwise log-likelihood then becomes

$$\begin{aligned}\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{Z}) =& \sum_{k,l}\sum_{i,j}\sum_{b<d} Z_{ik}Z_{jl}\Big\{A_{ij}^{(b)}A_{ij}^{(d)} \log\alpha_1(\boldsymbol{\Theta}_{kl}^{(bd)}) + A_{ij}^{(b)}(1-A_{ij}^{(d)})\log\alpha_2(\boldsymbol{\Theta}_{kl}^{(bd)}) \\ &+ (1-A_{ij}^{(b)})A_{ij}^{(d)}\log\alpha_3\left(\boldsymbol{\Theta}_{kl}^{(bd)}\right) + (1-A_{ij}^{(b)})(1-A_{ij}^{(d)})\log\alpha_4(\boldsymbol{\Theta}_{kl}^{(bd)})\Big\} \\ =&: \sum_{k,l}\mathcal{L}_{kl}(\boldsymbol{\Theta}, \boldsymbol{Z}).\end{aligned}$$

Denote $\mathcal{S}_{kl}$ as the pre-specified, shape-constrained set for $\boldsymbol{\Sigma}_{kl}$. Specifically, we focus on two scenarios, the sparse covariance matrix scenario with

$$\mathcal{S}_{kl} = \left\{\boldsymbol{X} \in \mathbb{R}^{M\times M} \mid \boldsymbol{X} \succ 0, \mathrm{diag}(\boldsymbol{X}) = \boldsymbol{1}_M, \mathrm{Supp}(\boldsymbol{X}) = T_{kl}\right\},$$

and the sparse precision matrix scenario with

$$\mathcal{S}_{kl} = \left\{\boldsymbol{X} \in \mathbb{R}^{M\times M} \mid \boldsymbol{X} \succ 0, \mathrm{diag}(\boldsymbol{X}) = \boldsymbol{1}_M, \mathrm{Supp}(\boldsymbol{X}^{-1}) = T_{kl}\right\}.$$

In both cases, $T_{kl} \subseteq [M]\times[M]$ represents the set of positions, known a priori, with $|T_{kl}| = s_{kl}^*$. Two examples for each scenario are the multilayer Ising model [7] and the autoregressive SBM [3]. Define the parameter space as

$$\begin{aligned}\boldsymbol{\Omega} = \Big\{&\boldsymbol{\omega} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{Z}) \mid \boldsymbol{Z} \in \{0,1\}^{N\times K},\ \boldsymbol{Z}\boldsymbol{1}_K = \boldsymbol{1}_N,\ c_l\rho_{N,M} \le \Phi(\mu_{kl}^{(b)}) \le c_u\rho_{N,M}, \\ &\boldsymbol{\Sigma}_{kl} \in \mathcal{S}_{kl},\ \text{and}\ \sup_{k,l}\|\,\mathrm{ndiag}(\boldsymbol{\Sigma}_{kl})\|_{\max} \le D_{N,M}\Big\},\end{aligned}$$

where $c_l < 1 < c_u$ are two constants and $\rho_{N,M}$ controls the network sparsity level. Note that the magnitudes of s_{kl}^* and $D_{N,M}$ specify the inter-layer dependence structures and the strength of dependence across different layers, respectively. Denote the true parameters as $\boldsymbol{\omega}^* = (\boldsymbol{\Theta}^*, \boldsymbol{Z}^*) = (\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*, \boldsymbol{Z}^*)$ and assume $\boldsymbol{\omega}^* \in \boldsymbol{\Omega}$. Lemma 0.1 shows that the pairwise likelihood function in (0.2) is Fisher consistent in $\boldsymbol{\Omega}$.

Lemma 0.1. *Let $\boldsymbol{e}(\boldsymbol{\omega}^*, \boldsymbol{\omega}) = \frac{1}{N^2M^2} \sum_{k,l} \mathbb{E}\big(\mathcal{L}_{kl}(\boldsymbol{\Theta}^*, \boldsymbol{Z}^*) - \mathcal{L}_{kl}(\boldsymbol{\Theta}, \boldsymbol{Z})\big)$, then it holds true that $\boldsymbol{e}(\boldsymbol{\omega}^*, \boldsymbol{\omega}) \geq 0$ for any $\boldsymbol{\omega} \in \boldsymbol{\Omega}$.*

Lemma 0.1 shows that $\boldsymbol{\omega}^*$ is a maximizer of $\mathbb{E}(\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{Z}))$, and thus justifies the use of the pairwise likelihood function in estimating $\boldsymbol{\omega}^*$. Therefore, we estimate $(\boldsymbol{\Theta}^*, \boldsymbol{Z}^*)$ via the constrained maximum pairwise log-likelihood estimate,

$$(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{Z}}) = \underset{(\boldsymbol{\Theta}, \boldsymbol{Z}) \in \boldsymbol{\Omega}}{\operatorname{argmax}} \mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{Z}). \tag{0.2}$$

We also adopt an alternative updating algorithm to solve the constrained optimization problem. Theoretically, we establish the asymptotic consistency of the proposed method for both parameter estimation and community detection under mild conditions.

We demonstrate how the inter-layer dependence structures and strength affect the accuracy of community detection in theory. In the autoregressive SBM, the proposed method exhibits a smaller misclassification rate than [3] when $\rho_{N,M} \gtrsim \frac{1}{\log(NM)}$ and $M \lesssim N$. In the multilayer Ising model [7] with $K \lesssim \log(NM), s^*_{kl} \asymp M^2, M \asymp N$, the required sparsity condition there is that $\rho_{N,M} \gg \left(\frac{1}{N}\right)^{\frac{1}{1+c}}$ for some constant $c > 0$, up to some logarithmic terms. In contrast, the proposed method can achieve $\rho_{N,M} \gg \frac{1}{N}$, up to some logarithmic terms, which achieves a better sparsity condition. Moreover, through extensive simulations and a real-world multilayer international trade network, we demonstrate the superior numerical performance of the proposed method compared to several popular competitors.

References

[1] Patrick J Heagerty and Subhash R Lele. A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association*, 93(443):1099–1111, 1998.

[2] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[3] Binyan Jiang, Jialiang Li, and Qiwei Yao. Autoregressive networks. *Journal of Machine Learning Research*, 24(227):1–69, 2023.

[4] Jing Lei, Kehui Chen, and Brian Lynch. Consistent community detection in multilayer network data. *Biometrika*, 107(1):61–73, 2020.

[5] Pradeep Ravikumar, Martin J Wainwright, and John D Lafferty. High-dimensional ising model selection using l1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

[6] Cristiano Varin, Nancy Reid, and David Firth. An overview of composite likelihood methods. *Statistica Sinica*, 21(1):5–42, 2011.

[7] Jingnan Zhang, Junhui Wang, and Xueqin Wang. Consistent community detection in inter-layer dependent multi-layer networks. *Journal of the American Statistical Association*, 119(548):3141–3151, 2024.

Department of Statistics, The Chinese University of Hong Kong, Sha Tin, Hong Kong
Email address: dapengshi@cuhk.edu.hk

ONLINE INFERENCE FOR LOW-RANK REINFORCEMENT LEARNING

WILL WEI SUN

Classification AMS 2020: 62L20, 90B50, 62E20

Keywords: online inference, online decision-making, low-rank models, reinforcement learning, stochastic gradient descent.

Reinforcement learning (RL) deals with how intelligent agents leverage contextual information and historical data to take actions in an uncertain environment in order to maximize the cumulative reward [1]. It has achieved phenomenal success in diverse fields, such as video games, robotics, autonomous driving, precision medicine, and recommendation systems. In modern applications, such context format can be rich and can often be formulated as a matrix or higher order tensor. This is evident in scenarios such as monitoring brain activity in real-time during clinical research, tracking dynamic user preferences in online recommender systems, and analyzing the evolving relationships in social network analysis. Consider neuroscience, where dynamic treatments may be tailored to a patient based on their neuroimaging. Here, the neuroimaging data forms a tensor state, while the treatment, such as dynamic sleep intervention, represents an action in the RL framework. Such high-dimensional higher-order tensor contexts necessitate the incorporation of low-rank structures in RL models.

Why inference in RL? While existing RL algorithms mainly focus on minimizing regret or choosing the best action with respect to some oracle policy, less attention has been paid to the statistical inference for RL models where the data are adaptively collected. In real-world applications of RL, we are often not just interested in obtaining the point estimate of the value function, but also a measure of the statistical uncertainty associated with the estimate. This is especially relevant to fields such as personalized medicine, mobile health and autonomous driving, where it is often risky to run a policy without a statistically sound estimate of its quality. For example, online A/B testing has been widely conducted by technological/pharmaceutical companies to compare a new product with an old one. Recent studies [2] have used various bandit or RL methods to form sequential online A/B testing procedures. In these online evaluation tasks, it is important to quantify the uncertainty of the point estimate for constructing a valid hypothesis testing. Moreover, the information obtained by conducting statistical inference of parameters or value functions, can eventually help experimenters to yield a better understanding in the used RL reward model, and this increase of knowledge can potentially improve the design of the experiments [3].

Why are new tools needed? When data is collected in an adaptive manner, even simple ordinary least squares can exhibit non-normal asymptotic behavior [3]. In this case, the confidence intervals constructed from traditional estimators induce bias and lead to wrong coverage. In extensive numerical studies, [4] empirically illustrate that common statistical hypothesis tests lead to as much as double the false positive rate and false negative rate using adaptive data collected in the bandit setting. While the use of adaptively collected data for inferential purposes has gained popularity in recent years, existing inferential methods are primarily developed under simple settings. These include adaptive linear regression [5], linear bandit [3], batch Markov decision process [6], and linear stochastic approximation under Markov noise [7]. While these contributions serve as crucial initial steps in statistical inference for adaptive data, practical RL applications involving higher-order tensor contexts call for more sophisticated inference tools.

In this talk, we discuss *provable online inferential tools tailored for low-rank reinforcement learning*. We first introduce an efficient online low-rank stochastic gradient descent (SGD) method and establishes its non-asymptotic rate of convergence. Building upon this foundation, we propose a simple yet powerful online debiasing approach for the sequential statistical inference of low-rank tensor learning. The entire online procedure studied in this context, encompassing both estimation and inference, eliminates the need for data splitting or storing historical data, making it suitable for on-the-fly hypothesis testing. We then progress to low-rank contextual bandit by incorporating online decision-making policies, where sequential decisions rely on higher-order contextual information. By conducting hypothesis testing on entries of the parameter tensor, one can assess the impact of a specific region of the tensor context on the reward. The challenges of this inference arise from *two sources of bias*: the first due to the low-rank structure of the parameter, and the second originating from the decision-making policy, as the chosen action depends on all historical data. We discuss an *online double debiasing* procedure for statistical inference within the low-rank contextual bandit framework, and establish the validity of the resulting confidence interval. Additionally, we identify an intriguing *tradeoff between parameter inference and regret minimization*, prompting a formulation of this trade-off as a minimax multi-objective optimization problem.

References

[1] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

[2] C. Shi, S. Luo, H. Zhu, and R. Song. An online sequential test for qualitative treatment effects. *Journal of Machine Learning Research*, 22(286):1–51, 2021.

[3] K. Zhang, L. Janson, and S. Murphy. Statistical inference with m-estimators on adaptively collected data. *Advances in Neural Information Processing Systems*, 34:7460–7471, 2021.

[4] J. J. Williams, J. Nogas, N. Deliu, H. Shaikh, S. S. Villar, A. Durand, and A. Rafferty. Challenges in statistical analysis of data collected by a bandit algorithm: An empirical exploration in applications to adaptively randomized experiments. *arXiv preprint arXiv:2103.12198*, 2021.

[5] K. Khamaru, Y. Deshpande, L. Mackey, and M. J. Wainwright. Near-optimal inference in adaptive linear regression. *arXiv preprint arXiv:2107.02266*, 2021.

[6] C. Shi, S. Zhang, W. Lu, and R. Song. Statistical inference of the value function for reinforcement learning in infinite-horizon settings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2021.

[7] P. Ramprasad, Y. Li, Z. Yang, Z. Wang, W. W. Sun, and G. Cheng. Online bootstrap inference for policy evaluation in reinforcement learning. *Journal of the American Statistical Association*, 2022 (just-accepted).

Purdue University
Email address: sun244@purdue.edu

THEORY AND APPLICATION OF HIGH-DIMENSIONAL SMOOTH TENSORS

MIAOYAN WANG

Keywords: Smooth tensors, Nonparametric Methods, Latent Permutation

Abstract. We consider the problem of structured tensor denoising in the presence of unknown permutations. Such data problems arise commonly in recommendation system, neuroimaging, community detection, and multiway comparison applications. Here, we develop a general family of smooth tensor models up to arbitrary index permutations; the model incorporates the popular tensor block models and Lipschitz hypergraphon models as special cases. We show that a constrained least-squares estimator in the block-wise polynomial family achieves the minimax error bound. A phase transition phenomenon is revealed with respect to the smoothness threshold needed for optimal recovery. In particular, we find that a polynomial of degree up to $(m-2)(m+1)/2$ is sufficient for accurate recovery of order-m tensors, whereas higher degree exhibits no further benefits. This phenomenon reveals the intrinsic distinction for smooth tensor estimation problems with and without unknown permutations. Furthermore, we provide an efficient polynomial-time Borda count algorithm that provably achieves optimal rate under monotonicity assumptions. The efficacy of our procedure is demonstrated through both simulations and Chicago crime data analysis.

Model. Let $\Theta \in \mathbb{R}^{d\times\cdots\times d}$ be an order-m d-dimensional tensor, $\pi\colon [d] \to [d]$ be an index permutation, and $\Theta(i_1,\ldots,i_m)$ the tensor entry indexed by $(i_1,\ldots,i_m)$. We sometimes also use shorthand notation $\Theta(\omega)$ for tensor entries with indices $\omega = (i_1,\ldots,i_m)$. Suppose we observe an order-m d-dimensional data tensor from the following model,

$$\mathcal{Y} = \Theta \circ \pi + \mathcal{E}, \tag{0.1}$$

where $\circ$ represents the function composition, $\pi\colon [d] \to [d]$ is an unknown latent permutation, $\Theta \in \mathbb{R}^{d\times\cdots\times d}$ is an unknown signal tensor under certain smoothness (to be specified in next paragraph), and $\mathcal{E}$ is a noise tensor consisting of zero-mean, independent sub-Gaussian entries with variance bounded by σ^2. The general model allows continuous- and binary-valued tensors. For instance, in binary tensor problems, the entries in $\mathcal{Y}$ are $\{0,1\}$-labels from Bernoulli distribution, and the entrywise variance of $\mathcal{E}$ depends on the mean. For ease of presentation, we assume $\sigma = 1$ throughout the paper. We call (0.1) the Gaussian model if the $\mathcal{E}$ consists of i.i.d. $\mathcal{N}(0,1)$ entries, and call (0.1) the sub-Gaussian model if $\mathcal{E}$ consists of independently (but not necessarily identically) distributed sub-Gaussian entries.

We now describe the smooth model on the signal Θ. Suppose that there exists a multivariate function $f\colon [0,1]^m \to \mathbb{R}$ underlying the signal tensor, such that

$$\Theta(i_1,\ldots,i_m) = f\left(\frac{i_1}{d},\ldots,\frac{i_m}{d}\right), \quad \text{for all } (i_1,\ldots,i_m) \in [d]^m. \tag{0.2}$$

For a multi-index $\kappa = (\kappa_1, \ldots, \kappa_m) \in \mathbb{N}^m$ and a vector $\boldsymbol{x} = (x_1, \ldots, x_m)^T$, we denote $|\kappa| = \sum_{i\in[m]} \kappa_i$, $\kappa! = \prod_{i\in[m]} \kappa_i!$, $\boldsymbol{x}^\kappa = \prod_{i\in[m]} x_i^{\kappa_i}$, and the derivative operator $\nabla_\kappa = \frac{\partial^{|\kappa|}}{\partial x_1^{\kappa_1}\cdots\partial x_m^{\kappa_m}}$. The generative function f in (0.2) is assumed to be in the α-Hölder smooth family [5].

Definition 0.1 (α-Hölder smooth). *Let $\alpha > 0$ and $L > 0$ be two positive constants. A function $f\colon [0,1]^m \to \mathbb{R}$ is called α-Hölder smooth, denoted as $f \in \mathcal{F}(\alpha, L)$, if*

$$\sum_{\kappa : |\kappa| = \lceil \alpha - 1 \rceil} \frac{1}{\kappa!} |\nabla_\kappa f(\boldsymbol{x}) - \nabla_\kappa f(\boldsymbol{x}_0)| \le L \|\boldsymbol{x} - \boldsymbol{x}_0\|_\infty^{\alpha - \lceil \alpha - 1 \rceil} \tag{0.3}$$

holds for every $\boldsymbol{x}, \boldsymbol{x}_0 \in [0,1]^m$.

The Hölder smooth function class is one of the most popular function classes considered in the nonparametric regression literature [3, 2]. In addition to the function class $\mathcal{F}(\alpha, L)$, we also define the smooth tensor class based on discretization (0.2),

$$\mathcal{P}(\alpha, L) = \left\{ \Theta \in \mathbb{R}^{d \times \cdots \times d} \colon \Theta \text{ is generated from (0.2) and } f \in \mathcal{F}(\alpha, L) \right\}. \tag{0.4}$$

Combining (0.1) and (0.2) yields our proposed *permuted smooth tensor model*. The unknown parameters are the smooth tensor $\Theta \in \mathcal{P}(\alpha, L)$ and latent permutation $\pi \in \Pi(d, d)$. The model is visualized in Figure 1(a) for the case $m = 2$ (matrices).

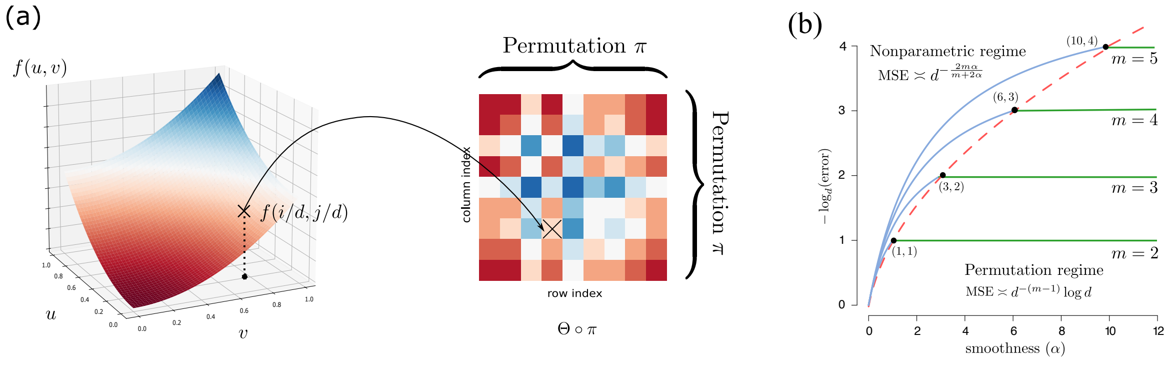


FIGURE 1. (a): Illustration of order-m d-dimensional permuted smooth tensor models with $m = 2$. (b): Phase transition of mean squared error (MSE) (on $-\log_d$ scale) as a function of smoothness α and tensor order m. Bold dots correspond to the critical smoothness level above which higher smoothness exhibits no further benefits to tensor estimation.

Results Summary. We develop a suite of statistical theory, efficient algorithms, and related applications for permuted smooth tensor models. Our contributions are summarized below. First, we develop a general permuted α-smooth tensor model of arbitrary smoothness level $\alpha > 0$. We establish the statistically optimal error rate and its dependence on model complexity. Specifically, we express the optimal rate as a function of tensor order m, tensor dimension d, and the smoothness level α, given by

$$\text{Rate}(d) := d^{-\frac{2m\alpha}{m+2\alpha}} \vee d^{-(m-1)} \log d. \tag{0.5}$$

Our framework substantially generalizes earlier works which focus on only matrices with $m = 2$ [2, 3] or Lipschitz function with $\alpha = 1$ [1, 4]. The generalization enables us to obtain results previously impossible: i) As tensor order m increases, we demonstrate the failure of pervious clustering-based algorithms [1, 2], and we develop a new block-wise polynomial algorithm for tensors of order $m \geq 3$; ii) As smoothness α increases, we demonstrate that the error rate converges to a fast rate $\mathcal{O}(d^{-(m-1)})$, thereby disproving the conjectured lower bound $\mathcal{O}(d^{-2m/(m+2)})$ posed by earlier work [1]. The results showcase the accuracy gain of our new approach, as well as the intrinsic distinction between matrices and higher-order tensors.

Second, we discover a phase transition phenomenon with respect to the smoothness needed for optimal recovery in the model (0.1) and (0.2). Figure 1(b) plots the dependence of estimation error in terms of smoothness level α for tensors of order m. We characterize two distinct error behaviors determined by a critical smoothness threshold. Specifically, the accuracy improves with α in the regime $\alpha \leq m(m-1)/2$, whereas the accuracy becomes a constant of α in the regime $\alpha > m(m-1)/2$. The results imply a polynomial of degree $(m-2)(m+1)/2 = [m(m-1)/2 - 1]$ is sufficient for accurate recovery of order-m tensors of arbitrary smoothness in the model (0.1) and (0.2)., whereas higher degree brings no further benefits. The phenomenon is distinctive from matrix problems [3, 2] and classical *non-permuted* smooth function estimation [5], thereby highlighting the fundamental challenges in our new setting. These statistical contributions, to our best knowledge, are new to the literature of permuted smooth tensor problems.

Third, we propose two estimation algorithms with accuracy guarantees: the least-squares estimation and Borda count estimation. The least-squares estimation, although being computationally hard, reveals the fundamental model complexity in the problem. The result serves as the benchmark and a useful guide to the algorithm design. Furthermore, we develop an efficient polynomial-time Borda count algorithm that provably achieves a minimax optimal rate under an extra Lipschitz monotonic assumption. The algorithm handles a broad range of data types, including continuous and binary observations.

Lastly, we illustrate the efficacy of our method through both simulations and data applications. A range of practical settings are investigated in simulations, and we show the outperformance of our method compared to alternative approaches. Application to Chicago crime data is presented to showcase the usefulness of our method. We identify the key global pattern and pinpoint local smooth structure in the denoised tensor. Our method will help practitioners efficiently analyze tensor datasets in various areas. Toward this end, the package and all data used are released at CRAN link `https://cloud.r-project.org/web/packages/SmoothTensor/index.html`.

References

[1] Balasubramanian, K. (2021). Nonparametric modeling of higher-order interactions via hypergraphons. *Journal of Machine Learning Research 22*, 1–25.

[2] Gao, C., Y. Lu, and H. H. Zhou (2015). Rate-optimal graphon estimation. *The Annals of Statistics 43*(6), 2624–2652.

[3] Klopp, O., A. B. Tsybakov, and N. Verzelen (2017). Oracle inequalities for network models and sparse graphon estimation. *The Annals of Statistics 45*(1), 316–354.

[4] Li, Y., D. Shah, D. Song, and C. L. Yu (2019). Nearest neighbors for matrix estimation interpreted as blind regression for latent variable model. *IEEE Transactions on Information Theory 66*(3), 1760–1784.

[5] Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. Springer Science & Business Media.

UNIVERSITY OF WISCONSIN - MADISON
Email address: `miaoyan.wang@wisc.edu`

REGRET MINIMIZATION AND STATISTICAL INFERENCE IN ONLINE DECISION MAKING WITH HIGH-DIMENSIONAL COVARIATES

DONG XIA

Classification AMS 2020:

Keywords: linear contextual bandit, sparsity, ε-greedy, inference, regret

We Investigate regret minimization, statistical inference, and their interplay in high-dimensional online decision-making based on the sparse linear contextual bandit model. We integrate the ε-greedy bandit algorithm for decision-making with a hard thresholding algorithm for estimating sparse bandit parameters and introduce an inference framework based on a debiasing method using inverse propensity weighting. Under a margin condition, our method achieves either $O(\sqrt{T})$ regret or classical $O(\sqrt{T})$-consistent inference, indicating an unavoidable trade-off between exploration and exploitation. If a diverse covariate condition holds, we demonstrate that a pure-greedy bandit algorithm—i.e., exploration-free—combined with a debiased estimator based on average weighting can simultaneously achieve optimal $O(\log T)$ regret and $O(\sqrt{T})$-consistent inference. We also show that a simple sample mean estimator can provide valid inference for the optimal policy's value. Numerical simulations and experiments on Warfarin dosing data validate the effectiveness of our methods.

To the best of our knowledge, this work is the first to investigate regret minimization, statistical inference, and their interplay in high-dimensional online decision-making based on the sparse-LCB model. Our contributions are summarized as follows:
General Inference Framework and Tradeoff with Regret. We propose a novel statistical inference framework for adaptively collected high-dimensional data. Our approach integrates the ε-greedy bandit algorithm with hard-thresholding (HT), resulting in a biased estimator due to the adaptive data collection and implicit regularization introduced by the HT algorithm. To mitigate this bias, we introduce an online debiasing technique based on IPW that maintains low computational and storage complexity. Under a margin condition with parameter ν, the debiased estimator is asymptotically normal, enabling the construction of confidence intervals and hypothesis tests for both individual arm parameters and their differences. Additionally, we identify a trade-off between regret performance and the estimator's asymptotic variance, which affects inference efficiency by determining the width of confidence intervals and the p-values of test statistics. Specifically, when the algorithm achieves a regret upper bound of $O(T^{1-\gamma} + T^{(\gamma-1)(1+\nu)/2})$ with margin parameter ν, and some user-specified $\gamma \in [0,1)$—which characterizes the exploration probability, the estimator's asymptotic variance is $O(T^{-(1-\gamma)})$. For example, when $\nu = 1$, setting $\gamma = \frac{1}{2} + o(1)$ yields a regret bound of $O(T^{1/2})$ and an estimator variance of $O(T^{-1/2})$, which does not attain the classic $\sqrt{T}$-consistency; setting $\gamma = 0$ yields a trivial regret bound of $O(T)$ and an

asymptotically normal estimator which is $\sqrt{T}$-consistent. While IPW is effective for debiasing, it unfortunately inflates the variance of the final estimator.

Simultaneous Optimal Inference and Regret. We demonstrate that optimal inference efficiency and regret performance can be simultaneously achieved under an additional *covariate diversity* assumption, commonly employed in high-dimensional bandit literature ([Bastani et al.(2021)Bastani, Bayati, and Khosravi, Ren and Zhou(2024)] and references therein). This assumption is motivated by the observation that when covariates are sufficiently diverse, an exploration-free algorithm (i.e., setting the exploration probability $\varepsilon = 0$ in the ε-greedy algorithm) can still adequately explore each arm. This automatic exploration facilitates debiasing through a simple average weighting (AW) approach, bypassing IPW and thereby avoiding variance inflation. Specifically, our approach achieves an optimal $O(\log T)$ regret upper bound, and the resulting estimators of arm parameters are asymptotically normal with a variance of $O(T^{-1})$, thereby attaining the classic $\sqrt{T}$-consistency and optimal inference efficiency. Additionally, we introduce an inference procedure for the optimal policy's value, often referred to as the Q-value, within this framework. We provide a straightforward method to assess the maximum total reward achievable by the optimal policy.

Empirical Result. We evaluate the empirical performance of our algorithm and inference framework through numerical simulations and a real-world data experiment. Specifically, we apply this framework to the aforementioned Warfarin dosing problem. Our approach identifies several significant variables that determine the appropriate dosage, with findings that are consistent with existing medical literature while also offering novel insights.

References

[Bastani et al.(2021)Bastani, Bayati, and Khosravi] Hamsa Bastani, Mohsen Bayati, and Khashayar Khosravi. Mostly exploration-free algorithms for contextual bandits. *Management Science*, 67(3):1329–1349, 2021.

[Ren and Zhou(2024)] Zhimei Ren and Zhengyuan Zhou. Dynamic batch learning in high-dimensional sparse linear contextual bandits. *Management Science*, 70(2):1315–1342, 2024.

Department of Mathematics, HKUST, Clear Water Bay, Sai Kung, Hong Kong SAR
Email address: madxia@ust.hk

COINTEGRATION BETWEEN TWO INTRINSICALLY STATIONARY SPATIAL PROCESSES

RONGMAO ZHANG, CHANGXIONG CHI, PINGFANG ZHU, YUANYUAN JI AND QIWEI YAO

Classification AMS 2023: 62M30, 62M40

Keywords: Cointegration tests, confidence intervals, intrinsic processes, intrinsic stationarity, spatial block bootstrap, spatial cointegration, wild bootstrap.

The concept of the intrinsic processes proposed by Matheron (1973) provides an elegant mathematical framework for modeling nonstationary spatial phenomena. It can be viewed as a direct analogue of taking differences of nonstationary time series in order to achieving stationarity. But it is applicable to spatial data observed on irregular grids. The goal of this paper is to establish the inference methods and the relevant theory for identifying the cointegration between two simple intrinsic processes. We apply the least squares estimation, similar to Engle and Granger (1987). However the asymptotic property of the estimation is much more complex, depending on the underlying processes as well as the manner in which the observations were taken. We propose some bootstrap approximations for the asymptotic distribution of the estimators. It turns out that the wild bootstrap procedure is adaptive automatically to varying convergence rates under the different schemes of taking the observations. Therefore it paves the way for constructing practically feasible confidence intervals for cointegration coefficients. A new and easy-to-use statistical tests is constructed for testing the cointegration. The proposed methods, as well as the associated asymptotic results under various settings, are illustrated in simulation. The application to a real data example is also reported.

1. Intrinsic Processes

Let $X(\mathbf{s})$ be a real-valued spatial process defined on $\mathbf{s} \equiv (u, v) \in \mathcal{S}$, where $\mathcal{S}$ is a subset of R^2.

Definition 1.1. *Let $k \geq 0$ be an integer. A finite vector $(\lambda_{\mathbf{s}_1}, \cdots, \lambda_{\mathbf{s}_m})$ is called a k-increment coefficient vector if*

$$\sum_{i=1}^{m} \lambda_{\mathbf{s}_i} u_i^{k_1} v_i^{k_2} = 0 \text{ \textit{for any integers} } k_1, k_2 \geq 0, \text{ and } k_1 + k_2 \leq k, \tag{1.1}$$

where $m \geq 2$ is an arbitrary integer, $\{\mathbf{s}_i = (u_i, v_i), i = 1, \cdots, m\} \subset \mathcal{S}$, and $\{\lambda_{\mathbf{s}_i}\}$ are real numbers. Furthermore, we call $Y(\mathbf{s}) \equiv \sum_i \lambda_{\mathbf{s}_i} X(\mathbf{s}_i + \mathbf{s})$, $\mathbf{s}_i + \mathbf{s} \in \mathcal{S}$ a k-increment process.

A k-increment coefficient vector is defined on a finite set of locations $\{\mathbf{s}_i\}$. But it is also a location-shift-invariant in the sense that for any $\mathbf{s} = (u, v)$, (1.1) implies

$$\sum_{i=1}^{m} \lambda_{\mathbf{s}_i} (u_i + u)^{k_1} (v_i + v)^{k_2} = 0 \text{ for any integers } k_1, k_2 \geq 0, \text{ and } k_1 + k_2 \leq k.$$

Furthermore, for any given location set $\{\mathbf{s}_i,\, i \in [m]\}$, its k-increment coefficient vectors form a linear subspace of R^m. For example, all the 2-increment coefficient vectors consists of the linear space spanned by the columns of matrix $\mathbf{I}_m - \mathbf{S}(\mathbf{S}^\top\mathbf{S})^-\mathbf{S}^\top$, where $\mathbf{I}_m$ denotes the $m \times m$ identity matrix, $\mathbf{S}$ is the $m \times 6$ matrix with $(1, u_i, v_i, u_i^2, u_iv_i, v_i^2)$ as its m-th row. Thus the 2-increment process $Y(\mathrm{s})$ filters out all the polynomial components upto the order 2 of $X(\mathbf{s})$. In general, a k-increment process filters out all the polynomial components upto the order k. Note that a 0-increment coefficient vector can be viewed a difference operator across space, and a 0-increment process can be viewed as a differenced process. In this spirit, a k-increment coefficient vector can be viewed as a $(k+1)$-th difference operator over space, and a k-increment process $Y(\mathrm{s})$ is resulted from differencing $\mathbf{X}(\mathrm{s})$ the $(k+1)$ times.

Now we are ready to introduce the concept of intrinsic processes. Recall that X is stationary (or, more precisely, weakly stationary), if $E\{X(\mathbf{s})^2\} < \infty$, and

$$EX(\mathbf{s}) \equiv \mu, \quad \mathrm{Cov}\{\mathrm{X}(\mathbf{s}+\mathbf{h}), \mathrm{X}(\mathbf{s})\} = \mathrm{K}(\mathbf{h}) \quad \text{for any } \mathbf{s},\ \mathbf{s}+\mathbf{h} \in \mathcal{S}, \tag{1.2}$$

where $K(\cdot)$ is the covariance function of the process.

Definition 1.2. *$X(\cdot)$ is called an intrinsic process of order k, denoted by $X \in \mathrm{IP(k)}$, if all its k-increment processes are stationary.*

Among intrinsic processes the IP(0) processes, which is also called intrinsic stationarity, play an important role in catering for the nonstarionary spatial features. For intrinsically stationary $X(\mathbf{s})$, it holds that

$$E\{X(\mathbf{s}+\mathbf{h}) - X(\mathbf{s})\} = \psi(\mathbf{h}) \quad \text{and} \quad \mathrm{Var}\{\mathrm{X}(\mathbf{s}+\mathbf{h}) - \mathrm{X}(\mathbf{s})\} = 2\nu(\mathbf{h}),$$

where function $\nu(\cdot)$ is called semi-variogram. With additional condition $\psi(\cdot) \equiv 0$, the spatial prediction under the framework of the ordinary kriging only depends on the semivariogram. For stationary $X(\cdot)$, it holds that

$$\nu(\mathbf{s}) = K(\mathbf{0}) - K(\mathbf{s}),$$

where $K(\cdot)$ is given in (1.2). Note that an I(1) time series is an IP(0) process defined on the one-dimensional integer grid.

2. COINTEGRATION MODEL

Let $X(\cdot)$ and $Y(\cdot)$ be two IP(0) processes defined in $\mathcal{S}$. We call that X and Y are *cointegrated* if a linear combination of X and Y is stationary, i.e.

$$Y(\mathbf{s}) = \alpha + \beta X(\mathbf{s}) + \varepsilon(\mathbf{s}), \tag{2.1}$$

where $\beta \neq 0, \alpha$ are constants, and $\varepsilon(\mathbf{s}), \mathrm{s} \in \mathcal{S}$, is stationary with mean zero.

Based on the observation $\{Y(\mathrm{s}_i), X(\mathrm{s}_i)\}$, $i = 1, \cdots, n$, we adopt the OLS method to estimate α and β:

$$\widehat{\alpha} = \bar{Y} - \widehat{\beta}\bar{X} \quad \text{and} \quad \widehat{\beta} = \sum_{i=1}^n \{Y(\mathrm{s}_i) - \bar{Y}\}\{X(\mathrm{s}_i) - \bar{X}\} \Big/ \sum_{i=1}^n \{X(\mathrm{s}_i) - \bar{X}\}^2, \tag{2.2}$$

where $\bar{Y} = n^{-1}\sum_i Y(\mathbf{s}_i)$ and $\bar{X} = n^{-1}\sum_i X(\mathbf{s}_i)$. Then it also holds that

$$\widehat{\alpha} - \alpha = \bar{\varepsilon} - \bar{X}(\widehat{\beta} - \beta) \quad \text{and} \quad \widehat{\beta} - \beta = \sum_{i=1}^n \{\varepsilon(\mathrm{s}_i) - \bar{\varepsilon}\}\{X(\mathrm{s}_i) - \bar{X}\} \Big/ \sum_{i=1}^n \{X(\mathrm{s}_i) - \bar{X}\}^2, \tag{2.3}$$

where $\bar{\varepsilon} = n^{-1} \sum_i \varepsilon(\mathbf{s}_i)$. Then the residuals are defined as

$$\widehat{\varepsilon}(\mathbf{s}_i) = Y(\mathbf{s}_i) - \widehat{\alpha} - \widehat{\beta} X(\mathbf{s}_i), \quad i \in [n]. \tag{2.4}$$

The cointegration is declared if this residual sequence behave like a stationary spatial process. To test this, we need to understand the behviour of the OLS estimators $\widehat{\beta}$ and $\widehat{\alpha}$, which is much more complex that that of nonstationary time series.

3. Asymptotic properties of OLS estimators $\widehat{\beta}$ and $\widehat{\alpha}$

Unlike the cointegration of the regularly sampled time series, there is no uniform asymptotic theory for the cointegration over space. We establish below a generic (and less explicit) limiting theorem for the estimators $\widehat{\beta}$ and $\widehat{\alpha}$ defined in (2.2). Nevertheless it shows that $\widehat{\alpha}$ always enjoys the standard root-n convergence rate. However the convergence rates of $\widehat{\beta}$ varies substantially, depending on how the observations were taken over the space. Indeed 'increasing domain' and 'fixed domain' samplings lead to different convergence rates for $\widehat{\beta}$. Furthermore the increasing speed in 'increasing domain' sampling also affects the convergence rate.

We always assume that the observations are taken at the n locations $\mathbf{s}_1, \cdots, \mathbf{s}_n \in \mathcal{S} \equiv \mathcal{S}_n$. Let $d_n^2 = \sum_{i=1}^n \mathrm{E}\{X(\mathbf{s}_i)^2\}$, $H_n(\mathbf{t}) = \sqrt{n} X([\mathbf{nt}])/d_n$ and $G_n(\mathbf{t}) = \frac{1}{\sqrt{n}} \sum_{\mathbf{s}_j \le [\mathbf{nt}]} \varepsilon(\mathbf{s}_j)$.

Theorem 3.1. (1) *If $EX(\mathbf{s}) = 0$ for all $\mathbf{s}$, and there exist two processes $H(\cdot), G(\cdot)$ on $[0,1]$ such that*

$$\Big(H_n(t), G_n(s), \frac{1}{d_n} \sum_{i=1}^n \{X(\mathbf{s}_i)\varepsilon(\mathbf{s}_i) - \mathrm{E}X(\mathbf{s}_i)\varepsilon(\mathbf{s}_i)\}\Big)$$
$$\Longrightarrow \; (H(t), G(s), \int_0^1 H(t-)\, dG(t)),$$

then

$$d_n(\widehat{\beta} - \beta) \xrightarrow{D} \frac{\int_0^1 H(t-) dG(t) - G(1) \int_0^1 H(t)\, dt}{\int_0^1 H^2(t) dt - (\int_0^1 H(t)\, dt)^2} \equiv \eta,$$

$$\sqrt{n}(\widehat{\alpha} - \alpha) \xrightarrow{D} G(1) - \Big(\int_0^1 H(t)\, dt\Big)\eta.$$

(2) *Let* $S_n = \Big(\frac{1}{d_n^2} \sum_{i=1}^n X^2(\mathbf{s}_i), \frac{1}{d_n} \sum_{i=1}^n (X(\mathbf{s}_i) - \bar{X})\varepsilon(\mathbf{s}_i)\Big)$*. If $EX(\mathbf{s}) \neq 0$, and*

$$\Big(S_n, H_n(1), G_n(1)\Big) \xrightarrow{D} (c_0, V, H(1), G(1))$$

for some constant $c_0 > 0$ and random variable V, then

$$d_n(\widehat{\beta} - \beta) \xrightarrow{D} V/c_0, \quad \text{and}$$

$$\sqrt{n}(\widehat{\alpha} - \alpha) \xrightarrow{D} G(1) - H(1)V/c_0.$$

Note. blue $d_n = \big(\sum_{i=1}^n \mathrm{E}\{X(\mathbf{s}_i)\}^2\big)^{1/2}$ depends on $X(\cdot)$ only.

4. Wild bootstrap adaptive to unknown convergence rates of β

Assumption: $X(\cdot)$ and $\varepsilon(\cdot)$ are independent.

Wild Bootstrap Algorithm:

Step 1. Compute residuals $\tilde{\varepsilon}(\mathbf{s}_i) = Y(\mathbf{s}_i) - \widehat{\alpha} - \widehat{\beta}X(\mathbf{s}_i)$, and $\widehat{\varepsilon}(\mathbf{s}_i) = \tilde{\varepsilon}(\mathbf{s}_i) - n^{-1}\sum_{1\le j\le n}\tilde{\varepsilon}(\mathbf{s}_j)$, $i \in [n]$.

Step 2. Set $\widehat{Y}^w(\mathbf{s}_i) = \widehat{\alpha} + \widehat{\beta}X(\mathbf{s}_i) + \delta_i\widehat{\varepsilon}(\mathbf{s}_i)$, $i \in [n]$, where δ_i are i.i.d. with mean 0 and variance $\sigma_n^2 = 1 + 2n^{-1}\sum_{1\le i<j\le n}\mathrm{Corr}\{\varepsilon(\mathrm{s_i}), \varepsilon(\mathrm{s_j})\}$.

Step 3. Let
$$\widehat{\alpha}^w = \bar{Y}^w - \widehat{\beta}^w\bar{X}, \quad \widehat{\beta}^w = \frac{\sum_{i=1}^n\{Y^w(\mathbf{s}_i) - \bar{Y}^w\}\{X(\mathbf{s}_i) - \bar{X}\}}{\sum_{i=1}^n\{X(\mathbf{s}_i) - \bar{X}\}^2}.$$

Then as $n \to \infty$,
$$\mathcal{L}\{d_n(\widehat{\beta}^w - \widehat{\beta})\} \asymp \mathcal{L}\{d_n(\widehat{\beta} - \beta)\}, \quad \mathcal{L}\{\sqrt{n}(\widehat{\alpha}^w - \widehat{\alpha})\} \asymp \mathcal{L}\{\sqrt{n}(\widehat{\alpha} - \alpha)\}.$$

Repeat Steps 2-3 above M times, leading to $(\widehat{\alpha}_i^w, \widehat{\beta}_i^w),\ i \in [M]$.

Hence an approximate $(1-\pi)$ confidence interval for β can be taken as $(2\widehat{\beta} - \widehat{\beta}^w_{1-\pi/2},\ \ 2\widehat{\beta} - \widehat{\beta}^w_{\pi/2})$, where $\widehat{\beta}^w_\pi$ denotes the $[\pi M]$-th smallest value among $\widehat{\beta}^w_1, \cdots, \widehat{\beta}^w_M$.

References

[1] Engle, R.F. and Granger, C.W.J. (1987). Co-integration and error correction: representation, estimation, and testing. *Econometrica*, 55, 251-276.

[2] Matheron, G. (1973). The intrinsic random functions and their applications. *Advances in Applied Probability*, 5, 439-68.

Qiwei Yao, London School of Economics and Political Science, Houghton Street, London WC2A 2AE, United Kingdom

Email address: Q.Yao@lse.ac.uk

REVISIT TENSOR DECOMPOSITION: STATISTICAL OPTIMALITY AND COMPUTATIONAL GUARANTEES

ANRU ZHANG

Classification AMS 2020: 62H12, 62F30, 62H25, 28Q25

Keywords: Tucker decomposition, CP decomposition, High-dimensional statistics, Signal-to-noise ratio, Statistical inference, Minimax optimality, Computational complexity, Alternating least squares, Low-rank tensors, Tensor PCA, Statistical-computational trade-off, Robust initialization

Tensor decomposition is a foundational tool in modern data analysis, enabling the extraction of structured, low-dimensional representations from high-dimensional, multi-way data. In this talk, we revisit two of the most widely used tensor decomposition frameworks, Tucker decomposition and Canonical Polyadic (CP) decomposition, through the lens of statistical optimality and computational guarantees. Our focus is on both the fundamental limits and practical algorithms for reliable use of tensor methods in noisy, high-dimensional settings.

We begin with Tucker decomposition, which models a low-rank tensor through multilinear projections along each mode [1]. This approach is particularly suited for applications in computational imaging and social sciences, where data are high-order and often corrupted by noise. We analyze the Tucker model in the presence of additive Gaussian noise, where the underlying signal tensor exhibits low multilinear rank. Our results characterize the three-phase behavior of statistical estimation under varying signal-to-noise ratios (SNR): (i) in the strong SNR regime, the Higher-Order Orthogonal Iteration (HOOI) algorithm achieves minimax-optimal rates for estimating the singular subspaces and the tensor itself; (ii) in the weak SNR regime, no consistent estimator exists; and (iii) in the moderate SNR regime, a statistical-computational gap emerges—consistent estimation is possible in theory but computationally intractable under standard complexity assumptions.

We further explore inference procedures in Tucker decomposition [2]. Building on recent developments, we establish asymptotic distributions for singular subspace estimators derived from alternating minimization, allowing for the construction of confidence regions. Importantly, unlike matrix-based settings where debiasing is often necessary, our results show that no debiasing is required for valid inference in tensor models—underlining a key distinction introduced by the multilinear structure and the tensor-specific computational landscape.

Next, we turn to CP decomposition, where a tensor is represented as a sum of rank-one components [3]. Despite its wide empirical use, the theoretical understanding of CP decomposition, especially under noise, non-orthogonality, and higher-rank scenarios, has remained limited. We address this gap by analyzing the Alternating Least Squares (ALS) algorithm in a signal-plus-noise model. We show that ALS, when properly initialized, achieves non-asymptotic, minimax-optimal error bounds for tensors of arbitrary order, dimension, and rank. We propose a robust initialization method—Tucker-based Approximation with Simultaneous Diagonalization (TASD)—which compresses the tensor and stabilizes subsequent optimization. When used with ALS, the resulting estimator (TASD-ALS) is both statistically consistent and computationally efficient, achieving optimal estimation rates in practice.

Additionally, we provide a rigorous convergence analysis of ALS. We prove that in the rank-one setting, ALS achieves optimal error bounds in just one or two iterations. For general rank, we uncover a two-phase convergence pattern: an initial quadratic phase followed by a linear refinement, with rates determined by coherence properties of the underlying components. These findings give the first formal justification of the fast empirical convergence observed for ALS in structured tensor settings.

In summary, this talk bridges a significant gap between statistical theory and algorithmic practice in tensor decomposition. Our results provide sharp insights into the limits of estimation and inference, while offering provably effective algorithms that scale to modern high-dimensional, multi-modal data.

References

[1] Zhang, A., & Xia, D. (2018). Tensor SVD: Statistical and computational limits. *IEEE Transactions on Information Theory*, 64(11), 7311-7338.

[2] Xia, D., Zhang, A. R., & Zhou, Y. (2022). Inference for low-rank tensors—no need to debias. *The Annals of Statistics*, 50(2), 1220-1245.

[3] Tang, R., Chhor, J., Klopp, O., & Zhang, A. R. (2025). Revisit CP Tensor Decomposition: Statistical Optimality and Fast Convergence. *arXiv preprint arXiv:2505.23046*.

Department of Biostatistics & Bioinformatics and Department of Computer Science, Duke University, 2424 Erwin Road, Durham, NC 27710

Email address: anru.zhang@duke.edu

MODELING NON-UNIFORM HYPERGRAPHS USING DETERMINANTAL POINT PROCESSES

EMMA JINGFEI ZHANG

Classification AMS 2020: 62A09, 91D30

Keywords: hypergraphs; higher-order interactions; determinantal point process.

Existing network analyses have primarily focused on pairwise interactions, where each edge in the network consists of two nodes. However, higher-order interactions, which involve multiple nodes simultaneously, are ubiquitous across many real-world scenarios. For example, in academic collaborations, researchers often co-author papers in teams of three or more. Existing studies have demonstrated the importance of higher-order interactions in contexts such as neural systems [3], genetic networks [2], and the spread of epidemics [1].

High-order interactions among a set of nodes can be naturally represented using *hypergraphs*, a generalization of traditional graphs where an edge, known as a *hyperedge*, is a set that includes all interacting nodes. In our work, we model the observed hypergraph as a collection of independent realizations from a random set distribution $\mathcal{P}$, where each hyperedge corresponds to a node subset.

Let $V = [n]$ denote the set of n nodes, and $D = \{e_1, e_2, \ldots, e_m\}$ the set of m hyperedges. Each hyperedge can be represented as a subset of V, that is, $e_s \subset [n]$ for $s \in [m]$. We assume that e_s for $s \in [m]$ follows

$$e_s \overset{\text{i.i.d.}}{\sim} pr_L,$$

where L is a kernel matrix and pr_L is as defined as

$$pr_L(E = e) = \frac{\det(L_e)}{\det(L + I)}. \tag{0.1}$$

The distribution in (0.1) is referred to as a determinantal point process (DPP) and the matrix L is referred to as a *kernel matrix*. Modeling hyperedges using determinantal point processes has several benefits. First, it naturally accommodates non-uniform hyperedges and multi-hyperedges, which greatly extends model flexibility. Second, as the normalizing constant can be easily derived, pr_L defines a tractable distribution over all possible hyperedges, facilitating estimation, inference, and sampling. Third, the kernel matrices L enhances model interpretability.

We model L as the sum of a symmetric matrix and a skew symmetric matrix, and estimate model parameters using projected gradient ascent over the log-likelihood, subject to constraints. In theory, we establish that under mild regularity conditions, the maximum likelihood estimator (MLE) is consistent and asymptotically normal. The proofs are nontrivial considering the special manifold of the parameter space which arises from the model configuration.

We demonstrate the effectiveness and flexibility of the proposed non-symmetric DPP (NDPP) hypergraph model in simulations and real data analysis. Simulation results show that NDPP achieves accurate parameter estimation, with estimation errors decreasing as the number of hyperedges increases, validating our theoretical results. Compared to the symmetric DPP model, NDPP shows superior performance when hyperedges are generated under its own model, and comparable performance under the DPP model. In real data analyses across four hypergraphs, including contact-high-school, email-Eu, NDC-substances, and tags-math-sx, NDPP consistently outperforms DPP in hyperedge prediction tasks (AUC and MPR) in cases where node similarity is more plausible, while performing comparably in settings favoring node diversity. These results show NDPP's flexibility in modeling both similarity and diversity among nodes in non-uniform hypergraphs.

This is a joint work with Yichao Chen (University of Michigan) and Ji Zhu (University of Michigan).

References

[1] Federico Battiston, Enrico Amico, Alain Barrat, Ginestra Bianconi, Guilherme Ferraz de Arruda, Benedetta Franceschiello, Iacopo Iacopini, Sonia Kéfi, Vito Latora, Yamir Moreno, et al. The physics of higher-order interactions in complex systems. *Nature Physics*, 17(10):1093–1098, 2021.

[2] Anna Ritz, Allison N Tegge, Hyunju Kim, Christopher L Poirel, and TM Murali. Signaling hypergraphs. *Trends in biotechnology*, 32(7):356–362, 2014.

[3] Shan Yu, Hongdian Yang, Hiroyuki Nakahara, Gustavo S Santos, Danko Nikolić, and Dietmar Plenz. Higher-order interactions characterized in cortical activity. *Journal of neuroscience*, 31(48):17514–17526, 2011.

Goizueta Business School, Emory University, Atlanta, Georgia, U.S.A.
Email address: emma.zhang@emory.edu

A DYNAMIC NETWORK AUTOREGRESSIVE MODEL FOR TIME-VARYING NETWORK-LINK DATA

JINGNAN ZHANG

Classification AMS 2025: Frontiers of Statistical Network Analysis: Inference, Tensors and Beyond

Keywords: Global homogeneity, group structure, network-linked data, dynamic network.

Network-linked data, which refers to a group of units that are observed connected by a network and have a set of available attributes, has attracted much attention in the past few decades (Michell and West, 1996; Lee et al., 2010; Li et al., 2019, 2023; Huang et al., 2021). Yet its extension, time-varying network-link data, has received less investigation.

Existing methods for time-varying network-link data usually assume that units' attributes evolve over time, whereas the network remains unchanged as time increases. (Zhu et al., 2017; Wu, 2019; Zhu and Pan, 2020; Zhu et al., 2022; Chen et al., 2023; Zhu et al., 2023; Li et al., 2023). Zhu et al. (2017) firstly proposed a network vector autoregressive model (NAR) to incorporate network structure. Specifically, they assume that

$$Y_{it} = \mu + \boldsymbol{X}_i^\top \boldsymbol{\gamma} + \eta_0 n_i^{-1} \sum_{j=1}^{N} a_{ij} Y_{j(t-1)} + \eta_1 Y_{i(t-1)} + \epsilon_{it}, \tag{0.1}$$

where $\boldsymbol{Y}_t = (Y_{1t}, \dots, Y_{Nt})^\top$ is the high-dimensional response vector with N being the number of nodes in network $\mathcal{G}$, the node-specific covariate vector $\boldsymbol{X}_i$ is independent and identically distributed random, $\boldsymbol{A} = (a_{ij})_{i,j=1}^N \in \{0,1\}^{N \times N}$ is the adjacent matrix of $\mathcal{G}$ with $a_{ij} = 1$ if there exists an edge between nodes i and j and $a_{ij} = 0$ otherwise, $n_i = \sum_{j=1}^n a_{ij}$ is the degree of node i, and $(\mu, \boldsymbol{\gamma}, \eta_0, \eta_1)$ are parameters to be estimated. Since Zhu et al. (2017), many extensions of NAR model have been studied. For example, Wu (2019) extends model (0.1) to a time-varying setting by allowing $(\mu, \boldsymbol{\gamma}, \eta_0, \eta_1)$ to change with t. Zhu et al. (2022) further extends to the functional varying coefficient setting. Another extension route is to assume that there exists some group structure among N nodes to capture the heterogeneity of nodes (Zhu and Pan, 2020; Chen et al., 2023; Zhu et al., 2023), which means that parameters are the same within each group but different across different groups. Li et al. (2023) studied a grouped time-varying NAR model by assuming the time-varying functional coefficients share some group structure.

However, all aforementioned methods have the following drawbacks. The first one is that network $\mathcal{G}$ does not change over time. In real world, edges among nodes usually change frequently and drastically as time increases (Matias and Miele, 2017; Liu et al., 2018). The second one is that they assume that $\boldsymbol{A}$ is deterministic. In practice, it is well known that network data are collected with errors (Le and Li, 2022). It is common to assume that network data are generated by some parametric model, such as the Erdős Rényi model (Erdős et al., 1960) and the stochastic block model (Holland et al., 1983).

The third one is that the heterogeneity captured by group structure is not sufficient. Nodes within the same group should still behave differently, which corresponds to the degree-corrected stochastic block model (Karrer and Newman, 2011). Besides, as argued by Li et al. (2019) and Le and Li (2022), the parametric form of autoregressive neighborhood average in model (0.1) may be inappropriate to model the network effect to network-linked data, and thus leads to unsatisfactory performance. The last one is the assumption that $\boldsymbol{X}_{it}$'s are identically and independently distributed across $1 \leq i \leq N$ and $1 \leq t \leq T$ may be inappropriate for real data.

In this paper, we propose a novel dynamic network autoregressive model to tackle the above problems for time-varying network-linked data. Specifically, we consider that networks are also evolving as time changes. Then, the dynamic networks are modeled with a tensor CANDECOMP/PARAFAC(CP) decomposition method (Kolda and Bader, 2009), where node and time features of networks are captured by some embedding vectors in low-dimensional Euclidean space. By assuming node-embedding vectors concentrate around some centers, we allow heterogeneity for nodes within the same group. Next, we reformulate the NAR model (0.1) with the help of node and time-embedding vectors. Nodes with similar embedding vectors will have similar contributions to the response variable Y_{it}. Moreover, we consider a flexible framework for the effect of covariate vector $\boldsymbol{X}_{it}$, where both within-group and global homogeneities are allowed. We also allow non-random covariate vector $\boldsymbol{X}_{it}$.

The main contribution of the proposed model is the development of a novel framework to model time-varying network-linked data, which mainly integrates a tensor decomposition method and the NAR model (0.1). Instead of considering a deterministic network without changing over time, we model dynamic networks via tensor decomposition. To the best of our knowledge, this is the first attempt to consider dynamic networks for network-linked data. More importantly, we propose a new dynamic network autoregressive model, which incorporates node-embedding and time-embedding vectors as dynamic network impact factors. It is more natural than the neighborhood average effect in literature. Node-embedding and time-embedding vectors and the group structure are estimated using the tensor power update algorithm (Zhang et al., 2023). To solve the resultant optimization task for the dynamic network autoregressive model, we employ a group lasso-type penalty and develop an efficient alternative update algorithm. Further, we establish the asymptotic consistencies for the proposed method whether the global effect of covariate vector exists or not. The superior numerical performance of the proposed method is supported by extensive simulated examples and a real application on time-varying network-linked fund data.

References

Chen, E. Y., Fan, J., and Zhu, X. (2023). Community network auto-regression for high-dimensional time series. *Journal of Econometrics*, 235(2):1239–1256.

Erdős, P., Rényi, A., et al. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60.

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137.

Huang, D., Zhu, X., Li, R., and Wang, H. (2021). Feature screening for network autoregression model. *Statistica Sinica*, 31:1239.

Karrer, B. and Newman, M. E. (2011). Stochastic blockmodels and community structure in networks. *Physical review E*, 83(1):016107.

Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM review*, 51(3):455–500.

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: Statistical Methodology*, 84(5):1851–1885.

Lee, L.-f., Liu, X., and Lin, X. (2010). Specification and estimation of social interaction models with network structures. *The Econometrics Journal*, 13(2):145–176.

Li, D., Peng, B., Tang, S., and Wu, W. (2023). Inference of grouped time-varying network vector autoregression models. *arXiv preprint arXiv:2303.10117*.

Li, T., Levina, E., and Zhu, J. (2019). Prediction models for network-linked data. *The Annals of Applied Statistics*, 13(1):132–164.

Liu, F., Choi, D., Xie, L., and Roeder, K. (2018). Global spectral clustering in dynamic networks. *Proceedings of the National Academy of Sciences*, 115(5):927–932.

Matias, C. and Miele, V. (2017). Statistical clustering of temporal networks through a dynamic stochastic block model. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(4):1119–1141.

Michell, L. and West, P. (1996). Peer pressure to smoke: the meaning depends on the method. *Health education research*, 11(1):39–49.

Wu, B. (2019). Time-varying network vector autoregression model.

Zhang, Y., Zhang, J., Sun, Y., and Wang, J. (2023). Change point detection in dynamic networks via regularized tensor decomposition. *Journal of Computational and Graphical Statistics*, (just-accepted):1–22.

Zhu, X., Cai, Z., and Ma, Y. (2022). Network functional varying coefficient model. *Journal of the American Statistical Association*, 117(540):2074–2085.

Zhu, X. and Pan, R. (2020). Grouped network vector autoregression. *Statistica Sinica*, 30(3):1437–1462.

Zhu, X., Pan, R., Li, G., Liu, Y., and Wang, H. (2017). Necwork vector autoregression. *The Annals of Statistics*, 45(3):1096–1123.

Zhu, X., Xu, G., and Fan, J. (2023). Simultaneous estimation and group identification for network vector autoregressive model with heterogeneous nodes. *Journal of Econometrics*, page 105564.

International Institute of Finance, School of Management, University of Science and Technology of China, Hefei, Anhui Province, 230026.

Email address: jnzhang@ustc.edu.cn

HIGHER-ORDER ACCURATE TWO-SAMPLE NETWORK INFERENCE AND NETWORK HASHING: A SCIENTIFIC REPORT

YUAN ZHANG

1. Introduction and Problem Statement

Two-sample hypothesis testing for network comparison is a fundamental statistical problem with applications in neuroscience, genomics, and social network analysis. The central challenge involves determining whether two collections of network observations arise from the same distribution or represent distinct populations with different structural characteristics.

Traditional network comparison methods face several limitations: dependence on strong distributional assumptions, lack of finite-sample guarantees, inability to handle heterogeneous network properties, and computational inefficiency for large-scale applications. The paper addresses these challenges through a unified framework that handles repeated network observations, accommodates unknown node correspondence, achieves higher-order finite-sample accuracy, and provides fast computation with theoretical guarantees.

2. Methodology and Theoretical Contributions

2.1. **Core Framework.**

The proposed method introduces a comprehensive statistical toolbox with the following key features:

Unified Test Statistic: The methodology establishes a test statistic based on network structural features that maintains consistent performance across different data configurations, adapting seamlessly to scenarios with or without node matching information.

Higher-Order Finite-Sample Accuracy: Unlike traditional approaches focusing on asymptotic properties, this method achieves higher-order accuracy in finite samples through carefully designed bias correction and variance estimation techniques.

Adaptive Design: The method automatically adapts to different network characteristics without requiring users to pre-specify sparsity or scale parameters, enhancing practical usability across diverse applications.

2.2. **Network Hashing Framework.**

A significant innovation is the development of a network hashing framework for large-scale network databases:

Hash Algorithm: The algorithm compresses network structural information into fixed-length hash codes while preserving distance relationships between networks, enabling efficient similarity-based retrieval.

Fast Querying: The framework enables rapid similarity queries in large-scale databases, transforming query complexity from linear to near-constant time with important implications for network data mining and pattern recognition.

2.3. **Theoretical Guarantees.** The paper provides rigorous theoretical analysis with several key results:

Power Optimality: The method achieves power optimality under the minimax framework, attaining the highest possible detection probability given error rate constraints.

Finite-Sample Theory: Comprehensive analysis includes exact distributions of test statistics, non-asymptotic error rate bounds, and precise characterization of power functions.

Minimax Optimality: The proposed tests achieve minimax optimal rates for network comparison problems, establishing fundamental limits for this class of inference problems.

3. Experimental Validation

3.1. **Simulation Studies.** Extensive simulations validate the methodology across diverse network models including random graphs, small-world networks, scale-free networks, and stochastic block models. Performance benchmarking reveals significant advantages in both accuracy and computational speed compared to existing methods. The finite-sample theoretical guarantees are confirmed through simulation studies across realistic sample size regimes.

3.2. **Real Data Applications.** The method demonstrates practical utility through two real-world applications:

Brain Network Analysis: Application to brain connectivity data successfully identifies differences in network connection patterns between populations, showcasing potential in neuroscience research.

Social Network Analysis: Analysis of social network data validates utility in social science research, providing tools for understanding network evolution and structural characteristics.

Both applications reveal previously unidentified network structures, demonstrating the method's capability to uncover subtle but meaningful differences.

3.3. **Computational Performance.** Experimental results show substantial improvements in computational efficiency and scalability. The hashing framework enables efficient memory utilization, making the approach practical for large network databases that might otherwise exceed memory constraints.

4. Innovations and Impact

4.1. **Key Innovations. Theoretical Advances**: The unified framework handles scenarios with and without node correspondence, achieving finite-sample higher-order accuracy and establishing minimax optimality.

Methodological Contributions: Adaptive algorithm design reduces parameter tuning burden while the network hashing mechanism provides novel solutions for large-scale network data management.

Practical Benefits: Significant computational improvements, memory efficiency, and user-friendly implementation facilitate adoption across research communities.

4.2. **Applications and Future Directions.** The research has broad applications in biomedical research (brain networks, gene regulatory networks), social sciences (network evolution analysis), and engineering (communication networks, recommendation systems). Future extensions include dynamic network analysis, multilayer network handling, and integration with machine learning approaches.

5. Conclusions

The work by Shao et al. represents a comprehensive solution to fundamental challenges in network comparison through two-sample hypothesis testing. Key contributions include:

- Establishment of unified testing frameworks with finite-sample higher-order accuracy
- Proof of minimax optimality and comprehensive theoretical guarantees
- Development of innovative network hashing for large-scale applications
- Substantial improvements in computational efficiency while maintaining statistical optimality
- Demonstration of practical utility across diverse real-world applications

The comprehensive nature of this work—combining rigorous theory, practical methodology, and extensive validation—establishes it as a significant contribution to network statistics with lasting impact on both theoretical understanding and practical applications. As network data becomes increasingly prevalent across scientific disciplines, this research provides powerful analytical tools that will likely inspire subsequent developments in network statistical methodology.

Ohio State University, USA
Email address: yzhanghf@stat.osu.edu

FROM SCORE ESTIMATION TO SAMPLING

HARRISON ZHOU

Recent impressive advances in the algorithmic generation of high-fidelity images, audio, and video can be largely attributed to the success of score-based diffusion models. A crucial step in their implementation is score matching, which involves estimating the score function of the forward diffusion process from training data. In this work, we establish the rate-optimal estimation of the score function for smooth, compactly supported densities and explore its applications to estimation of density, transport, and optimal transport.

Yale University, USA
Email address: harrison.zhou@yale.edu

NONPARAMETRIC INFERENCE ON NETWORK EFFECTS WITH DEPENDENT EDGES: OPTIMALITY, TWO-SAMPLE, MULTIPLE STRATA

WEN ZHOU

Testing network effects in weighted directed networks is a foundational problem in econometrics, sociology, and psychology. Yet, the prevalent edge dependency poses a significant methodological challenge. Most existing methods are model-based and come with stringent assumptions, limiting their applicability. In response, we introduce a novel, fully nonparametric framework that requires only minimal regularity assumptions. While inspired by recent developments in U-statistic literature, our approach notably broadens their scopes. Specifically, we identified and carefully addressed the challenge of indeterminate degeneracy in the test statistics – a problem that aforementioned tools do not handle. We established Berry-Esseen type bounds for the accuracy of type-I error rate control. With original analysis, we also proved the minimax power optimality of our test. Simulations underscore the superiority of our method in computation speed, accuracy, and numerical robustness compared to competing methods.

New York University, USA
Email address: w.zhou@nyu.edu

HYPERBOLIC NETWORK LATENT SPACE MODEL WITH LEARNABLE CURVATURE

JI ZHU

Network data is ubiquitous in various scientific disciplines, including sociology, economics, and neuroscience. Latent space models are often employed in network data analysis, but the geometric effect of latent space curvature remains a significant, unresolved issue. In this work, we propose a hyperbolic network latent space model with a learnable curvature parameter. We theoretically justify that learning the optimal curvature is essential to minimizing the embedding error across all hyperbolic embedding methods beyond network latent space models. A maximum-likelihood estimation strategy, employing manifold gradient optimization, is developed, and we establish the consistency and convergence rates for the maximum-likelihood estimators, both of which are technically challenging due to the non-linearity and non-convexity of the hyperbolic distance metric. We further demonstrate the geometric effect of latent space curvature and the superior performance of the proposed model through extensive simulation studies and an application using a Facebook friendship network.

UNIVERSITY OF MICHIGAN, USA
Email address: jizhu@umich.edu