Adjusted Chi-square Test for Degree-Corrected Block Models

Arash Amini
Department of Statistics
University of California at Los Angeles

We propose a general goodness-of-fit test for degree-corrected stochastic block models (DCSBM). The test is based on an adjusted chi-square statistic for measuring equality of means among sub-groups of n multinomial distributions with $d_1, \dots, d_n$ observations. In the context of network models, the number of multinomials, n, grows much faster than the number of observations, d_i, hence the setting deviates from classical asymptotics. We show that a simple adjustment allows the statistic to converge in distribution, under null, as long as the harmonic mean of $\{d_i\}$ grows to infinity. This result applies to large sparse networks where the role of d_i is played by the degree of node i. Our distributional results are nonasymptotic, with explicit constants, providing finite-sample bounds on the Kolomogorv-Smirnov distance to the target distribution. When applied sequentially, the test can also be used to determine the number of communities. The test operates on a (row) compressed version of the adjacency matrix, conditional on the degrees, and as a result is highly scalable to large sparse networks. We incorporate a novel idea of compressing the columns based on a $(K+1)$-community assignment when testing for K communities. This approach increases the power in sequential applications without sacrificing computational efficiency, and we prove its consistency in recovering the number of communities. Since the test statistic does not rely on a specific alternative, its utility goes beyond sequential testing and can be used to simultaneously test against a wide range of alternatives outside the DCSBM family. The test can also be easily applied to Poisson count arrays in clustering or biclustering applications, as well as bipartite and directed networks. We show the effectiveness of the approach by extensive numerical experiments with simulated and real data. In particular, applying the test to the Facebook-100 dataset, a collection of one hundred social networks, we find that a DCSBM with a small number of communities (say < 25) is far from a good fit in almost all cases. Despite the lack of fit, we show that the statistic itself can be used as an effective tool for exploring community structure, allowing us to construct a community profile for each network. This is joint work with Linfan Zhang.

Prevalence of Neural Collapse During the Terminal Phase of Deep Learning Training

David Donoho
Department of Statistics
Stanford University

Modern deep neural networks for image classification have achieved superhuman performance. Yet, the complex details of trained networks have forced most practitioners and researchers to regard them as blackboxes with little that could be understood. This paper considers in detail a now-standard training methodology: driving the cross-entropy loss to zero, continuing long after the classification error is already zero. Applying this methodology to an authoritative collection of standard deepnets and datasets, we observe the emergence of a simple and highly symmetric geometry of the deepnet features and of the deepnet classifier; and we document important benefits that the geometry conveys – thereby helping us understand an important component of the modern deep learning training paradigm. This is joint work with Vardan Papyan, U Toronto and XY Han, Cornell. It covers a paper which appeared in September 2021 in Proc Natl Acad Sci.

Community Network Autoregression

Jianqing Fan
Department of Operations Research and Financial Engineering
Princeton University

Network data is prevalent in many contemporary big data applications in which the nodes can be broadly defined such as individuals, economic entities, documents, or medical disorders in social, economic, text, or health networks. Yet, a simple question how to use community structures in high-dimensional vector autoregression models remains largely unexplored. To answer the question, we introduce a novel network vector autoregressive model to reduce the number of coefficients in high-dimensional vector autoregression. The model utilizes the network structure to characterize the dependence and intra-community homogeneity of the high dimensional time series and include the non-community-related latent factors to account for unknown cross-sectional dependence. We propose new econometrics methods, leveraging on the community structure, and establish related asymptotic inference tools. The advantages of the CNAR model are nicely illustrated by synthetic and real datasets. This is joint joint work with Elynn Chen and Xuening Zhu.

Universal Rank Inference via Residual Subsampling with Application to Large Networks

Yingying Fan
Department of Operations Management and Data Science
University of Southern California

Determining the precise rank is an important problem in many large-scale applications with matrix data exploiting low-rank plus noise models. In this paper, we suggest a universal approach to rank inference via residual subsampling (RIRS) for testing and estimating rank in a wide family of models, including many popularly used network models such as the degree corrected mixed membership model as a special case. Our procedure constructs a test statistic via subsampling entries of the residual matrix after extracting the spiked components. The test statistic converges in distribution to the standard normal under the null hypothesis, and diverges to infinity with asymptotic probability one under the alternative hypothesis. The effectiveness of RIRS procedure is justified theoretically, utilizing the asymptotic expansions of eigenvectors and eigenvalues for large random matrices recently developed in Fan, Fan, Han and Lv (2019a, b). The advantages of the newly suggested procedure are demonstrated through several simulation and real data examples. This work is joint with Xiao Han and Qing Yang.

Counting Cycles in Networks

Zheng Tracy Ke
Department of Statistics
Harvard University

In many network models, the quantity of interest (community structure, mixed-memberships) is a low-rank signal matrix, masked by noise. The spectrum of the signal matrix plays a fundamental role in network analysis and is of major interest. We propose to recover the spectrum by counting short cycles in the adjacency matrix. The cycle counts provide a good estimate for the moments of the spectrum, which can thus be used to estimate the spectrum. Compared to empirical spectrum, the proposed estimators are more accurate in a wide range of parameter settings. The idea can also be adapted to solve many other problems. One of such problems is global testing, where the goal is to test whether the network only has one communities or multiple communities. We find that counting cycles with a centered adjacency matrix gives rise to an easy-to-use testing statistic that is asymptotically $N(0, 1)$ under null and achieves the optimal phase transition. The test is competitive in a wide range of network settings, where we allow severe degree heterogeneity and mixed-memberships.

Data Splitting for Graphical Model Selection With FDR Control

Jun S. Liu
Department of Statistics
Harvard University

Simultaneously finding multiple influential variables and controlling the false discovery rate (FDR) for statistical and machine learning models is a fundamental problem with a long history. We here examine how the simple old idea of data splitting (DS) can be leveraged for controlling FDRs in learning graphical models. As one may have anticipated, a DS procedure simply estimates two nearly independent graphical models from two datasets of the half-size created by random splitting, and constructs a contrast statistic. The FDR control can be achieved by taking advantage of the statistic's property that, for any unconnected pair of nodes, its statistic's sampling distribution is symmetric about 0. Furthermore, by repeated sample splitting, we propose Multiple Data Splitting (MDS) to stabilize the selection result and boost the power. Interestingly, MDS not only helps overcome the power loss caused by data splitting with the FDR still under control, but also results in a lower variance for the estimated FDR compared with all other considered methods. DS and MDS are straightforward conceptually, easy to implement algorithmically, and efficient computationally. Simulation results show that both DS and MDS control the FDR well and have very competitive power. They are particularly powerful especially when the signals are weak. Our preliminary tests on generalized linear models and neural networks also also show promises. The presentation is based on joint work with Chenguang Dai, Buyu Lin, and Xin Xing.

Global and Individualized Community Detection in Inhomogeneous Multilayer Networks

Zongming Ma
Department of Statistics
University of Pennsylvania

In this talk, we discuss community detection in a stylized yet informative inhomogeneous multilayer network model. In our model, layers are generated by different stochastic block models, the community structures of which are (random) perturbations of a common global structure while the connecting probabilities in different layers are not related. Focusing on the symmetric two block case, we establish minimax rates for both global estimation of the common structure and individualized estimation of layer-wise community structures. Both minimax rates have sharp exponents. In addition, we provide an efficient algorithm that is simultaneously asymptotic minimax optimal for both estimation tasks under mild conditions. The optimal rates depend on the parity of the number of most informative layers, a phenomenon that is caused by inhomogeneity across layers. This talk is based on joint work with Shuxiao Chen and Sifan Liu.

Bringing Perspective into Dynamic Network Analysis

Nynke Niezink
Department of Statistics & Data Science
Carnegie Mellon University.

Many studies have shown that individuals differ in how they perceive and cognitively represent the networks they are embedded in. However, in the analysis of network dynamics, this is not taken into account. Instead, current models assume individuals to make network decisions, creating and dissolving ties, based on a shared network representation. But what if we choose to abandon the assumption of such a shared representation? In the talk, we will explore this question and reflect on the dynamics of network perceptions.

Community Detection with Dependent Connectivity

Annie Qu
Department of Statistics
University of California Irvine

In network analysis, within-community members are more likely to be connected than between-community members, which is reflected in that the edges within a community are intercorrelated. However, existing probabilistic models for community detection such as the stochastic block model (SBM) are not designed to capture the dependence among edges. In this paper, we propose a new community detection approach to incorporate intracommunity dependence of connectivities through the Bahadur representation. The proposed method does not require specifying the likelihood function, which could be intractable for correlated binary connectivities. In addition, the proposed method allows for heterogeneity among edges between different communities. In theory, we show that incorporating correlation information can achieve a faster convergence rate compared to the independent SBM, and the proposed algorithm has a lower estimation bias and accelerated convergence compared to the variational EM. Our simulation studies show that the proposed algorithm outperforms the existing multi-network community detection methods assuming conditional independence among edges. We also demonstrate the application of the proposed method to agricultural product trading networks from different countries and to brain fMRI imaging networks. This is joint work with Yubai Yuan.

Vintage Factor Analysis with Varimax Performs Statistical Inference

Karl Rohe
Department of Statistics
University of Wisconsin

Psychologists developed Multiple Factor Analysis to decompose multivariate data into a small number of interpretable factors without any *a priori* knowledge about those factors [Thurstone, 1935]. In this form of factor analysis, the Varimax "factor rotation" is a key step to make the factors interpretable [Kaiser, 1958]. Charles Spearman and many others objected to factor rotations because the factors seem to be rotationally invariant [Thurstone, 1947, Anderson and Rubin 1956]. These objections are still reported in all contemporary multivariate statistics textbooks. This is an engima because this vintage form of factor analysis has survived and is widely popular because, empirically, the factor rotation often makes the factors easier to interpret. We argue that the rotation makes the factors easier to interpret because, in fact, the Varimax factor rotation performs statistical inference. We show that Principal Components Analysis (PCA) with the Varimax rotation provides a unified spectral estimation strategy for a broad class of modern factor models, including the Stochastic Blockmodel and a natural variation of Latent Dirichlet Allocation (i.e., "topic modeling"). In addition, we show that Thurstone's widely employed sparsity diagnostics implicitly assess a key "leptokurtic" condition that makes the rotation statistically identifiable in these models. Taken together, this shows that the know-how of Vintage Factor Analysis performs statistical inference, reversing nearly a century of statistical thinking on the topic. With a sparse eigensolver, PCA with Varimax is both fast and stable. Combined with Thurstone's straightforward diagnostics, this vintage approach is suitable for a wide array of modern applications. This is in collaboration with Muzhe Zeng.

Inference of Causal Relations with Interventions

Xiaotong Shen
Department of Statistics
University of Minnesota

The inference of causal relations between primary variables is challenging in the presence of unknown interventions. In this article, we infer multiple causal relations while identifying relevant interventions. In particular, we derive conditions for multiple unknown interventions to yield an identifiable model. For inference, we identify the ancestral relations and the interventions for each hypothesis-specific primary variable. Towards this end, we propose a likelihood ratio test based on data perturbation, in which the identification effect is accounted for by perturbing original data to assess the uncertainty associated with identifying ancestors and interventions. For testing the presence and strengths of causal relations in a pathway, we show that the proposed tests achieve desired statistical properties in terms of controlling Types I and II error in a higher-dimensional situation. Numerical examples will be given to demonstrate the utility and effectiveness of the proposed procedure. This work is joint with Chunlin Li and Wei Pan of the University of Minnesota.

Causal Network Construction via Directed Acyclic Mixed Graphs

Peter X.K. Song
Department of Biostatistics
University of Michigan

Directed acyclic mixed graphs (DAMGs) provide a useful representation of network topology with both directed and undirected edges subject to the restriction of no directed cycles in the graph. This graphical framework may arise in many biomedical studies, for example, when a directed acyclic graph (DAG) of interest is contaminated with undirected edges induced by some unobserved confounding factors (e.g., unmeasured environmental factors). Directed edges in a DAG are widely used to evaluate causal relationships among variables in a network, but detecting them is challenging when the underlying causality is obscured by some shared latent factors. We develop an effective structural equation model (SEM) method to extract reliable causal relationships from a DAMG. The proposed approach, termed structural factor equation model (SFEM), uses the SEM to capture the network topology of the DAG while accounting for the undirected edges in the graph with a factor analysis model. The latent factors in the SFEM enable the identification and removal of undirected edges, leading to a simpler and more interpretable causal network. The proposed method is evaluated and compared to existing methods through extensive simulation studies, and illustrated through the construction of gene regulatory networks related to breast cancer. This is a joint work with Drs. Yan Zhou and Xiaoquan Wen.

Factor Models for High-Dimensional Tensor Time Series

Cun-Hui Zhang
Department of Statistics
Rutgers University

Large tensor data are now routinely collected in a wide range of applications due to rapid development of information technologies and their broad implementation in our era. Often such observations are taken over time, forming tensor time series. We present a factor model approach for analyzing high-dimensional dynamic tensor time series and multi-category dynamic transport networks. Two estimation procedures are developed along with iterative projection algorithms to improve them. Theoretical results provide guaranteed convergence rates and proves the benefit of the iterative projections. Simulation results support the theory. Real applications are used to illustrate the model and its interpretations. This is joint work with Rong Chen, Yuefeng Han and Dan Yang.