A102 · Learning Sparse Gaussian Graphical Models from Correlated Data

Presenter: Zeyuan Song — Tufts Medical Center · Tufts University

Authors: Zeyuan Song, Anastasia Leshchyk, Sophia Gunn, Stefano Monti, Gina Marie Peloso, Ching-Ti Liu, Kathryn Lunetta, Paola Sebastiani

Session: Day 2 · Fri Oct 2, 2026 · 2:15–4:15 PM

Gaussian graphical models (GGMs) are widely used in biomedical research to characterize conditional dependence structures among biological, clinical, and social factors. Contemporary biomedical studies increasingly rely on clustered and longitudinal data that produce correlated observations, and ignoring these correlations can generate networks with false-positive edges. To address this challenge, we recently introduced a cluster-based bootstrap algorithm for learning GGMs from correlated data. While this approach effectively controls false-positive edges, it often produces overly connected networks because partial correlations of any magnitude are retained. Here, we propose a statistical framework that retains only edges with partial correlations exceeding a prespecified threshold. By constructing GGMs over a sequence of thresholds ranging from 0 to 1, we can observe the changes of network structure as weaker edges are progressively removed, allowing stronger connections to reveal biologically meaningful modules. We applied this approach to polygenic risk scores, serum metabolomics, and transcriptomic hallmark aging clocks from the Long Life Family Study. By visualizing networks across different levels of sparsity and evaluating their graphical properties, we demonstrate how variables cluster into coherent modules within biological superclasses while identifying key variables that bridge distinct modules. This dynamic network framework provides an interpretable approach for dissecting the organization of complex biological systems and understanding the relationships among interconnected biological processes.

Learning Sparse Gaussian Graphical Models from Correlated Data

Zeyuan Song[1,6,*], Anastasia Leshchyk[1], Sophia Gunn[2], Stefano Monti[3,4], Gina Marie Peloso[5], Ching-Ti Liu[5], Kathryn Lunetta[5], Paola Sebastiani[1,6,7]

1: Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, MA, USA (Zeyuan.Song@tuftsmedicine.org; Anastasia.Leshchyk@tuftsmedicine.org; psebastiani@tuftsmedicalcenter.org)
2: The New York Genome Center, New York, NY, USA (sgunn@nygenome.org)
3: Section of Computational Biomedicine, Boston University School of Medicine, Boston, MA, USA
4: Bioinformatics Program, Boston University, Boston, MA, USA (smonti@bu.edu)
5: Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA (gpeloso@bu.edu, ctliu@bu.edu, klunetta@bu.edu)
6: Tufts University School of Medicine, Boston, MA, USA
7: Data Intensive Study Center, Tufts University, Medford, MA, USA

*: Corresponding author at: Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, MA (Zeyuan.Song@tuftsmedicine.org)

Abstract
Gaussian graphical models (GGMs) are widely used in biomedical research to characterize conditional dependence structures among biological, clinical, and social factors. Contemporary biomedical studies increasingly rely on clustered and longitudinal data that produce correlated observations, and ignoring these correlations can generate networks with false-positive edges. To address this challenge, we recently introduced a cluster-based bootstrap algorithm for learning GGMs from correlated data. While this approach effectively controls false-positive edges, it often produces overly connected networks because partial correlations of any magnitude are retained. Here, we propose a statistical framework that retains only edges with partial correlations exceeding a prespecified threshold. By constructing GGMs over a sequence of thresholds ranging from 0 to 1, we can observe the changes of network structure as weaker edges are progressively removed, allowing stronger connections to reveal biologically meaningful modules. We applied this approach to polygenic risk scores, serum metabolomics, and transcriptomic hallmark aging clocks from the Long Life Family Study. By visualizing networks across different levels of sparsity and evaluating their graphical properties, we demonstrate how variables cluster into coherent modules within biological superclasses while identifying key variables that bridge distinct modules. This dynamic network framework provides an interpretable approach for dissecting the organization of complex biological systems and understanding the relationships among interconnected biological processes.

Keyword: Gaussian Graphical Models, Correlated Data, Sparse Networks, Omics, Polygenic Risk Scores, Metabolomics, Aging Clocks