國立臺灣大學理學院統計與數據科學研究所

碩士論文

Institute of Statistics and Data Science

College of Science

National Taiwan University

Master's Thesis

利用層次圖模型進行社群偵測及圖結構之估計

A Unified Framework for Graph Estimation and Community Detection using Hierarchical Graphical Models

劉宸熙

Chen-Hsi Liu

指導教授: 楊鈞澔 博士

Advisor: Chun-Hao Yang Ph.D.

中華民國 113 年 8 月

August, 2024

doi:10.6342/NTU202401933

Acknowledgements

能完成這篇碩士論文，我必須感謝父母以及指導教授楊鈞澔博士。在求學期間，包括大學以及研究所，父母都贊成我的選擇並給予經濟上的支持。讓我能心無旁騖地鑽研學問。論文能順利地完工，特別感謝楊鈞澔博士的指導與鞭策。楊鈞澔博士總是有問必答，而且速度還很快。迅速解決問題的速度可匹敵某貓型機器人。研究遇上瓶頸的時候，經常依靠楊鈞澔博士提出的建議研究才能有所突破。最後感謝在求學路途上任何幫助過我的人們，感謝之情，無由表達，還是謝天罷。

doi:10.6342/NTU202401933

doi:10.6342/NTU202401933

摘要

圖或者說網絡在社群偵測和圖模型中分別擔任輸入和輸出的角色。由於理解社群結構可提高對圖結構的理解，因此在使用圖模型獲得圖結構之估計後，人們渴望識別潛在的分組。不同於先使用圖模型再對其估計值進行社群偵測，我們的層次圖模型同時估計圖結構和社群結構。該模型將常態-威夏特模型的部分特徵與貝氏社群偵測相融合。最後，我們為後驗推斷開發了一種高效的吉布斯取樣。

關鍵字：社群偵測、圖模型、共變異數選擇、貝氏推論、無限關係模型

doi:10.6342/NTU202401933

doi:10.6342/NTU202401933

Abstract

Graphs or networks respectively serve as input and output in community detection and graphical models. As understanding community structure enriches our comprehension of graphs, there is a desire to identify potential groupings after obtaining a graph estimate using a graphical model. Rather than sequentially applying a graphical model followed by community structure detection, our hierarchical graphical model concurrently estimates both the graph and community structures. This model blends aspects of the normal-Wishart model with Bayesian community detection. Finally, we develop an efficient Gibbs sampler for posterior inference.

Keywords: Community detection, Graphical model, Covariance selection, Bayesian inference, Infinite relationship model

doi:10.6342/NTU202401933

doi:10.6342/NTU202401933

Contents

doi:10.6342/NTU202401933

doi:10.6342/NTU202401933

List of Figures

doi:10.6342/NTU202401933

doi:10.6342/NTU202401933

List of Tables

doi:10.6342/NTU202401933

doi:10.6342/NTU202401933

Chapter 1 Introduction

1.1 Gaussian Graphical Model

Graphical model is a powerful tool for illustrating the conditional dependencies among numerous variables. Its applications span various domains, like protein-signaling (Sachs et al., 2005), disease diagnosis (Sedgewick et al., 2018), financial flows (Giudici and Spelta, 2016), social network analysis (Goodreau, 2007) and image segmentation (Zhang and Ji, 2010). A Gaussian graphical model (GGM) posits that the joint distribution of variables is a Gaussian distribution. The covariance matrix Σ is restricted by the Markov property (Lauritzen, 1996) induced by the graph structure. In essence, GGM is a **covariance estimation** problem. A Bayesian approach entails placing priors on both the covariance matrix and the graph, with many opting for priors linked to the Wishart distribution due to its conjugacy properties. In GGM, the presence or absence of edges constrains elements of the inverse covariance matrix Σ^{-1} to be zero. Consequently, inferring the graph structure is a **covariance selection** problem. Intuitively, we can compute the inverse of the sample covariance matrix, $\hat{\Sigma}^{-1}$, then shrink those components with small values towards zero. In frequentist and Bayesian approaches, selection or shrinkage procedures commonly use penalty functions and shrinkage priors, respectively. As the graph space is discrete, many priors assign a nonzero probability to the exclusion of an edge, resembling

doi:10.6342/NTU202401933

a shrinkage prior. There are several objective choices for priors on the graph: a uniform distribution where all graphs have the same probability, and a binomial distribution on edge connectivity:

$$P(G) \propto p^{E^+}(1-p)^{E^-}, \tag{1.1}$$

where p is the link probability, E^+ and E^- are the number of links and nonlinks of edge.

1.2 Community Detection in Graphs

In network analysis, community detection, also known as graph clustering or network clustering, is increasingly gaining popularity. In Figure 1.1, we show a collaboration network among scientists. Each node in this network corresponds to a scientist, and the shape of each node indicates their particular research area. Edges connecting nodes indicate that the two scientists have collaborated on at least one paper together. The presence of community structure enhances our understanding of interactions between vertices and offers an alternative perspective on vertex relationships. While it may seem intuitive that a node with numerous connections is significant, even in the absence of community structure, nodes that bridge different clusters play a crucial role. These nodes represent interdisciplinary scientists, whose connections, although fewer, are equally essential.

Imagine Figure 1.1 without a group label, community detection aims to figure out how vertices are organized into groups, called communities or clusters. According to Fortunato and Hric (2016), there are five primary types of methods used to identify communities.

doi:10.6342/NTU202401933

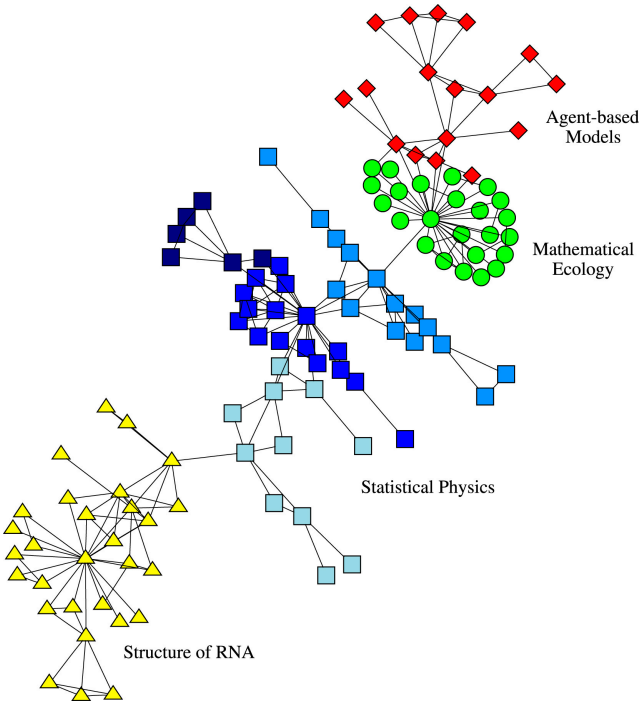


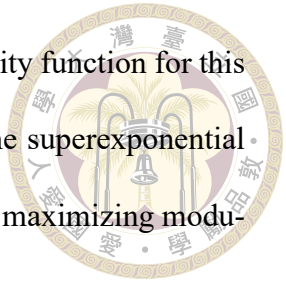
Figure 1.1: Collaboration network of scientists working at the Santa Fe Institute. Nodes represent scientists, and the shape of each node reflects their research areas. Edges connect coauthors. Reprinted figure from Girvan and Newman (2002).

1. Spectral: The eigenvectors of graph matrices, such as the adjacency matrix and Laplacian, encapsulate group information. Spectral clustering involves transforming each node into Euclidean space based on these eigenvectors (Luxburg, 2007). Once the entire graph is mapped into Euclidean space, standard clustering algorithms like k-means can be applied.

2. Statistical inference: Statisticians often address problems using probability models. The Stochastic Block Model (SBM, Nowicki and Snijders (2001)) stands as the predominant generative model for graphs with community structure. It describes the generative process of a graph by randomly partitioning vertices into different groups. While its likelihood resembles (1.1), the link probability in SBM is not a constant p; rather, it varies based on the group index of the nodes.

3. Optimisation: This approach aims to maximize the quality function that describes

doi:10.6342/NTU202401933



the goodness of clusters within a graph. A commonly used quality function for this purpose is modularity (Newman and Girvan, 2004). Due to the superexponential growth in the number of possible graphs with increasing nodes, maximizing modularity often necessitates approximations.

4. Dynamics process: Propose a dynamic process on a graph and make inferences based on the realization of the process (Zhou, 2003). For example, consider suggesting a random walk starting from node i, whose outcome is a sequence of nodes. If a node j appears multiple times in the sequence, it indicates proximity to node i and suggests a higher likelihood of belonging to the same community as node i.

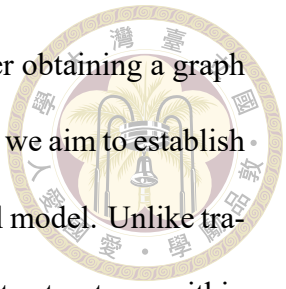
5. Dynamics clustering: This approach requires a sequence of graphs, denoted as $G_1, G_2, ..., G_n$, to represent the growth of the graph over time. Imagine graphs as plants, the root, stem, leaf, and flower correspond to different communities within the graph. As the plant grows, its organs extend from the existing body, mirroring the expansion of communities within the evolving graph structure. Hence, we can utilize the disparity between G_t and G_{t+1} to detect clusters (Asur et al., 2007).

In addition to the aforementioned, there exists a plethora of diverse techniques in community detection. We will not delve extensively into all of these methods, aside from the probability model, which plays a pivotal role in our main work.

1.3 Motivation and Methodology

Community detection and graphical models share a common component: graphs. Graphs respectively serve as input and output in community detection and graphical models. By exploring the community structure, we can gain deeper insights into the graph.

doi:10.6342/NTU202401933



Thus, it is unsurprising that we can perform community detection after obtaining a graph estimator $\hat{G}$ using a graphical model. Rather than a two-step approach, we aim to establish a one-step approach by incorporating group structure into the graphical model. Unlike traditional community detection, which focuses on uncovering community structures within a given graph, our task resembles traditional clustering throughout a hierarchical graphical model.

To estimate a graph with community structure, Tan et al. (2015) proposed cluster graphical LASSO, a two-stage method. First, they identify the community structure based on the empirical covariance matrix. Then, they apply the graphical LASSO (Friedman et al., 2007) to each block. Sun et al. (2014) made a Bayesian GGM with block structure on the inverse covariance matrix Σ^{-1}. However, both models have a significant drawback: they ignore the edges between different communities. For example, in the collaboration network, connections between different clusters represent interdisciplinary programs. If these links are absent, identifying interdisciplinary scientists becomes challenging.

Castelletti et al. (2018) made the prior (1.1) more flexible by introducing a beta prior on the link probability. This motivates the idea that we can establish statistical properties through a hierarchical prior on the graph. Additionally, Mørup and Schmidt (2012) proposed the Bayesian community detection (BCD) model, which improves upon the stochastic block model by incorporating a hierarchical structure on link probability. If we assert that connections within clusters should be denser than those between them, it's logical to consider the BCD model as a generative prior for (1.1) to introduce group structure.

The main contributions of the thesis are twofold. Firstly, instead of estimating $\hat{G}$ and subsequently conducting community detection, we propose a one-step approach that com-

doi:10.6342/NTU202401933

bines graphical modeling and community detection. This results in a Bayesian hierarchical model capable of simultaneously estimating both the graph and community structure. Secondly, we develop an efficient Gibbs sampler for inference. We observe that Gibbs sampling can be divided into two parts: a graphical part and a BCD part, both of which can be generated by existing algorithms respectively.

The rest of the thesis is organized as follows. In Chapter 2, we briefly review the relevant preliminaries, including graphs, graphical models, and probability models in community detection. Chapter 3 outlines our primary contributions, including a detailed description of the proposed model and the MCMC method employed for inference. In Chapter 4, we compare our method to alternative algorithms through simulations and apply it to ovarian cancer data. Finally, we discuss our findings and provide a summary in Chapter 5.

doi:10.6342/NTU202401933

Chapter 2 Preliminaries

2.1 Graph

First, we establish our notation and define basic graph terminology.

1. Graph,vertices and edge: A graph G comprises two parts: the vertex set V and the edge set E. V is a finite set naming each node, such as {Alice, Bob, ...}. For simplicity, we assume $V = \{1, 2, \ldots, p\}$ with p vertices. An edge $e \in E$ is a pair of vertices $(u, v) \in V \times V$.

2. Directed and undirected: We define an edge $(u, v) \in E$ as directed if the reversed edge $(v, u) \notin E$. If both (u, v) and (v, u) lie in E, the edge is undirected. A graph G is considered a directed (undirected) graph if its edge set only contains directed (undirected) edges. In this paper, we solely focus on undirected graphs.

3. Neighbor: We define u as a neighbor of v if (u, v) is an undirected edge in E. Let $nb(v)$ denote the set of all neighbors of v.

4. Adjacency matrix: Consider a graph $G = (V, E)$ with p nodes. The adjacency matrix A is a $p \times p$ matrix with elements of 0 or 1. Its component A_{ij} equals one if $(i, j) \in E$, and zero otherwise. A straightforward observation is that A is symmetric if and only if the graph is undirected.

doi:10.6342/NTU202401933

5. Subgraph: A subgraph G_A of G contains a subset of vertices $A \subseteq V$ and an induced edge set $E_A = E \cap (A \times A)$.

6. Completeness and clique: A graph is complete if $(u, v) \in E$ for any $u \neq v$, meaning there is an edge between every pair of nodes. A clique C is a maximal complete subgraph of G. By slight abuse of notation, we sometimes treat $C = \{1, 2, 3\}$ as a vertex set.

7. Path: A path in G is a sequence of distinct vertices,$v_0, v_1, \ldots, v_k$ such that $(v_{i-1}, v_i) \in E$ for $i = 1, \ldots, k$.

8. Separated: We say a subset $S \subseteq V$ separates u and v, or u and v are separated by S, if all paths from u to v intersect S. Let A and B be subsets of V. We say S separates A and B if S separates every pair $u \in A$ and $v \in B$.

9. Decomposable: We say (A, S, B) decomposes the graph G if $V = A \cup B$, $S = A \cap B$, S separates A and B, and S induces a complete subgraph. A graph with such a decomposition is termed a decomposable graph.

Decomposable graphs, also known as chordal graphs, play a crucial role in graphical models, as sampling non-decomposable graphs demands significant computational resources. This aspect will be further elaborated on in Section 3.

2.2 Graphical Model

Consider a random variable $y = (y_1, y_2, \ldots, y_p)$ whose joint distribution follows a p-dimensional Gaussian distribution with covariance matrix Σ and zero mean. It is well-known that the vanishing of Σ_{ij} implies independence between y_i and y_j. Similarly, for

doi:10.6342/NTU202401933

the inverse covariance matrix $\Omega = \Sigma^{-1}$, often referred to as the precision matrix, a similar result holds. If Ω_{ij} is zero, then $y_i \perp y_j \mid y_{-\{i,j\}}$, indicating that y_i and y_j are conditionally independent given all other variables. In terms of graphical representation, the absence of an edge between y_i and y_j signifies this conditional independence. Hence, we can express this relationship as:

$$y_i \perp y_j \mid y_{-\{i,j\}} \iff \Omega_{ij} = 0 \iff \text{no edge between } y_i \text{ and } y_j.$$

Looking from another angle, given a graph G, the precision matrix is constrained by its Markov property. The precision matrix Ω lies in $M^+(G)$, the set of symmetric positive-definite matrices with zero entries for $(i,j) \notin E$. A straightforward Bayesian model proposes priors on both the precision matrix and the graph structure as follows:

$$y \mid \Omega \sim N(0, \Omega^{-1}),\ \Omega \mid G \sim \pi(\Omega \mid G),\ G \sim \pi(G). \tag{2.1}$$

Several priors on the covariance matrix have been proposed. For instance, Dawid and Lauritzen (1993) suggested the Hyper-Inverse Wishart (HIW) prior for $\Sigma^{-1} \in M^+(G)$, where G is a decomposable graph. Denoting $\Sigma \mid G \sim HIW(b, D)$, where b and D represent the degrees of freedom and shape matrix, respectively, for the standard Wishart distribution $W(b, D)$. The density of the HIW distribution relies on the perfect sequence of cliques (Lauritzen, 1996). Let $\mathcal{C} = \{C_1, C_2, \ldots, C_t\}$ denote the clique set, and $\mathcal{S} = \{S_2, S_3, \ldots S_t\}$ represent the separator set, where

$$S_j = C_j \cap (C_1 \cup C_2 \ldots \cup C_{j-1}).$$

doi:10.6342/NTU202401933

With $\mathcal{C}$ and $\mathcal{S}$ defined, the density of the HIW distribution is given by

$$p_G(\Sigma \mid b, D) = \frac{\prod_{C\in\mathcal{C}} p(\Sigma_C \mid b, D)}{\prod_{S\in\mathcal{S}} p(\Sigma_S \mid b, D)}, \tag{2.2}$$

where $p(\Sigma_C \mid b, D)$ denotes the inverse-Wishart density with parameter b, D:

$$p(\Sigma_C \mid b, D) \propto |\Sigma_C|^{-(b/2+|C|)} \exp\left(-\frac{1}{2} tr\left(\Sigma_C^{-1} D_C\right)\right).$$

Here, Σ_C is the submatrix induced by the vertex set C, $|\Sigma|$ represents the determinant, and $|C|$ indicates the cardinality of the set. The G-Wishart distribution serves as a generalized version of the HIW distribution, expanding its support to non-decomposable graphs. We denote $\Omega \sim W_G(b, D)$ or $W_A(b, D)$ with its density given by

$$p_G(\Omega \mid b, D) = I(b, D, G)^{-1} |\Omega|^{\frac{b-2}{2}} \exp(-\frac{1}{2} tr(\Omega D)) \mathbf{1}_{\Omega\in M^+(G)},$$

where $I(b, D, G)^{-1}$ is the normalizing constant. This prior is conjugate to the Gaussian distribution and is proper for $b > 1$. When combined with either the HIW or G-Wishart distribution, model (2.1) is referred to as the normal-Wishart model. Since the Wishart distribution is the conjugate prior, many priors on the covariance matrix are Wishart distributions with some modifications. For example, Kundu et al. (2019) proposed the regularized Wishart distribution, which shrinks the small non-diagonal elements of Ω without imposing a prior on G. Additionally, Cao et al. (2016) achieved selection consistency using the DAG-Wishart prior.

In model (2.1), our objective is to generate posterior samples of both Ω and G. However, when our primary interest is in G alone, sampling becomes more efficient by integrating out Ω. Let $Y = (Y_1, Y_2, \ldots, Y_n)^T$ denote the data matrix collecting n samples

doi:10.6342/NTU202401933

from y. The marginal posterior distribution of Ω is available in closed form

$$\begin{aligned}\pi(G \mid Y) &= \pi(G) \int_{M^+(G)} p(Y \mid \Omega, G) p(\Omega \mid G) d\Omega \\ &\propto \pi(G) \int_{M^+(G)} \frac{|\Omega|^{\frac{n}{2}} \exp\left(-\sum_{i=1}^{n} \frac{Y_i \Omega Y_i^T}{2}\right) \times |\Omega|^{\frac{b-2}{2}} \exp\left(-\frac{tr(\Omega D)}{2}\right)}{I(b, D, G)} d\Omega \\ &\propto \pi(G) \frac{I(b^*, D^*, G)}{I(b, D, G)}, \end{aligned} \tag{2.3}$$

where $b^* = b + n$ and $D^* = D + Y^T Y$. Though the normalizing constant has an explicit form (Uhler et al., 2018), computing it for general cases proves challenging. This constant has a simple closed form only when the graph is decomposable, and as indicated by (2.2), it equals the ratio of the normalizing constants of a series of standard Wishart distributions.

Giudici and Green (1999) conducted MCMC sampling exclusively within the space of decomposable graphs due to this advantage. While restricting the graph space indeed renders the ratio computable, it fails to accurately reflect the true distribution. Another drawback is the necessity to check whether a graph is decomposable by drawing G from the proposal distribution every time. In Giudici and Green (1999), the proposal is not symmetric and requires computing the number of decomposable graphs G_{t+1} that can be reached from G_t, imposing a significant computational burden and reducing efficiency. To sample from the normal-Wishart model, encompassing non-decomposable graphs, a straightforward approach is to approximate the normalizing constant $I(b, D, G)$ rather than obtaining its exact value. For more details, refer to Wang and Li (2012).

doi:10.6342/NTU202401933

2.3 SBM and IRM

In the introduction, we introduced several community detection methods, despite the lack of a formal definition for "community." Numerous definitions of "community" have been discussed, but none are universally accepted. Many of these definitions revolve around a central concept: that connections within a community should be denser and stronger than those between communities. From an engineering standpoint, metrics such as edge density and degree within and between communities are potential candidates for characteristic measures. In statistics, we would like to use the probability model to describe edge density.

Let L represent the number of clusters, η denote an $L \times L$ matrix with $\eta_{ij} \in [0, 1]$ indicating the probability of a link between clusters, and $z = (z_1, z_2, ..., z_p)$ be a partition of p vertices. For example, with $p = 4$, $z = (1, 1, 2, 3)$ indicates nodes 1 and 2 are in group one, node 3 is in group 2, and node 4 is in group 3. Note that the representations $(1, 1, 2, 3)$ and (d, d, α, θ) are equivalent. To maintain clarity in notation, we let z_i be an integer between 1 and L. SBM model is given by:

$$A_{ij} \sim Ber(\eta_{z_i z_j}). \tag{2.4}$$

Adopting a Bayesian approach, we introduce a beta-binomial model:

$$A_{ij} \sim Ber(\eta_{z_i z_j}),\ \eta_{lm} \sim Beta(\beta, \beta). \tag{2.5}$$

A major drawback of SBM is that K must be assigned beforehand, yet in real networks, determining the number of potential communities is challenging. One possible approach

doi:10.6342/NTU202401933

is first to estimate K and then apply SBM. Alternatively, it's more straightforward to construct a hierarchical model by introducing another prior on the partition z. This model referred to as a special case of the Infinite Relation Model (IRM, Kemp et al. (2006)) is defined as follows:

$$A_{ij} \sim Ber(\eta_{z_i z_j}),\ \eta_{lm} \sim Beta(\beta, \beta),\ z \sim CRP(\alpha). \tag{2.6}$$

Here, CRP stands for the Chinese Restaurant Process. CRP is a discrete-time stochastic process related to the Dirichlet process. Its distribution can be expressed through conditional probability as follows:

$$P(z_i = l \mid z_1, z_2, \ldots, z_{i-1}) = \begin{cases} \frac{n_l}{i-1+\alpha} & 1 \le k \le L \\ \frac{\alpha}{i-1+\alpha} & k = L+1 \end{cases} \tag{2.7}$$

or by its probability mass function:

$$P(z) = \frac{\alpha^L \Gamma(\alpha)}{\Gamma(p+\alpha)} \prod_{l=1}^{L} \Gamma(n_l). \tag{2.8}$$

Here, n_l represents the size of cluster l and p is the length of z. The conditional distribution (2.7) describes the process as follows: whenever a customer enters the restaurant, they sit at an occupied table with rate n_k or at an empty table with rate $\alpha > 0$.

doi:10.6342/NTU202401933

doi:10.6342/NTU202401933

Chapter 3 Graphical community detection

3.1 Graphical Community Detection model

Mørup and Schmidt (2012) introduced a cluster gap γ to the IRM, reflecting the idea that connections should be stronger within communities. This modification, restricting the non-diagonal elements of η dominated by the diagonal elements, leads to the Bayesian Community Detection model:

$$\begin{aligned}
&\text{Cluster structure} : z \sim CRP(\alpha) \\
&\text{Within Link Probability} : \eta_{ll} \sim Beta(\beta, \beta) \\
&\text{Cluster Gap} : \gamma_l \sim Beta(\theta, \theta),\ x_{lm} = \min\left[\eta_{ll}\gamma_l, \eta_{mm}\gamma_m\right] \\
&\text{Between Link Probability} : \eta_{lm} \sim BetaInc(\beta, \beta, x_{lm}) \\
&\text{Link} : A_{ij} \sim Ber(\eta_{z_i z_j}).
\end{aligned} \tag{3.1}$$

Here, the term $BetaInc(\beta, \beta, x)$ represents the constrained beta distribution in the interval $[0, x]$. The use of x_{lm} and $BetaInc$ ensures that η_{lm} is smaller than both η_{ll} and η_{mm}, resulting in a higher internal link probability compared to external links. Combining (2.1)

doi:10.6342/NTU202401933

and (3.1), we obtain the full model:

$$
\begin{gathered}
y \mid \Omega \sim N(0, \Omega^{-1}),\ \Omega \mid A \sim W_A(b, D),\ A_{ij} \sim Ber(\eta_{z_i z_j}) \\
\text{Cluster structure} : z \sim CRP(\alpha) \\
\text{Within Link Probability} : \eta_{ll} \sim Beta(\beta, \beta) \\
\text{Cluster Gap} : \gamma \sim Beta(\theta, \theta), x_{lm} = \min\left[\eta_{ll}\gamma_l, \eta_{mm}\gamma_m\right] \\
\text{Between Link Probability} : \eta_{lm} \sim BetaInc(\beta, \beta, x_{lm}).
\end{gathered} \tag{3.2}
$$

We name the resultant model (3.2) graphical community detection (GCD). The joint likelihood with n observation is given by:

$$
\begin{aligned}
& P(Y, \Omega, A, z, \eta, \gamma \mid \psi = (b, D, \alpha, \beta, \theta)) \\
= & \left[\prod_{i=1}^{n} (2\pi)^{-\frac{p}{2}} \mid \Omega \mid^{\frac{1}{2}} \exp\left(-\frac{1}{2} Y_i^T \Omega Y_i\right)\right] \\
\times & \left[I(b, D, A)^{-1} \mid \Omega \mid^{\frac{b-2}{2}} \exp\left(-\frac{1}{2} tr(\Omega D)\right)\right] \\
\times & \left[\prod_{l=1}^{L+1} \prod_{j=l}^{L+1} \eta_{lj}^{n_{lj}^{+}} (1-\eta_{lj})^{n_{lj}^{-}}\right] \times \left[\prod_{l=1}^{L+1} \frac{\eta_{ll}^{\beta-1}(1-\eta_{ll})^{\beta-1}}{B(\beta, \beta)}\right] \\
\times & \left[\prod_{l=1}^{L+1} \frac{\gamma_l^{\theta-1}(1-\gamma_l)^{\theta-1}}{B(\theta, \theta)}\right] \times \left[\prod_{l=1}^{L} \prod_{j=l+1}^{L+1} \frac{\eta_{lj}^{\beta-1}(1-\eta_{lj})^{\beta-1}}{B_{x_{lj}}(\beta, \beta)}\right] \\
\times & \left[\frac{\alpha^L \Gamma(\alpha)}{\Gamma(p+\alpha)} \prod_{l=1}^{L} \Gamma(n_l)\right].
\end{aligned}
$$

$B(\beta, \beta), B_x(\beta, \beta)$ are the beta function and incomplete beta function. n_{lj}^{+} and n_{lj}^{-} represent the number of links and nonlinks between cluster l and j. We use $\eta \in M_{(L+1)\times(L+1)}$ and $\gamma \in \mathbb{R}^{L+1}$ because we require pseudo-parameters for potential new clusters when updating

doi:10.6342/NTU202401933

cluster z. After some calculation, we obtain:

$$
\begin{aligned}
& P(Y, \Omega, A, z, \eta, \gamma \mid \psi) \\
\propto & \left[I(b, D, A)^{-1} \mid \Omega \mid^{\frac{b^*-2}{2}} \exp\left(-\frac{1}{2} tr(\Omega D^*) \right) \right] \\
\times & \left[\prod_{l=1}^{L+1} \eta_{ll}^{n_{ll}^+ + \beta - 1} (1 - \eta_{ll})^{\eta_{ll}^- + \beta - 1} \right] \times \left[\prod_{l=1}^{L+1} \gamma_l^{\theta-1} (1 - \gamma_l)^{\theta-1} \right] \\
\times & \left[\prod_{l=1}^{L} \prod_{j=l+1}^{L+1} \frac{\eta_{lj}^{n_{lj}^+ + \beta - 1} (1 - \eta_{lj})^{n_{lj}^- + \beta - 1}}{B_{x_{lj}}(\beta, \beta)} \right] \times \left[\frac{\alpha^L \Gamma(\alpha)}{\Gamma(p + \alpha)} \prod_{l=1}^{L} \Gamma(n_l) \right].
\end{aligned}
\tag{3.3}
$$

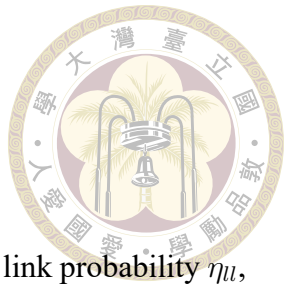
As our main focus lies on A and z, it's more efficient to integrate out the other parameters. Let $\dot{\eta}$ represent the diagonal part of η. Integrating out $\dot{\eta}$ and γ is challenging due to the presence of the term $B_{x_{lj}}(\beta, \beta)$ in the denominator. Non-diagonal element of η can be integrated as incomplete beta function and (3.3) becomes

$$
\begin{aligned}
& P(Y, \Omega, A, z, \dot{\eta}, \gamma \mid \psi) \\
\propto & \left[I(b, D, A)^{-1} \mid \Omega \mid^{\frac{b^*-2}{2}} \exp\left(-\frac{1}{2} tr(\Omega D^*) \right) \right] \\
\times & \left[\prod_{l=1}^{L+1} \eta_{ll}^{n_{ll}^+ + \beta - 1} (1 - \eta_{ll})^{\eta_{ll}^- + \beta - 1} \right] \times \left[\prod_{l=1}^{L+1} \gamma_l^{\theta-1} (1 - \gamma_l)^{\theta-1} \right] \\
\times & \left[\prod_{l=1}^{L} \prod_{j=l+1}^{L+1} \frac{B_{x_{lj}}(n_{lj}^+ + \beta, n_{lj}^- + \beta)}{B_{x_{lj}}(\beta, \beta)} \right] \times \left[\frac{\alpha^L \Gamma(\alpha)}{\Gamma(p + \alpha)} \prod_{l=1}^{L} \Gamma(n_l) \right].
\end{aligned}
\tag{3.4}
$$

With (3.4), we can derive the full conditional distribution and update each parameter individually.

doi:10.6342/NTU202401933

3.2 Gibbs Sampler



We begin by addressing the BCD parameter (η, γ, z). For within link probability η_{ll}, the conditional distribution is proportional to:

$$f(\eta_{ll} \mid \dots) \propto \left[\eta_{ll}^{n_{ll}^{+}+\beta-1}(1-\eta_{ll})^{n_{ll}^{-}+\beta-1}\right] \times \left[\prod_{j\neq l} \frac{B_{x_{lj}}(n_{lj}^{+}+\beta, n_{lj}^{-}+\beta)}{B_{x_{lj}}(\beta,\beta)}\right].$$

We then utilize the Metropolis-Hastings (MH) algorithm with an appropriate proposal distribution to update η_{ll}. Specifically, we employ independent sampling $\pi(\eta_{ll}^{*} \mid \eta_{ll}) \sim Beta(\beta, \beta)$ as the proposal distribution. In the special case where $l = L+1$, as there are no individuals in cluster $L+1$, the parameters $n_{(L+1)j}^{+}$ and $n_{(L+1)j}^{-}$ are zero for all $j = 1, 2, \dots, L+1$. Consequently, the pseudo-parameter can be directly generated from $Beta(\beta, \beta)$.

Similarly, for the cluster gap γ, the conditional distribution is given by:

$$f(\gamma_l \mid \dots) \propto \left[\gamma_l^{\theta-1}(1-\gamma_l)^{\theta-1}\right] \times \left[\prod_{j\neq l} \frac{B_{x_{lj}}(n_{lj}^{+}+\beta, n_{lj}^{-}+\beta)}{B_{x_{lj}}(\beta,\beta)}\right].$$

We select $\pi(\gamma_l^{*} \mid \gamma_l) \sim Beta(\beta, \beta)$ as our proposal distribution. Furthermore, $\gamma(L+1)$ follows $Beta(\theta, \theta)$ for the same reason. During each update, we perform 10 samplings, following the approach outlined in Mørup and Schmidt (2012). To update (Ω, A), we require the non-diagonal part of η in addition to $\dot{\eta}$. According to the likelihood (3.3), we have:

$$f(\eta_{lj} \mid \dots) \sim BetaInc(n_{lj}^{+}+\beta-1, n_{lj}^{-}+\beta-1, x_{lj}).$$

Directly drawing η_{lj}^{*} from an unconstrained beta distribution until it is less than x_{lj} might

doi:10.6342/NTU202401933

encounter difficulties when x_{lj} is extremely small. Therefore, we utilize inverse transform sampling.

For the cluster z, determining the posterior distribution from (3.4) is not straightforward, as the transition from z_i to z_i^* may impact n_{lj}^+, n_{lj}^- and L. We derive the formula of $p(z_k \mid \ldots)$ as follows:

$$\begin{aligned} & P(z_k = l \mid z_{-k}, \eta, \gamma, A, Y, \psi) \\ = & \frac{P(z_k = l, z_{-k}, \eta, \gamma, A, Y \mid \psi)}{P(z_{-k}, \eta, \gamma, A, Y \mid \psi)} \\ = & \frac{P(z_k = l, z_{-k})}{P(z_{-k})} \frac{P(\eta, \gamma, A, Y \mid z_k = l, z_{-k}, \psi)}{P(\eta, \gamma, A, Y \mid z_{-k}, \psi)}. \end{aligned}$$

The first term represents the conditional distribution of CRP (2.7). As for the second term, its denominator is a constant for l and can thus be disregarded. Hence, we arrive at:

$$P(z_k = l \mid z_{-k}, \eta, \gamma, A, Y, \psi) \propto \begin{cases} n_l \times P(\eta, \gamma, A, Y \mid z_k = l, z_{-k}, \psi) & 1 \leq l \leq L \\ \alpha \times P(\eta, \gamma, A, Y \mid z_k = l, z_{-k}, \psi) & l = L + 1 \end{cases} \tag{3.5}$$

The term $P(\eta, \gamma, A, Y \mid z_k = l, z_{-k}, \psi)$ represents (3.4) without the PMF of CRP. This illustrates why we require parameters for an additional group; it simplifies the computation of $P(z_k = L + 1 \mid z_{-k}, \eta, \gamma, A, Y)$. Otherwise, $P(z_k = L + 1 \mid z_{-k}, \eta, \gamma, A, Y)$ becomes nonsensical if $\eta \in M_{L \times L}$ and $\gamma \in \mathbb{R}^L$. When computing (3.5), many duplicate terms can

doi:10.6342/NTU202401933

be canceled out. We can derive the representation with $n_{L+1} = 1$:

$$P(z_k = l \mid \ldots) \propto \alpha^L n_l \times \eta_{ll}^{n_k(l)^+} (1-\eta_{ll})^{n_k(l)^-} \times \prod_{m \neq l} \frac{B_{x_{lm}}(N_k(l,m)^+ + n_k(m)^+ + \beta, N_k(l,m)^- + n_k(m)^- + \beta)}{B_{x_{lm}}(N_k(l,m)^+ + \beta, N_k(l,m)^- + \beta)}, \tag{3.6}$$

where $n_k(l)^+$ and $n_k(l)^-$ denote the number of links and nonlinks between node k and cluster l. $N_k(l,m)^+$ and $N_k(l,m)^-$ represent the number of links and nonlinks between between cluster l and m excluding node k.

For A and Ω, the joint posterior distribution is given by

$$P(\Omega, A \mid \ldots) \propto \left[I(b,D,A)^{-1} \mid \Omega \mid^{\frac{b^*-2}{2}} \exp\left(-\frac{1}{2} tr(\Omega D^*)\right)\right] \times \left[\prod_{l=1}^{L+1} \eta_{ll}^{n_{ll}^+ + \beta - 1} (1-\eta_{ll})^{\eta_{ll}^- + \beta - 1}\right] \times \left[\prod_{l=1}^{L} \prod_{j=l+1}^{L+1} \frac{B_{x_{lj}}(n_{lj}^+ + \beta, n_{lj}^- + \beta)}{B_{x_{lj}}(\beta, \beta)}\right], \tag{3.7}$$

as number of links and nonlinks n^+ and n^- are related to the adjacency matrix A. Using the MH algorithm, where each step involves adding or deleting an edge, leads to the Reversible Jump Markov Chain Monte Carlo (RJMCMC) proposed by Green (1995), which is fundamental in Bayesian graphical models. We may integrate out Ω in the likelihood (3.4) to obtain the marginalized graph posterior:

$$P(A \mid \ldots) \propto \left[\frac{I(b^*, D^*, A)}{I(b, D, A)}\right] \times \left[\prod_{l=1}^{L+1} \eta_{ll}^{n_{ll}^+ + \beta - 1} (1-\eta_{ll})^{\eta_{ll}^- + \beta - 1}\right] \times \left[\prod_{l=1}^{L} \prod_{j=l+1}^{L+1} \frac{B_{x_{lj}}(n_{lj}^+ + \beta, n_{lj}^- + \beta)}{B_{x_{lj}}(\beta, \beta)}\right]. \tag{3.8}$$

The reason $P(A \mid \ldots)$ aligns with (2.3) is that the hierarchical model (3.2) reduces to the normal-Wishart model (2.1) when the BCD parameters are fixed. Therefore, updating

doi:10.6342/NTU202401933

(Ω, A) in each iteration can be done as in the normal-Wishart model. Castelletti et al. (2018) and Consonni et al. (2017) applied MCMC with the marginalized graph posterior and have to face

$$I(b^*, D^*, A^*)I(b, D, A)/I(b, D, A^*)I(b^*, D^*, A)$$

when computing MH ratio. This term can be split into two ratios of normalizing constants: $I(b, D, A)/I(b, D, A^*)$ and $I(b^*, D^*, A^*)/I(b^*, D^*, A)$.

3.3 Approximation for normalizing constant

The normalizing constant can be computed using (2.2) if A is decomposable. For non-decomposable graphs, Uhler et al. (2018) provides an explicit formula for the general $I(b, D, A)$ in Theorem 3.3. However, this formula is too complex to be practically employed. Initially, RJMCMC was designed solely for decomposable graphs due to this complexity. When setting D to be the identity matrix, the formula becomes more manageable, as shown in Corollary 3.4 (Uhler et al., 2018). Nevertheless, it remains computationally expensive and cannot handle $I(b^*, D^*, A)$. Since the exact formula doesn't offer realistic assistance, the next best option is approximation. Let's briefly discuss some approximate methods for computing $I(b, D, A)$.

1. Monte Carlo approximation: Atay-Kayis and Massam (2005) decomposed $I(b, D, A)$ into a determined part with a expected value term. Assume $D = I_p$ and writing $\Omega = \Phi^T\Phi$ as Cholesky decomposition, we obtain:

$$I(b, I_p, A) = \left[\prod_{i=1}^{p} \pi^{\frac{v_i}{2}} 2^{\frac{b+v_i}{2}} \Gamma\left(\frac{b+v_i}{2}\right)\right] \times E\left(e^{-\frac{Q}{2}}\right), \tag{3.9}$$

doi:10.6342/NTU202401933

where $v_i = \mid nb(i) \cap \{i+1, \ldots, p\} \mid$ and

$$Q = \sum_{(i,j) \notin E, i<j} \phi_{ij}^2.$$

Given that $\phi_{ij} \sim N(0,1)$ for $(i,j) \in E$, $\phi_{ii}^2 \sim \chi_{b+v_i}^2$ and Q is a function of them, approximating $I(b, I_p, A)$ reduces to a simple Monte Carlo estimation task on $E\left(e^{-\frac{Q}{2}}\right)$.

2. Laplace approximation: Lenkoski and Dobra (2011) proposed a Laplace approximation to $I(b, D, A)$, given by:

$$\widehat{I(b, D, A)} = (2\pi)^{\frac{|\mathcal{V}|}{2}} \exp\left\{f(\hat{\Omega})\right\} |H(\hat{\Omega})|^{-\frac{1}{2}}$$

where

$$f(\Omega) = \frac{b-2}{2} log|\Omega| - \frac{1}{2} tr(D\Omega), \ \mathcal{V} = \{(i,j) \in V \times V \mid i \leq j, i = j \text{ or } (i,j) \in E\},$$

$\hat{\Omega}$ is the mode of $W_G(b, D)$ and H is the Hessian matrix. The mode $\hat{\Omega}$ can be computed through the iterative proportional scaling algorithm (Speed and Kiiveri, 1986).

3. Exchange algorithm: This approach is also known as the auxiliary variable approach. Murray et al. (2006) considered the joint distribution with an extra variable $p(x, \theta \mid y)$, which preserves the target posterior distribution $p(\theta \mid y)$. Through this strategy, one can cancel out the intractable normalizing constant in the MH ratio using a well-designed proposal distribution.

Based on the reformulation in (3.9), Reza Mohammadi and Letac (2023) approximated the ratio of $I(b, D, A^{-e})/I(b, D, A)$ instead of $I(b, D, A)$ itself, where A^{-e} is A with the

doi:10.6342/NTU202401933

edge e removed. The approximation takes the form:

$$\frac{I(b, I_p, A^{-e})}{I(b, I_p, A)} \approx \frac{1}{2\sqrt{\pi}} \frac{\Gamma\left(\frac{b+d}{2}\right)}{\Gamma\left(\frac{b+d+1}{2}\right)}, \tag{3.10}$$

where d denotes the number of length-two paths connecting the endpoints of e. In the special case when G is decomposable, the approximation (3.10) becomes an equation. Reza Mohammadi and Letac (2023) incorporated (3.10) into the BDMCMC algorithm, which is a continuous-time version of RJMCMC. Their algorithm is implemented in the R package **bdgraph**. We utilize the function **bdgraph** with fifteen iterations to update A in our Gibbs sampling.

doi:10.6342/NTU202401933

doi:10.6342/NTU202401933

Chapter 4 Simulations and real data analysis

4.1 Simulation result

To demonstrate the rationality of simultaneous estimation of graph and community, we compare our model against the two-stage approach, specifically the G-Wishart prior with birth-death MCMC (Mohammadi and Wit, 2015) combined with Bayesian community detection (Mørup and Schmidt, 2012). Additionally, we compare our graph estimation results to those obtained using the Bayesian graphical LASSO (Wang, 2012) and its adaptive version. The aforementioned graphical models are accessible through the **R** packages **BDgraph**, **BayesianGLasso**, and **abglasso**, while Mørup and Schmidt (2012) provided **Matlab** code in the article. Since G-Wishart prior with the birth-death MCMC approach relies on the function 'bdgraph' in the package **BDgraph**, we will refer to this method as bdgraph. We perform the simulation result in the following scenarios:

- Dimension $p = 16, 32$; sample size $n = 3p$.
- G-Wishart parameter $b = 3$; $D = I_p$.
- Two communities, $L = 2$, of equal size with $\eta_{11} = \eta_{22} = 0.9$.

doi:10.6342/NTU202401933

- Between-cluster probability η_{12} ranges over $0, 0.05, 0.1, 0.15, 0.2$.

The synthetic data is generated according to the likelihood specified in (3.2): (1) Sample $A_{ij} \sim Ber(\eta_{z_i z_j})$ independently. Figure 4.1 shows an example of A under different η_{12} sampling. (2) $\Omega \mid A \sim W_A(b, D)$ (3) Generate n samples following the Gaussian distribution, $y \mid \Omega \sim N(0, \Omega^{-1})$.

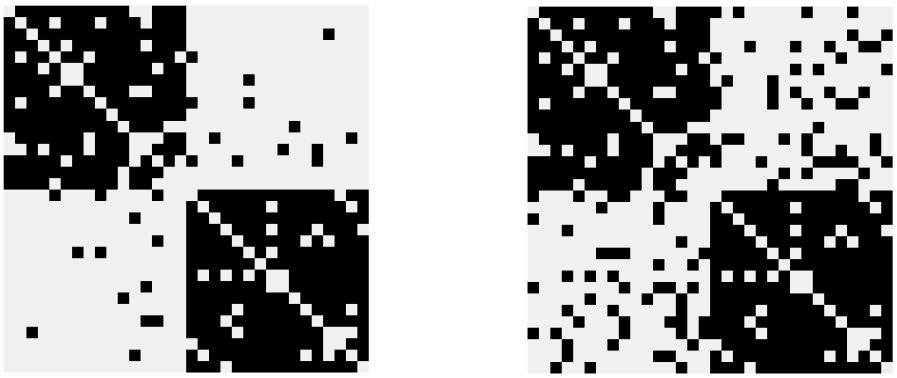


Figure 4.1: Adjacency matrix with $\eta_{12} = 0.05$ (left) and $\eta_{12} = 0.20$ (right) in the data generating process. The black dot indicates the edge and the white dot indicates the absence of edge.

Our inference involves estimating the graph A and the community structure Z. In graph estimation, two common approaches are Maximum A Posteriori (MAP) estimation and the inclusion probability method. To obtain the MAP estimator, we can either use the mode of posterior samples or record the posterior likelihood and identify the highest one. However, since the graph space grows super-exponentially for p, using the mode of posterior samples requires a long chain for good performance. The latter method is only feasible for decomposable graphs due to the normalizing constant $I(b, D, A)$; hence it cannot be applied here.

Next, the inclusion probability of an edge $e = (u, v)$ is given by $P(A_{uv} = 1 \mid Y)$, which can be estimated by the sample mean of A_{uv}. The graph estimator $\hat{A}$ consists of those edges with an inclusion probability higher than a given threshold. When the

doi:10.6342/NTU202401933

threshold is set to 0.5, this approach is known as the median probability model and has been shown to be predictively optimal by Barbieri and Berger (2004). Instead of a 0.5 threshold, we use the Area Under the Curve (AUC) to evaluate the performance of those above Bayesian graphical models.

For the community detection, we suggest the **mode in mode** estimator. We first take mode on the number of clusters $\hat{L} = l$ among all posterior samples. After collecting the posterior samples with $L = l$, we take mode again on the cluster z among these samples. We use the normalized mutual information (NMI) to measure the performance of the community estimation. NMI measures the similarity of two partitions and is defined as $\frac{2I(z_1,z_2)}{H(z_1)+H(z_2)}$, where the $I(\cdot,\cdot)$ is the mutual information and $H(\cdot)$ is the entropy. The value of NMI ranges from 0 (two partitions are independent) to 1 (two partitions are identical).

Our simulation results for AUC and NMI are presented in Tables 4.1 and 4.2, respectively, based on 100 repetitions. In each scenario, we set the G-Wishart parameters $b = 3; D = I_p$, consistent with the data generating process, and used $\beta = \theta = 1$. The parameter α varies with the dimension p, being $\alpha = 1$ when $p = 16$ and $\alpha = 0.8$ when $p = 32$. For each repetition, the MCMC iterations of each algorithm are as follows:

1. $p = 16$
    - GCD: 600 iterations, with the first 100 iterations burned.
    - BDMCMC: 6000 iterations, burning the first half, followed by BCD with 600 iterations, burning the first 100.
    - Bayesian graphical LASSO and its adaptive version: 6000 iterations, burning the first half.
2. $p = 32$

doi:10.6342/NTU202401933

- GCD: 600 iterations, with the first 100 iterations burned.
- BDMCMC: 8000 iterations, burning the first half, followed by BCD with 600 iterations, burning the first 100.
- Bayesian graphical LASSO and its adaptive version: 8000 iterations, burning the first half.

	$p = 16$				$p = 32$			
η_{12}	GCD	bdg+BCD	bglasso	abgl	GCD	bdg+BCD	bglasso	abgl
0	**0.97** (2.5)	0.88 (5.3)	0.81 (4.3)	0.90 (4.6)	**0.98** (0.6)	0.88 (3.1)	0.84 (2.5)	0.93 (2.3)
0.05	**0.93** (5.8)	0.81 (5.7)	0.81 (4.3)	0.83 (5.1)	**0.95** (1.2)	0.80 (3.9)	0.81 (2.2)	0.85 (3.0)
0.1	**0.87** (7.7)	0.78 (5.3)	0.79 (4.4)	0.79 (5.4)	**0.92** (4.1)	0.77 (2.6)	0.79 (2.6)	0.81 (3.2)
0.15	**0.83** (8.2)	0.76 (5.2)	0.77 (4.6)	0.77 (5.2)	**0.88** (5.1)	0.74 (2.6)	0.77 (2.1)	0.78 (3.1)
0.2	**0.79** (8.7)	0.74 (5.3)	0.76 (4.6)	0.75 (4.8)	**0.84** (6.3)	0.73 (2.6)	0.75 (2.3)	0.75 (3.0)

Table 4.1: The AUC of graph estimation (averaged over 100 repetitions), the standard deviation is given in scale of 10^{-3}. bdg,bglasso, and abgl refer to BDgraph, Bayesian graphical LASSO, and adaptive Bayesian graphical LASSO. A higher AUC means better performance.

	$p = 16$				$p = 32$			
η_{12}	GCD	bdg+BCD	bglasso	abgl	GCD	bdg+BCD	bglasso	abgl
0	**0.88** (1.2)	0.86 (2.0)	NA	NA	0.88 (1.1)	**1.00** (0.1)	NA	NA
0.05	**0.78** (1.8)	0.61 (3.7)	NA	NA	0.78 (1.1)	**0.95** (0.9)	NA	NA
0.1	**0.66** (2.4)	0.40 (3.8)	NA	NA	0.68 (1.4)	**0.87** (1.5)	NA	NA
0.15	**0.54** (2.5)	0.24 (3.1)	NA	NA	0.62 (1.5)	**0.70** (3.2)	NA	NA
0.2	**0.48** (2.7)	0.21 (2.9)	NA	NA	**0.52** (1.7)	0.45 (3.9)	NA	NA

Table 4.2: The NMI for community detection (averaged over 100 repetitions), the standard deviation is given in scale of 10^{-2}. A higher NMI means better performance.

For graph estimation, GCD consistently outperforms the other three methods across all scenarios. However, the difference diminishes as η_{12} increases, reflecting the greater difficulty in identifying the true community structure with higher η_{12}. When the ratio η_{11}/η_{12} and η_{22}/η_{12} approach 1, the distinction between the two groups becomes less clear as Figure 4.1 shows. Figure 4.2 and 4.3 show the posterior distribution generated by GCD. A smaller η_{12} captures the true L and enhances the clarity of the group structure, thereby

doi:10.6342/NTU202401933

improving the accuracy of graph estimation, and vice versa. GCD establishes a beneficial feedback loop between the graph and the community structure.

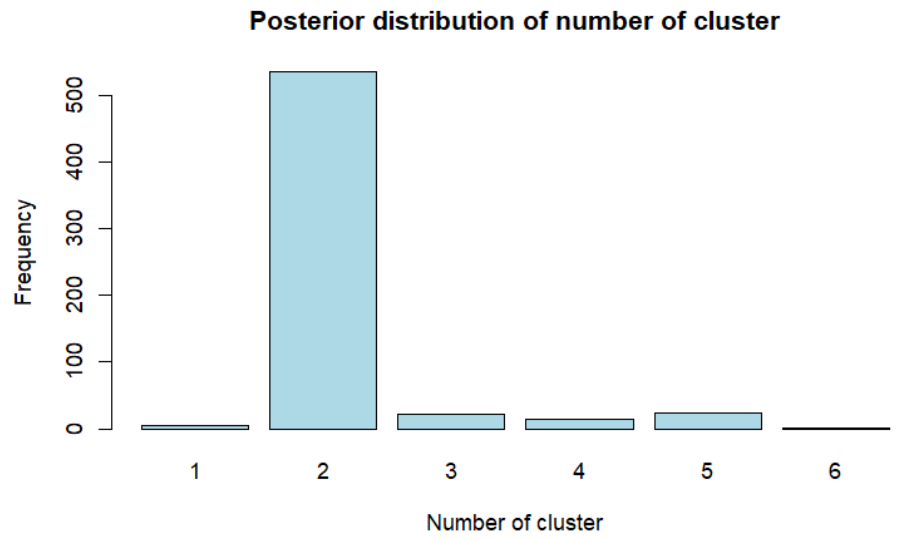


Figure 4.2: Posterior distribution of number of cluster under $p = 32, \eta_{12} = 0.05$. The length of the GCD chain is 600.

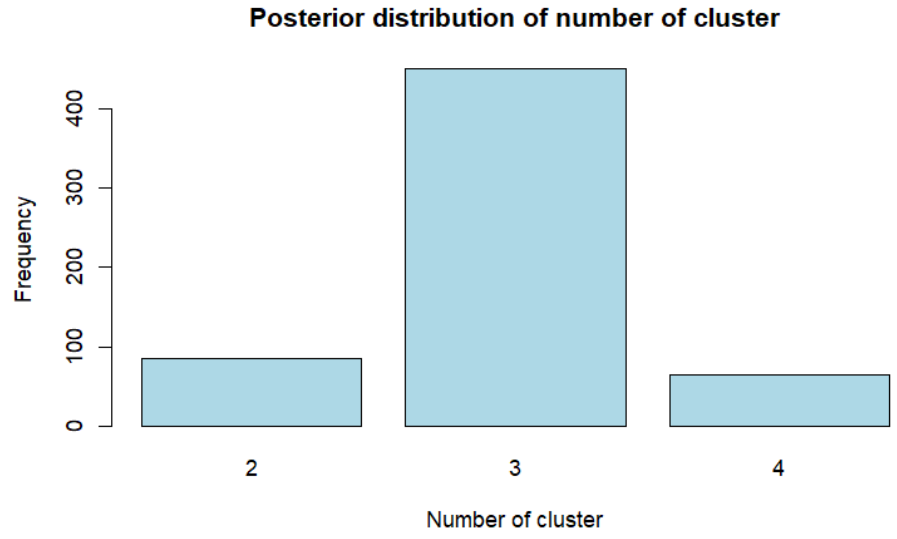


Figure 4.3: Posterior distribution of number of cluster under $p = 32, \eta_{12} = 0.20$. The length of the GCD chain is 600.

In community detection, GCD outperforms bdg+BCD for $p = 16$ but is surpassed

doi:10.6342/NTU202401933

for $p = 32$. GCD maintains stable performance as p increases, whereas the NMI of the two-stage approach improves. This enhancement can be attributed to the increased characteristic size. In our data generation process, matrix A exhibits similar patterns for $p = 16$ and $p = 32$ but a larger scale for $p = 32$. However, higher dimensions do not boost the NMI of GCD. The space of cluster structures expands more rapidly than the integer partition, which grows as $\exp\left\{\pi\sqrt{\frac{2p}{3}}\right\}/4p\sqrt{3}$. Hence, an effective mode estimator requires adequate support from a sufficiently long chain. Since the likelihood in BCD is available, they mitigate this issue by selecting the cluster with the highest posterior likelihood.

4.2 TCGA ovarian cancer

The Cancer Genome Atlas (TCGA) is a landmark cancer genomics program that began in 2006, spearheaded by the National Cancer Institute and the National Human Genome Research Institute in the United States. TCGA project has provided the most comprehensive genomic data resource from over 33 types of cancers. TCGA datasets often include various types of data e.g. gene expression, microRNA expression, DNA methylation profiles, protein expression, etc. Cancer is driven by complex interactions between multiple genes and signaling pathways. Graphical models can represent these interactions and dependencies, providing a comprehensive picture of the underlying biological processes. We apply GCD to the ovarian cancer data, one of the largest datasets in the TCGA project. By inferring the gene regulatory network and discovering potential group structure, researchers can identify key regulatory genes and potential biomarkers that are crucial for the development and progression of ovarian cancer.

The gene expression data consists of $n = 578$ samples with measurements for $p =$

doi:10.6342/NTU202401933

$13,104$ genes. Generally, a graphical model has no restriction on dimension and sample size but it's more practical to reduce the number of variables by prior knowledge. Referring to the study of Shutta et al. (2022), they first selected 156 genes where 59 genes are downregulated in mucinous ovarian tumors and other 97 genes are upregulated. Next, they applied the graphical LASSO to these 156 genes and made an interaction graph. We select 21 genes based on the topological properties of their estimated graph and apply our GCD to this subset with $(p, n) = (21, 578)$.

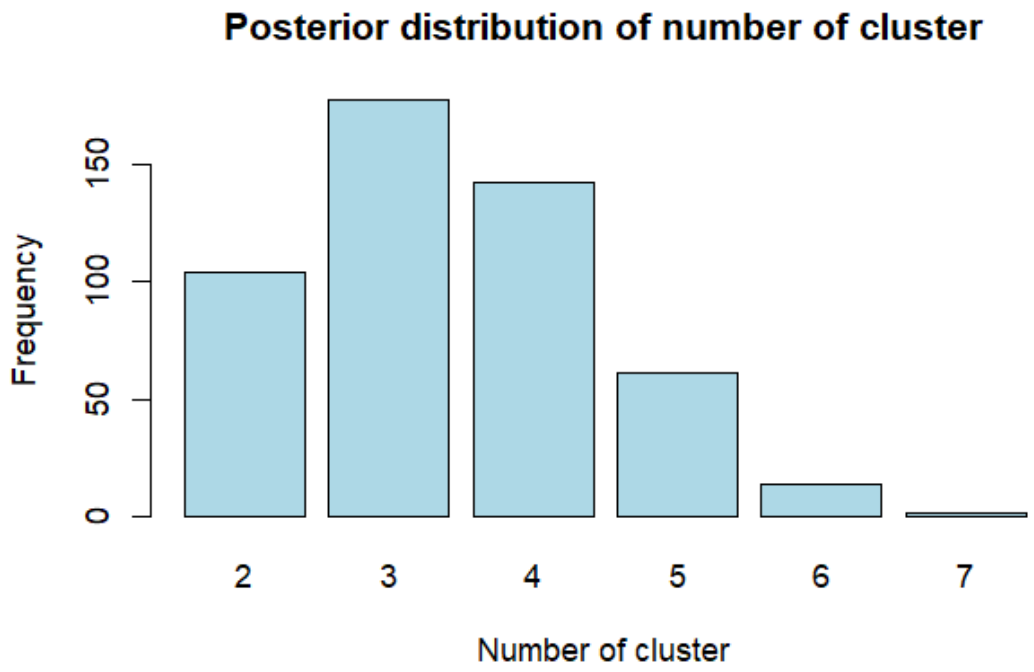


Figure 4.4: Posterior distribution of number of clusters generated by GCD in ovarian cancer data analysis. After burning the first 100 samples, the barplot shows the remaining 500 samples.

Instead of the mode in mode method, we take the mode of z directly. Although Figure 4.4 shows that the mode of L is 3, the frequency of $L = 3$ is scattered across many different z values, while the frequency of $L = 2$ is concentrated mostly on a single z.

Shutta et al. (2022) obtained the group structure by a two-stage approach with the graphical LASSO followed by the walktrap. Walktrap has been introduced in the dynamics

doi:10.6342/NTU202401933

process in the chapter 1. In addition to GCD, we also apply bdgraph + BCD method to the data. Figure 4.5 shows the analysis result of these three approaches. Graphical + walktrap method indicates four clusters while bdgraph + BCD and GCD only find two. The reason is that the analysis of graphical LASSO + walktrap is based on 156 genes. We extract the subgraph induced by the 21 genes from the whole graph. Overall, the three approaches all agree there are two communities among 21 genes. One group shows up in the upper left corner consisting of BASP1, DAB2, FLRT2, HEPH, and PDGFD, and four of them, BASP1, DAB2, FLRT2, and PDGFD are downregulated genes. In other words, the group in the bottom right corner containing 16 genes are all upregulated genes except for one. GCD finds the correct "**regulate structure**" without modeling the mean of gene expression.

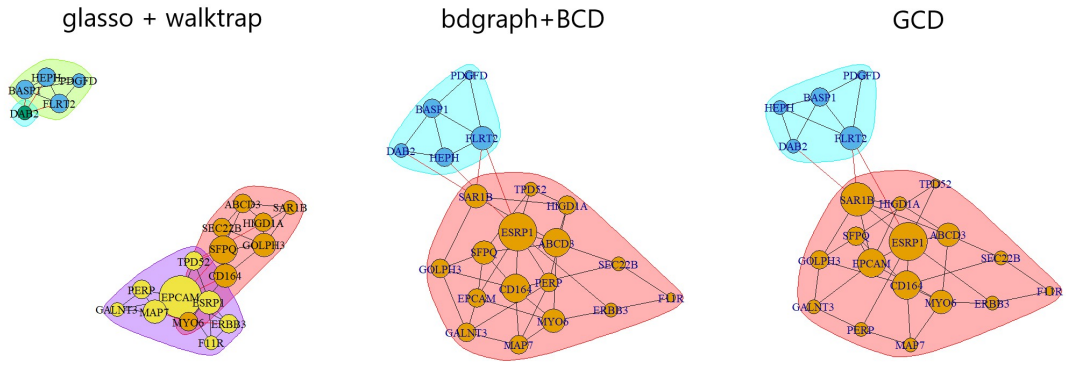


Figure 4.5: Graph and community estimation by graphical LASSO + walktrap, bdgraph + BCD, GCD.

doi:10.6342/NTU202401933

Chapter 5 Discussion

5.1 Discussion

In this article, we introduce a novel graphical model, GCD, designed for simultaneous estimation of both graph and community structures. Our model, GCD, integrates the normal-Wishart model and BCD. Specifically, we incorporate a group structure into the normal-Wishart model using a graph-generating process with blocks. The entire BCD model serves as a complex graph prior $\pi(G)$. Consequently, the Gibbs sampler can be partitioned into two components: one for the graph and another for the community. Each component can be executed by existing algorithms effectively. Given our interest in jointly estimating both the graph and community, it is reasonable to pursue their combined estimation. In our simulations, we demonstrate that the joint estimation of (z, A) outperforms separate estimations, particularly in scenarios with small dimensions p, and remains competitive in larger p cases. Moreover, GCD does not enforce $\Omega_{ij} = 0$ even if variables i and j do not belong to the same group. Notably, when between link probabilities are low, GCD exhibits significant advantages due to the mutually beneficial interactions between the graph and community structures.

doi:10.6342/NTU202401933

doi:10.6342/NTU202401933

References

Asur, S., Parthasarathy, S., and Ucar, D. (2007). An event-based framework for characterizing the evolutionary behavior of interaction graphs. In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '07, page 913–921, New York, NY, USA. Association for Computing Machinery.

Atay-Kayis, A. and Massam, H. (2005). A Monte Carlo method for computing the marginal likelihood in nondecomposable Gaussian graphical models. Biometrika, 92(2):317–335.

Barbieri, M. M. and Berger, J. O. (2004). Optimal predictive model selection. The Annals of Statistics, 32(3):870–897.

Cao, X., Khare, K., and Ghosh, M. (2016). Posterior graph selection and estimation consistency for high-dimensional Bayesian DAG models. The Annals of Statistics, 47:319–348.

Castelletti, F., Consonni, G., Vedova, M. L. D., and Peluso, S. (2018). Learning Markov equivalence classes of directed acyclic graphs: An objective Bayes approach. Bayesian Analysis, 13(4):1235–1260.

Consonni, G., Rocca, L. L., and Peluso, S. (2017). Objective Bayes covariate-adjusted sparse graphical model selection. Scandinavian Journal of Statistics, 44(3):741–764.

doi:10.6342/NTU202401933

Dawid, A. P. and Lauritzen, S. L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. The Annals of Statistics, 21(3):1272–1317.

Fortunato, S. and Hric, D. (2016). Community detection in networks: A user guide. Physics Reports, 659:1–44.

Friedman, J., Hastie, T., and Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical LASSO. Biostatistics, 9(3):432–441.

Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. Proceedings of the National Academy of Sciences, 99(12):7821–7826.

Giudici, P. and Green, P. (1999). Decomposable graphical Gaussian model determination. Biometrika, 86(4):785–801.

Giudici, P. and Spelta, A. (2016). Graphical network models for international financial flows. Journal of Business & Economic Statistics, 34(1):128–138.

Goodreau, S. M. (2007). Advances in exponential random graph (p*) models applied to a large social network. Social Networks, 29(2):231–248.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika, 82(4):711–732.

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., and Ueda, N. (2006). Learning systems of concepts with an infinite relational model. In Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1, AAAI'06, page 381–388. AAAI Press.

Kundu, S., Mallick, B. K., and Baladandayuthapani, V. (2019). Efficient Bayesian regularization for graphical model selection. Bayesian Analysis, 14(2):449–476.

doi:10.6342/NTU202401933

Lauritzen, S. L. (1996). Graphical models. Oxford University Press.

Lenkoski, A. and Dobra, A. (2011). Computational aspects related to inference in Gaussian graphical models with the G-Wishart prior. Journal of Computational and Graphical Statistics, 20(1):140–157.

Luxburg, U. (2007). A tutorial on spectral clustering. Statistics and Computing, 17(4):395–416.

Mohammadi, A. and Wit, E. C. (2015). Bayesian structure learning in sparse Gaussian graphical models. Bayesian Analysis, 10(1):109–138.

Murray, I., Ghahramani, Z., and MacKay, D. J. C. (2006). MCMC for doubly-intractable distributions. In Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence, UAI'06, page 359–366, Arlington, Virginia, USA. AUAI Press.

Mørup, M. and Schmidt, M. (2012). Bayesian community detection. Neural computation, 24:2434–2456.

Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. Phys. Rev. E, 69:026113.

Nowicki, K. and Snijders, T. A. B. (2001). Estimation and prediction for stochastic block-structures. Journal of the American Statistical Association, 96(455):1077–1087.

Reza Mohammadi, H. M. and Letac, G. (2023). Accelerating Bayesian structure learning in sparse Gaussian graphical models. Journal of the American Statistical Association, 118(542):1345–1358.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. (2005). Causal

doi:10.6342/NTU202401933

protein-signaling networks derived from multiparameter single-cell data. Science, 308(5721):523–529.

Sedgewick, A. J., Buschur, K., Shi, I., Ramsey, J. D., Raghu, V. K., Manatakis, D. V., Zhang, Y., Bon, J., Chandra, D., Karoleski, C., Sciurba, F. C., Spirtes, P., Glymour, C., and Benos, P. V. (2018). Mixed graphical models for integrative causal analysis with application to chronic lung disease diagnosis and prognosis. Bioinformatics, 35(7):1204–1212.

Shutta, K. H., De Vito, R., Scholtens, D. M., and Balasubramanian, R. (2022). Gaussian graphical models with applications to omics analyses. Statistics in Medicine, 41(25):5150–5187.

Speed, T. P. and Kiiveri, H. T. (1986). Gaussian Markov distributions over finite graphs. The Annals of Statistics, 14(1):138–150.

Sun, S., Zhu, Y., and Xu, J. (2014). Adaptive variable clustering in Gaussian graphical models. In Kaski, S. and Corander, J., editors, Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, volume 33, pages 931–939, Reykjavik, Iceland. PMLR.

Tan, K. M., Witten, D., and Shojaie, A. (2015). The cluster graphical LASSO for improved estimation of Gaussian graphical models. Computational Statistics and Data Analysis, 85:23–36.

Uhler, C., Lenkoski, A., and Richards, D. (2018). Exact formulas for the normalizing constants of Wishart distributions for graphical models. The Annals of Statistics, 46(1):90–118.

doi:10.6342/NTU202401933

Wang, H. (2012). Bayesian graphical LASSO models and efficient posterior computation. Bayesian Analysis, 7(4):867–886.

Wang, H. and Li, S. Z. (2012). Efficient Gaussian graphical model determination under G-Wishart prior distributions. Electronic Journal of Statistics, 6:168–198.

Zhang, L. and Ji, Q. (2010). Image segmentation with a unified graphical model. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(8):1406–1425.

Zhou, H. (2003). Network landscape from a Brownian particle's perspective. Phys. Rev. E, 67:041908.

doi:10.6342/NTU202401933